Heterogeneous Traffic Multiplexing in Next Generation Cellular Networks

Mohammed Al-Mekhlafi

A Thesis
In
The Department of
Information and Systems Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Information and Systems Engineering)
Concordia University
Montréal, Québec, Canada

January 2023

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mohammed Al-Mekhlafi**

Entitled: **Heterogeneous Traffic Multiplexing in Next Generation Cellular Networks**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________	Chair
Dr. Rajamohan Ganesan	
______________________________	External Examiner
Dr. Georges Kaddoum	
______________________________	External to Program
Dr. Jun Cai	
______________________________	Examiner
Dr. Amr Youssef	
______________________________	Examiner
Dr. Roch Glitho	
______________________________	Supervisor
Dr. Chadi Assi	
______________________________	Co-Supervisor
Dr. Ali Ghrayeb	

Approved ______________________________

Dr. Abdessamad Ben Hamza, Graduate Program Director

August 29, 2022 ______________________________

Dr. Mourad Debbabi, Dean

Gina Cody School of Engineering and Computer Science

ABSTRACT

Heterogeneous Traffic Multiplexing in Next Generation Cellular Networks

Mohammed Al-Mekhlafi, Ph.D.
Concordia University, 2023

The vision shaping the upcoming sixth-generation (6G) wireless cellular networks has recently gained considerable attention from researchers in academia and industry. 6G networks are expected to fulfill the limitations of the fifth-generation (5G) networks and support a wide range of new applications and services beyond those supported by 5G, namely, enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC) and massive machine-type communications (mMTC). Further, these emerging networks are thus mandated to support new emerging applications that concurrently demand multiple quality of service (QoS) requirements of data rate, reliability, latency, and connectivity. Due to the fundamental trade-off of such extremely diverse QoS requirements, the coexistence of these emerging applications has been identified as a major challenge in 6G networks and their predecessors. This dissertation aims at addressing the coexistence problem, specifically URLLC and eMBB traffic, by developing spectrally efficient multiplexing and scheduling solutions.

By considering different key enabling technologies, this dissertation provides unique research contributions to the coexistence problem that led to effective designs. In particular, coupling URLLC and eMBB through the Third Generation Partnership Project (3GPP) superposition/puncturing scheme naturally arises as a promising option due to the latter's tolerance in terms of latency and reliability. Moreover, reconfigurable intelligent surface (RIS) has been proposed as a potential low-cost and energy-efficient technology that can control the wireless propagation environment providing endless benefits in supporting coexisting 6G services.

Regarding the superposition scheme, this thesis investigates the joint scheduling of eMBB and URLLC traffic while minimizing the eMBB rate loss, considering

URLLC reliability and the eMBB QoS. In the context of puncturing, this thesis studied the interplay between the RIS configuration, URLLC reliability and eMBB rate by proposing proactive RIS configurations to guarantee the URLLC latency requirements. Although simulation results demonstrate that adopting the proposed scheme can further boost eMBB and URLLC traffic performance, the computational complexity of optimizing the RIS phase shifts is challenging. To this end, this thesis proposes two low-complexity methods for optimizing the RIS phase shift matrix. The first solution proposes reducing the number of optimization variables configuring the RIS to the number of users. The second algorithm is based on a closed-form expression for the RIS phase shift matrix. Finally, a new puncturing strategy is proposed to mitigate the impact on the eMBB transmission. The key idea of the proposed scheme is to puncture the eMBB data that has maximum symbol similarities with the URLLC leading to reducing the contaminated eMBB symbols. We study the performance of the proposed schemes in terms of the eMBB spectral efficiency, URLLC reliability and low complexity. We show analytically and through simulations the efficacy of the proposed schemes over their existing counterparts.

Acknowledgments

First and foremost, Alhamdulillah, all my gratitude goes to Allah. My complete faith in Him and His eternal blessings enabled me to complete this Thesis.

I would like to greatly thank my supervisors, Dr. Chadi Assi and Dr. Ali Ghrayeb. The knowledge, the hard work and the completion of my Ph.D. thesis would not be possible without their motivation, support and continuous guidance during my Ph.D. Thank you for your valuable time and unconditional availability through numerous discussions and meetings. I learned so much from them, and I greatly appreciate all the advice and wisdom.

I would also like to thank the advisory committee members of my Ph.D., Dr. Georges Kaddoum, Dr. Jun Cai, Dr. Amr Youssef, Dr. Roch Glitho and Dr. Walter Lucia. Thank you for your valuable comments, which have helped me substantially improve the quality of this dissertation.

My gratitude also goes to my colleague, Dr. Mohammed Arfaoui and Dr. Mohaned Chraiti . I thank them for their collaborations and their friendship, wishing them the best of luck in their future endeavours. I also thank my brothers, my wife, and my kids. I would finally like to express my deepest gratitude to my parents, to whom this work is dedicated. Without their unconditional support, kind words, and sound advice, I would not be the person I am today.

Contents

List of Figures

List of Tables

Abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth-Generation
6G	Sixth-Generation
AP	Access point
AWGN	Additive White Gaussian Noise
AO	Alternating Optimization
AR	Augmented Reality
BS	Base Station
BPSK	Binary Phase-shift keying
BER	Bit Error Rate
BLER	Block Error Rate
CSI	Channel State Information
DC	Difference of Convex
eMBB	Enhanced Mobile Broadband
FDMA	Frequency Division multiple Access
HARQ	Hybrid Automatic Repeat Request
IoT	Internet of Things
KKT	Karush–Kuhn–Tucker
LTE	Long-Term Evolution
mMTC	Massive Machine-Type Communication

MAC	Medium Access Control
MINLP	Mixed Integer Nonlinear Programming
MR	Mixed Reality
MIMO	Multiple-Input and Multiple-Output
mMIMO	Massive MIMO
NR	New Radio
NGMA	Next-Generation Multiple Access
NOMA	Non-Orthogonal Multiple Access
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RSMA	Rate Splitting Multiple Access
RIS	Reconfigurable Intelligent Surfaces
RB	Resource Block
SDR	Semidefinite Relaxation
SINR	Signal to Noise Ratio
SINR	Signal to Interference and Noise Ratio
SIC	Successive Interference Cancellation
SER	Symbol Error Rate
SCA	Successive Convex Approximation
TTI	Transmission Time Interval
UAV	Unmanned Aerial Vehicle
URLLC	Ultra-Reliable and Low Latency Communications
VR	Virtual Reality
XR	Extended Reality

Chapter 1

Introduction

1.1 Next-Generation Cellular Networks

Given the continually rising demand for data rates and connectivity, the upcoming sixth-generation (6G) wireless networks are expected to support a wide range of new applications and services beyond those supported by the current fifth-generation (5G) wireless networks, including enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC) and massive machine-type communications (mMTC) [5–7]. Thus, 6G networks should fulfill the limitations of 5G networks and support emerging applications that concurrently demand multiple quality of service (QoS) requirements in terms of reliability, rate, latency and connectivity [7], [5]. For example, high reliability, low latency and high data rates are services that are all needed for enabling extended reality (XR), which is one of the envisioned applications of 6G networks [5, 7]. As a result, supporting such heterogeneous services within the same network architecture, specifically, the coexistence of URLLC traffic with other service class, is marked as the major problem in 5G networks and it is going to be expanded in 6G networks.

In order to accommodate future heterogeneous services (enormous traffic demand, diverse and stringent quality of services (QoS) requirements, and massive connectivity), non-conventional technologies and networking architectures are recognized as core solutions for beyond 5G and 6G networks [5, 8]. Recently, next-generation multiple access (NGMA) schemes have gained much attention from researchers in academia and industry to keep track of the dramatic growth of the number of connected devices and the expected high demand for wireless data in next-generation wireless networks [6, 7, 9, 10]. For instance, non-orthogonal multiple access (NOMA) techniques have been widely studied as auspicious candidate due to its capability in supporting high number of users that is larger than the number of available orthogonal resources [11, 12]. Moreover, the research community has recently focused on rate splitting multiple access (RSMA) as a potential next-generation multiple access technique for upcoming wireless networks [13]. Furthermore, massive multiple-input multiple output (mMIMO) is also considered as one of the key components for 5G and 6G to achieve high spectral efficiency and coverage [1, 14]. Meanwhile, new operating frequency bands, such as mm-wave communications and visible light communications technologies, are explored to achieve higher data rates compared to the achieved rates of radio frequency bands [15, 16].

The above mentioned improvements in terms of the achievable data rates, latency, reliability, which are achieved only through the enhancement at the sender or the receiver sides (or both) may not be sufficient to fulfill the strict requirements demanded by future networks [17]. In this regard, reconfigurable intelligent surfaces (RISs) has been recently proposed as a promising low-cost and energy-efficient technology that is able to control the wireless propagation environment leading to enhance the spectral efficiency, latency and reliability [17–19].

On the other hand, since the reliability is coupled with the latency for URLLC applications, it is essential to reduce or eliminate the sources of latency that impact the URLLC service class. According to the Third Generation Partnership Project (3GPP) standard for the 5G new radio (NR), the URLLC latency must be less than 1 msec and the URLLC reliability must be higher than 0.99999 within a time period of less than 1 msec for the case of a URLLC packet of size 32 bytes [1, 20]. Accordingly, the latency, in the context of URLLC, measures the end-to-end time of delivering a packet, whereas the reliability is defined as the probability of the successful delivering of a packet within a limited time period. Based on this, the 3GPP standard proposes a shorter transmission time within the conventional transmission time interval, also known as the eMBB time-slot, in a way to accommodate the URLLC traffic. Precisely, the transmission time is reduced from 1 msec in Long Term Evolution (LTE) to 0.143 msec in 5G new radio, referred to as a mini-slot [1, 3, 20]. Moreover, key

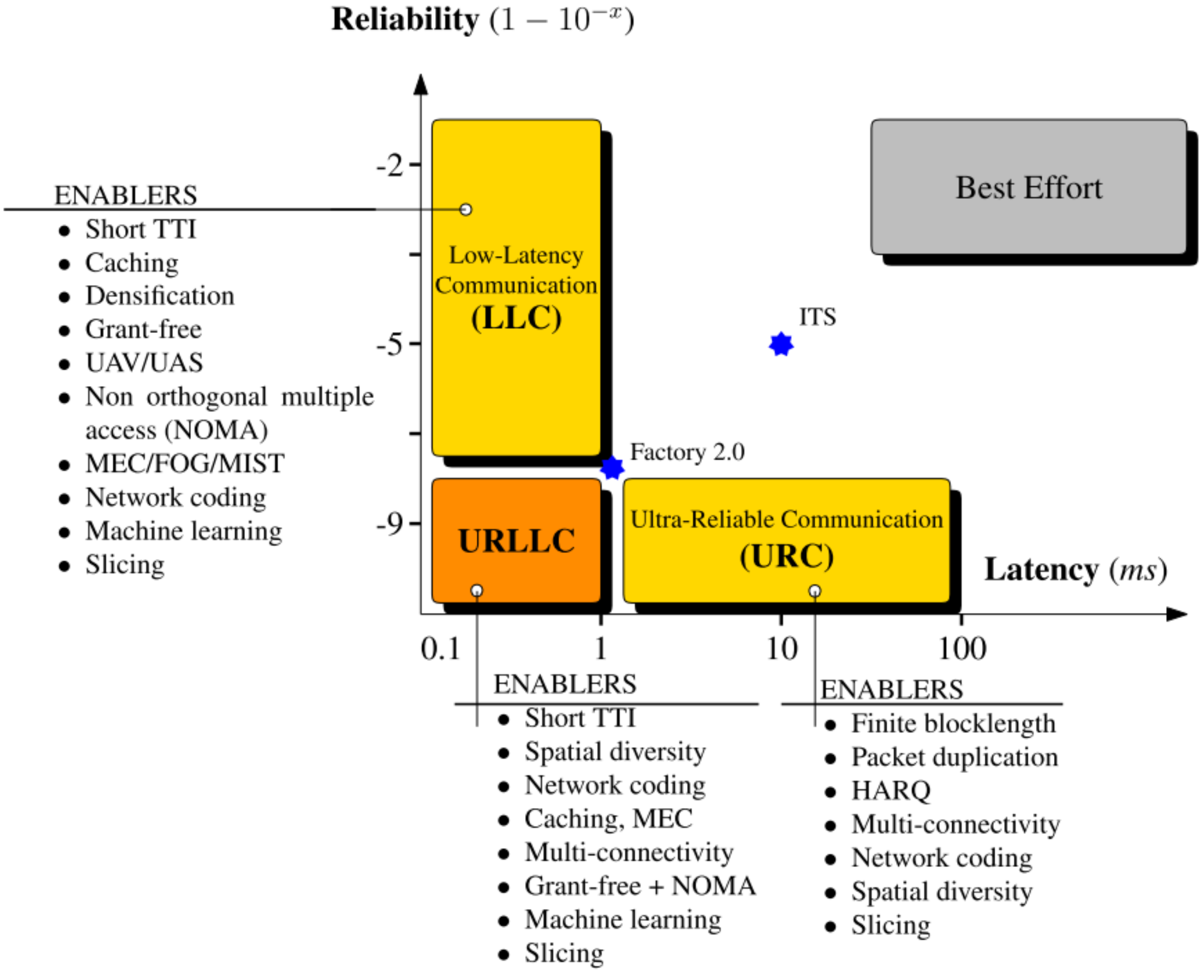


Fig. 1.1: Key enablers for low latency and high reliability [1].

enabling technologies, such as multiple waveform numerology, hybrid automatic repeat request (HARQ), and packet duplication, were introduced to accommodate the URLLC service class. Fig.1.1 presents an overview of the key enablers for low latency and high reliability.

Despite the numerous beneficial applications of URLLC, there are still several challenges which need to be tackled in 5G and beyond networks. These challenges are related to the extremely strict QoS requirements for URLLC and multiplexing of eMBB and URLLC traffic in which eMBB users and URLLC users coexist. Hence, maintaining such strict URLLC QoS requirements while guaranteeing resource availability is not an easy task, as the efficiencies of both spectrum and energy should not be compromised with URLLC.

This chapter presents the main potential applications of URLLC in communication systems. Then, it presents the main challenges associated with URLLC service as well as the key next-generation Enabling Technologies. Finally, the contributions of this dissertation are summarized.

1.2 URLLC Potential Applications

URLLC service class is the main enabler for several emerging applications that have various latency and reliability requirements, such as Industry 4.0, health care, intelligent transportation system (ITS), virtual reality (VR) etc, as shown in Fig.1.2. In this chapter, some URLLC applications are briefly discussed.

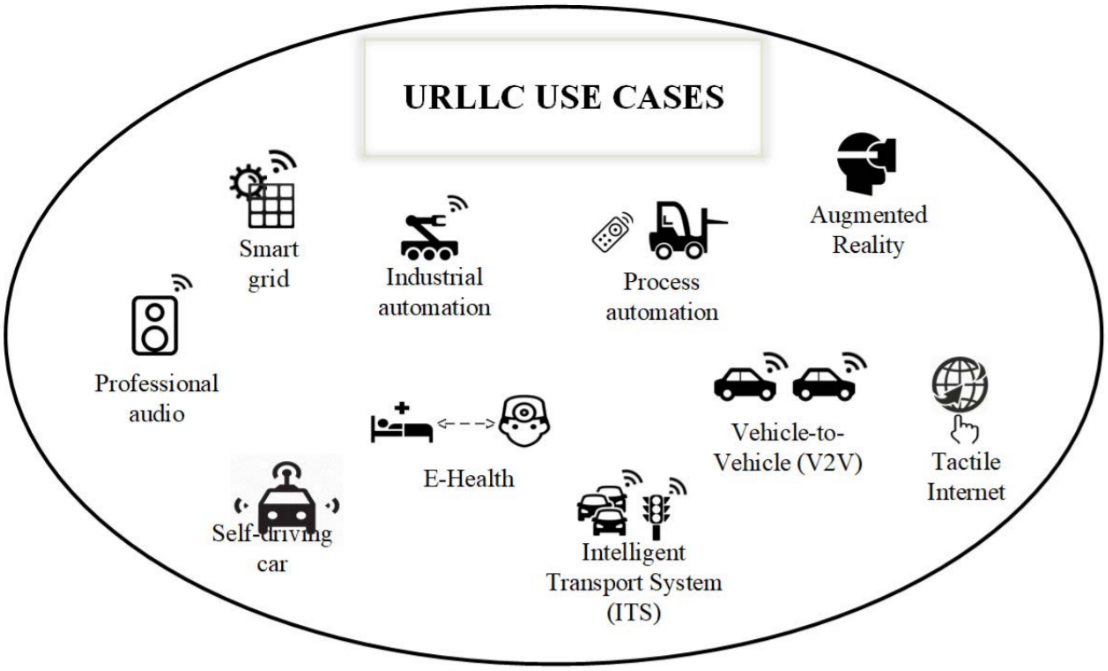


Fig. 1.2: URLLC use cases[2].

1.2.1 Industrial Automation

URLLC is the key enabling technology for the fourth industrial revolution, so-called Industry 4.0, wherein wireless communications replace wired connections [14, 21]. Wireless connections offer low manufacturing, installation, and maintenance costs, and deployment flexibility. In practice, industrial processes such as motion control, factory automation and process automation, require extremely reliable communications between sensors, actuators and controllers. Hence, enabling such applications necessitates extreme reliability of $1 - 10^{-9}$ with an end-to-end latency lower than 0.5 ms [14, 21].

1.2.2 Intelligent Transportation Systems

Another important application of URLLC is to empower technological transformations in the transportation industry [21]. These transformations include autonomous driving with safety and efficiency services. In practice, these shifts require vehicles to be fully networked and connected to collaboratively respond to the complicated road conditions instead of relying on local information. In other words, the information needs to be distributed between vehicles in a reliable manner within short time duration. Hence, the typical requirements of such application require reliability of $1 - 10^{-5}$ and end-to-end latency of 1-5 ms [14, 21, 22].

1.2.3 Health Care

The use cases of URLLC in health care involve remote diagnosis and remote treatment [21].The patient is constantly monitored remotely and communicated via devices that measure vital signs like blood pressure and body temperature. Remote treatment is automatically performed for patients who require an urgent response based on monitoring data. Remote surgical consultations, for example, can occur when a patient has a medical emergency and cannot wait to be carried to the hospital [21]. The typical requirements for remote surgery scenario are extreme reliability of $1 - 10^{-9}$ with an end-to-end latency less than 1 ms [14, 21, 22].

1.2.4 Other Potential Applications

Besides the above potential applications of URLLC, URLLC lies in the overlapped area of the internet of things (IoT) and tactile internet, and smart grid as illustrated in Fig.1.2. For instance, URLLC can provide a deterministic service guarantee with stringent latency and reliability requirements for operational and energy modules in the electrical grid, such as smart meters and devices. As previously stated, URLLC becomes conflated with both mMTC and eMBB. Augmented reality (AR) and mixed reality (MR), for example, necessitate larger data rates in addition to latency and reliability requirements [22, 23]. Table 1.1 summarizes URLLC uses cases and its requirements.

Table 1.1: URLLC use case and requirements [2].

Use case	Latency (ms)	Reliability (%)	Packet Size (bytes)
Smart grid	$3 \sim 20$	99.999	$80 \sim 1000$
Professional audio	2	99.99999	$3 \sim 1000$
Self-driving car	1	99	144
Industrial automation	$0.25 \sim 10$	99.9999999	$10 \sim 300$
Process automation	$50 \sim 100$	99.99	$40 \sim 100$
Health care	$1 \sim 30$	99.9999999	$28 \sim 1400$
Augmented reality	$0.4 \sim 2$	99.999	12k $\sim$ 16k
ITS	$5 \sim 10$	99.999	$50 \sim 200$
Tactile internet	1	99.99999	250

1.3 URLLC Challenges

1.3.1 QoS Requirements

Several delay components may impact the URLLC latency and reliability requirements. In practice, the latency components of the URLLC traffic include 1) transmission latency, which is the time required to transmit a packet; 2) processing latency, which represents the time required to perform the encoding and decoding at the transmitter and the receiver sides, respectively; 3) The retransmission latency that is required to perform HARQ in case of a decoding failure; and 4) The signalling latency required for a connection request, scheduling grant, channel training and feedback, and queueing.

All these sources of latency have to be controlled so that the anticipated URLLC QoS is maintained [20]. As a physical layer enabler, multiple waveform numerology has been proposed to support the latency requirements. Accordingly, the subcarrier spacing can have a bandwidth of $\{15, 30, 60, 120, 240\}$ kHz, while the transmission time interval (TTI) is also divided into mini-slots with sub-millisecond duration

$\{143, 66.77, 33.33, 16.67, 8.33\}$ μsec [24]. As a result, the URLLC TTI duration can be controlled based on the cell size and the operating frequency band. Although reducing the TTI duration enhances the service reliability and the service latency as queuing delay and the time needed for HARQ retransmissions, it involves more signal control overhead hence the availability of resources for other URLLC data transmissions is impacted [1].

1.3.2 Coexistence with eMBB

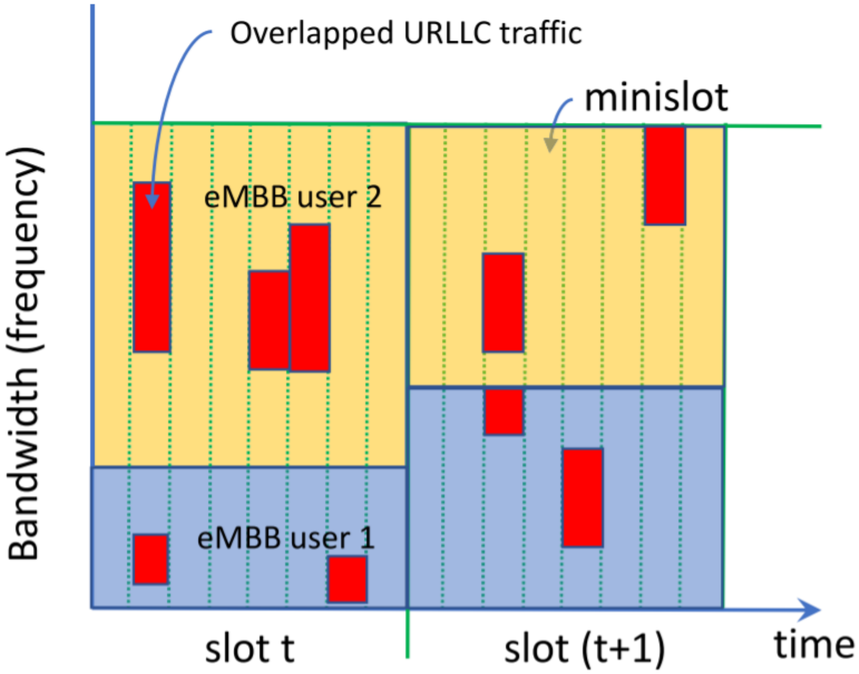


Fig. 1.3: Illustration of superposition/puncturing approach for multiplexing eMBB and URLLC [3].

Since the available bandwidth is limited, the eMBB and URLLC service classes could coexist in the same spectrum. I In order to accommodate both traffic, reservation based scheduling and instant scheduling (also called on-air resource allocation) have been proposed to enable the immediate scheduling of the URLLC traffic [20]. The reservation based scheduling consists of reserving a part of the resources to the URLLC load prior to its arrival whereas the instant scheduling consists of serving the URLLC packets immediately once arrived, immediately transmitted in the next mini-time slot, by interrupting the ongoing eMBB traffic [20]. Within the instant scheduling strategy, the puncturing/superposition schemes have been proposed by the 3GPP standard to promote the support of URLLC traffic [3, 20]. Puncturing refers to preempting part of the eMBB frequency resources by allocating zero power to the eMBB traffic to allocate the URLLC packets whereas superposition refers to multiplexing both eMBB and URLLC using superposition coding by allocating a fraction of the eMBB power to the URLLC packets. Because of the fundamental trade-off between latency, reliability, and spectral efficiency, achieving extreme spectral efficiency for eMBB and ultra-reliability and low latency for URLLC is a complicated scheduling task. In other words, when adopting the puncturing scheme, the punctured eMBB data is completely lost, whereas for the superposition scheme, the eMBB receivers can still recover their data using one of the interference cancellation techniques. However, the interference that is resulting from the eMBB signals and experienced at the URLLC receivers, impacts the strict URLLC reliability, even when the conventional interference cancellation techniques are employed.

1.4 Motivations and Contributions

Despite the numerous studies on accommodating heterogeneous services, specially URLLC and eMBB, in the same network architecture, there are still shortcomings in spectrally efficient scheduling and multiplexing coexisting services. This is the primary momentum for exploring possible ways to schedule URLLC traffic while protecting coexisting eMBB users. In this section, motivations and thesis contributions are summarized.

1.4.1 Joint eMBB/URLLC Scheduling Through Superposition

As discussed earlier in this chapter, by employing the superposition scheme, the eMBB receivers can recover their data using interference cancellation techniques. Hence, in practice, the superposition scheme could be more spectrally efficient than the puncturing scheme. However, the fundamental disadvantage of superposition is that the interference from eMBB signals received at URLLC receivers has an impact on the stringent URLLC reliability. In this regard, few works have considered superposition for the coexistence problem in downlink wireless networks [25, 26].

Motivated by the apparent shortcomings, Chapter 2 studies the trade-off between the eMBB spectral efficiency and the URLLC QoS while jointly scheduling eMBB and URLLC traffic using superposition and puncturing. To achieve this goal, a resource allocation problem is formulated to minimize the rate loss of the eMBB service and URLLC packet segmentation loss while satisfying the eMBB and URLLC QoS constraints. Since the formulated problem is a mixed-integer non-linear program (MINLP), which is generally hard to solve directly, we proposed low complexity one-to-one and many-to-many pairing algorithms for accommodating the URLLC packets over the eMBB users. In terms of performance, simulation results show that the proposed algorithm achieves a higher URLLC packet admission rate and lower rate loss for eMBB compared to the baseline methods. Moreover, it is shown that at least 30% more URLLC users can be served without degrading their QoS while keeping the impact on the eMBB rate minimal.

1.4.2 RIS Assisted Wireless Networks

Although there has been extensive research regarding jointly scheduling the eMBB and URLLC traffic, there is a clear gap in facilitating the low-latency and high-reliability requirements of the emerging 6G applications. Specifically, the key performance indicators, i.e., reliability and latency requirements, of emerging URLLC applications are much stricter, visioned to be 99.99999% reliability, and 0.1-millisecond latency [27]. Hence, the allocation of URLLC traffic based on the 5G enabling technologies is insufficient to satisfy these extreme requirements. In this regard, it has been shown that RIS enhances the channel quality of the targeted users, which is reflected as improvement to the service reliability, resource efficiency, and capacity [28]. This motivates us to investigate the integration of RISs to assist the coexisting URLLC and eMBB traffic in wireless networks. Consequently, we have conducted a comprehensive study on the benefits, challenges and possible directions of integrating RIS in wireless networks to assist URLLC and eMBB traffic simultaneously [29]. To the best of our knowledge, this is the first work to consider RIS in the context of the coexistence problem.

In Chapter 3, two optimization problems are formulated: a time slot basis eMBB allocation problem and a mini-time slot basis URLLC allocation problem. The eMBB allocation problem aims at maximizing the eMBB sum rate by jointly optimizing the power allocation at the BS and the RIS phase-shift matrix while satisfying the eMBB rate constraint. On the other hand, the URLLC allocation problem is formulated as a multi-objective problem to maximize the URLLC admitted packets and minimize the eMBB rate loss. This objective is achieved by jointly optimizing the power and frequency allocations along with the RIS phase-shift matrix. In order to avoid the violation of the URLLC latency requirements, we propose a novel framework in which the RIS phase-shift matrix that enhances the URLLC reliability is proactively designed at the beginning of the time slot. The simulation results show that the proposed framework has a low time complexity, which makes it practical for real-time and efficient multiplexing between eMBB and URLLC traffic. In addition, using only 60 RIS elements, we observe that the proposed scheme achieves around 99.99% URLLC packets admission rate compared to 95.6% when there is no RIS while also achieving up to 70% enhancement on the eMBB sum rate.

In fact, integrating RIS in wireless networks comes with its challenges that may impact the URLLC performance. Specifically, optimization and configuring the RIS phase shift is computationally complex leading to violating the latency and reliability requirements of the URLLC traffic. Hence, Chapter 4 explores new ways to reduce the complexity of optimizing the RIS phase shift matrix. To achieve this, we formulate the problem of minimizing the total transmit power while jointly optimizing the allocated power to users and the RIS phase shifts. In line with the existing literature and with the aid of alternating optimization, the problem is decomposed into power control and RIS phase shift sub-problems. Then, this chapter proposes two solutions to tackle the problem of RIS passive beamforming optimization by leveraging a linear transformation and element-wise Karush-Kuhn-Tucker (KKT), respectively. The main idea of the linear transformation approach is to reduce the number of optimization variables to the number of users associated with the RIS from the number of its elements which is, in general, very large. On the other hand, the main idea of the element-wise KKT-based approach is to obtain a closed-form expression for the RIS phase shifts for a multi-user scenario, similar to the single-user case. Simulation results show that the proposed solutions have a competitive performance in terms of optimality and complexity for large-scale RIS. Hence, our approach represents a general framework for configuring RIS elements in real scenarios. To the best of our knowledge, this is the first work considering the proposed methodologies considering the RIS phase shift optimization.

1.4.3 Data Similarity Based Puncturing

One of the critical shortcomings of the current puncturing scheme is that it severely impacts the eMBB spectral efficiency [3, 30]. Precisely, when puncturing is adopted, the preempted resources are lost. Consequently, it slows the eMBB traffic as it requires more overhead, including a puncturing indicator (PI) to inform the eMBB user of the punctured resources. At the same time, the whole information block can be re-transmitted if decoding errors occur due to the lost resources. Thus, it is essential to explore a new puncturing mechanism, such as the lost eMBB resources being minimized while achieving the URLLC QoS requirements.

Motivated by this limitation of the current puncturing mechanism, in Chapter 5, we propose a novel downlink URLLC-eMBB multiplexing technique that exploits possible similarities among URLLC and eMBB symbols to reduce the size of the lost eMBB symbols. We suggest that the base station (BS) scans the eMBB traffic' symbol sequences and punctures those that have the highest symbol similarity with that of the URLLC users to be served. As the eMBB and URLLC may use different constellation sizes, we introduce the concept of symbol region similarity to accommodate the different constellations. We assess the performance of the proposed scheme analytically, where we derive closed-form expressions for the symbol error rate (SER) of the eMBB and URLLC services. Besides the outstanding performance of the proposed scheme, the proposed strategy is based on a simple search strategy making it an efficient solution to be used in practice. To the best of our knowledge, this work is the first to consider symbol similarity as an efficient puncturing scheme.

The notations used in the thesis have no connection to each another. As a result, some symbols may appear in multiple chapters and have different meanings.

1.5 List of Publications

This thesis has led to the following publications:

1.5.1 Journal Papers

(J1) M. Almekhlafi, M. Chraiti, A. Arfaoui, C. Assi, A. Ghrayeb and A. Alloum, "A Downlink Puncturing Scheme for Simultaneous Transmission of URLLC and eMBB Traffic by Exploiting Data Similarity," in IEEE Transactions on Vehicular Technology, Dec. 2021.

(J2) M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi and A. Ghrayeb, "Joint Resource Allocation and Phase Shift Optimization for RIS-Aided eMBB/URLLC Traffic Multiplexing," in IEEE Transactions on Communications, Feb. 2022.

(J3) M. Almekhlafi, M. A. Arfaoui, C. Assi and A. Ghrayeb, "Enabling URLLC Applications Through Reconfigurable Intelligent Surfaces: Challenges and Potential," IEEE Internet of Things Magazine, Mar. 2022.

(J4) M. Almekhlafi, M. A. Arfaoui, C. Assi and A. Ghrayeb, "Superposition-Based URLLC Traffic Scheduling in 5G and Beyond Wireless Networks," in IEEE Transactions on Communications, Sept. 2022.

(J5) M. Almekhlafi, M. A. Arfaoui, C. Assi and A. Ghrayeb, "A Low Complexity Passive Beamforming Design for Reconfigurable Intelligent Surface (RIS) in 6G Networks," in IEEE Transactions on Vehicular Technology (Accepted).

(J6) S. Khisa, M. Almekhlafi, M. Elhattab and C. Assi, "Full Duplex Cooperative Rate Splitting Multiple Access for a MISO Broadcast Channel with two Users," in IEEE Communications Letters, Aug. 2022.

1.5.2 Conference Papers

(C1) M. Almekhlafi, M. A. Arfaoui, C. Assi and A. Ghrayeb, "Joint Resource and Power Allocation for URLLC-eMBB Traffics Multiplexing in 6G Wireless Networks," in Proc. IEEE (ICC), June 2021.

(C2) M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi and A. Ghrayeb, "Joint Scheduling of eMBB and URLLC Services in RIS-Aided Downlink Cellular Networks," in Proc. IEEE (ICCCN), July 2021.

Chapter 2

Joint eMBB/URLLC Scheduling Using Superposition

2.1 Background, Related Works, and Contributions

As discussed in Chapter 1, 6G cellular networks are expected to support some URLLC applications that focus on simultaneously demanding massive connectivity and high data rates instead of sparse and short packet transmissions[23]. As a result, these high requirements complicate the co-existence of emerging URLLC services with their eMBB and mMTC counterparts. To this end, the enabling of URLLC services has received considerable research interest in the last few years [1, 3, 20, 33]. In fact, applications belonging to the URLLC services class, such as IoTs, autonomous driving and virtual reality, require an extremely low end-to-end latency that is less than one millisecond and an ultra high reliability that is less than 10^{-6} packet error rate [1, 34]. Consequently, the data traffic of such applications needs immediate scheduling and transmission upon arrival at the base stations (BSs), which translates into

The content of this chapter leads to one submitted IEEE journal and one published conference [31, 32] requirement of immediate availability of spectral resources. In line with the immediate scheduling, superposition/puncturing schemes have been proposed by the 3GPP standard. Accordingly, the arriving URLLC packets are immediately transmitted in the next mini-time slot over the ongoing eMBB resources once received at the transmitting APs. This approach helps to satisfy the required URLLC latency.

Based on the superposition/puncturing scheme, several methods focusing on scheduling URLLC service with the aim of maximizing the total average data rate of eMBB users have been proposed [3, 25, 35–40]. The authors of [3] studied the joint eMBB and URLLC scheduling problem, and they considered linear, convex and threshold models for the eMBB rate loss resulting from the superposition/puncturing scheme. A resource allocation policy for a puncturing-based scheduler was proposed in [37], where the formulated problem considered the overhead associated with the URLLC load segmentation while maximizing the rate utility. In [35], a risk-sensitive approach was introduced to alleviate the puncturing effects on the eMBB users with low data rates. In [36], a deep reinforcement learning approach was proposed to allocate the URLLC traffic. A null-space-based spatial puncturing scheduler for joint URLLC/eMBB traffic was proposed in [38]. Work in [39, 40] proposed matching-based scheduling schemes to allocate the URLLC traffic by adopting puncturing mechanism. The works in [35–40] considered only the puncturing scheme for joint scheduling of eMBB and URLLC loads. Authors in [25] have formulated a URLLC traffic allocation problem by adopting a superposition or puncturing scheme.

In this chapter, different from the aforementioned works, we investigate the performance of a superposition/puncturing allocation scheme in a downlink system that consists of a single BS serving multiple eMBB and URLLC users. In this system, the joint scheduling of URLLC and eMBB services is performed using either the superposition or the puncturing schemes. Particularly, our goal is to propose a mini-time slot-basis URLLC allocation scheme with low complexity. Our approach performs frequency and power allocation to jointly minimize the rate loss of all eMBB users and the URLLC packet segmentation loss. We consider both the eMBB and URLLC QoS constraints that should be guaranteed while admitting the URLLC packets. We formulate this objective as an optimization problem, which is a mixed-integer nonlinear program (MINLP) that is generally hard to solve.

To achieve our goal, we first consider the case of one-to-one pairing wherein one

URLLC packet can be paired with only one eMBB user. We reformulate the problem as a bi-level optimization problem that consists of one inner and one outer problem. The inner problem aims to find the optimal power and frequency resources for each URLLC and eMBB pair, while the outer problem aims to find the optimal eMBB-URLLC pairing (assignment) policy. In the inner problem, we derive the feasibility conditions in terms of the frequency and the power resources, the eMBB and URLLC QoS requirements, and the eMBB and URLLC channel state information. Then, we derive the optimal frequency and power allocation scheme in closed-form expressions. Accordingly, the outer problem is reduced to a simple assignment problem that can be optimally solved by using a greedy algorithm which has a polynomial time complexity. Then, we generalize the algorithm for many-to-many pairing while minimizing the overhead due to URLLC packet segmentation. Simulation results show that the proposed algorithm achieves lower loss in eMBB rates, lower loss in overhead and higher packet admission rate for URLLC service class than the aforementioned baselines. In addition, the simulation results show that the eMBB QoS requirement, defined by the threshold of eMBB rate loss, has a negative impact on the admission of the URLLC packets. In fact, when the threshold of eMBB rate loss is fixed, the maximum URLLC load that can be accommodated will be also fixed. Hence, the threshold of eMBB rate loss, the URLLC load and the URLLC reliability requirement should be jointly considered when optimizing the eMBB spectral efficiency.

The rest of the chapter is organized as follows. Section 2.2 presents the system model. Section 2.3 presents the problem formulation. Section 2.4 presents the proposed solution approach. Sections 2.5 and 2.6 present the simulation results and the conclusion, respectively.

2.2 System Model

2.2.1 System Settings

We consider a downlink radio access network (RAN) which consists of a single BS that has $B > 1$ resource blocks (RBs), each with a bandwidth W. As shown in Fig. 2.1, the BS serves simultaneously $E \geq 1$ eMBB users and $U \geq 1$ URLLC users. For all $e \in \{1, \ldots, E\}$ and $u \in \{1, \ldots, U\}$, h_e and g_u denote the downlink channel

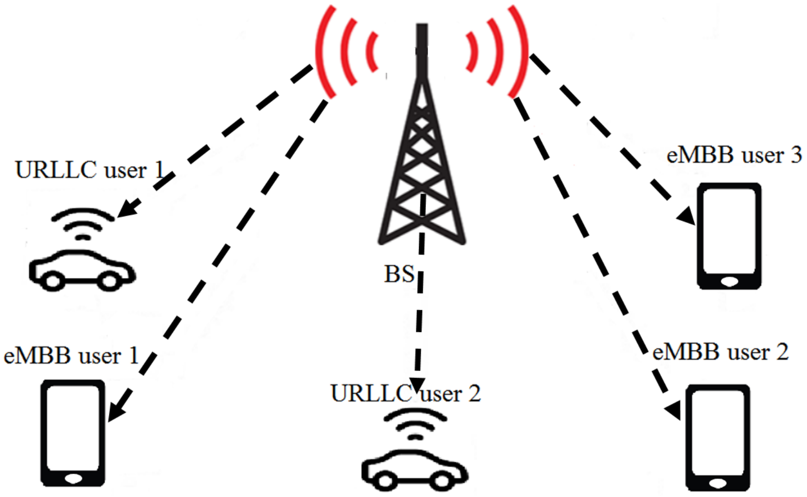


Fig. 2.1: System model

gains of the eth eMBB user and the uth URLLC user, respectively. Time is divided into slots. In addition, in order to support the latency requirement of the URLLC traffic, each time slot is further divided into N mini-time slots, where each mini-time slot has a duration δ [41]. Let $\mathcal{D}$ be a random variable whose distribution indicates the URLLC packet generation per URLLC user per mini-time slot; where $\mathbb{E}\{D\} = p$ is the average URLLC packet per user per mini-time slot. Accordingly, the average URLLC load is $p \times N \times U$ packet per eMBB time slot. Also, we assume the URLLC packet has a size of ζ information bits.

We assume that a frequency-division-multiple-access (FDMA) is employed as a multiple access technique for all eMBB users at the beginning of each time slot. Based on this, for all $e \in \{1, \ldots, E\}$, we denote by $\phi_e \leq B$ the number of frequency resources allocated to the eth eMBB user. On the other hand, the arriving URLLC traffic at each mini-time slot is immediately multiplexed with the eMBB traffic and then transmitted at the next time mini-time slot using the superposition/puncturing scheme. The superimposed resources of the eMBB are allocated based on power domain, i.e., the power is divided between the eMBB and the URLLC users that are sharing the same resources. In order to guarantee the required reliability and latency for the URLLC traffic, it is assumed that the BS allocates more power to the URLLC user. In fact, allocating more power to the URLLC traffic guarantees lower bit-error-rate (BER), and therefore, higher reliability for users in this service class. Now, since more power is allocated to the URLLC traffic, the SIC procedure is no longer needed, which will cancel the SIC processing delay at the URLLC users. Noting that we will use superposition for both superposition and puncturing as puncturing is a special case of superposition when the eMBB power allocation factor is zero.

2.2.2 Signal Model

We assume that the BS assigns its resources at the beginning of each time slot to the eMBB users using orthogonal resources. Accordingly, for all $e \in \{1, \ldots, E\}$, the achievable rate, in bits/s/Hz, of the eth eMBB user is expressed as

$$r_e = \log_2 \left(1 + \gamma_e\right), \quad bit/sec/Hz \tag{2.1}$$

where $\gamma_e = \frac{P_e |h_e|^2}{\sigma_{0,e}^2}$, in which P_e and $\sigma_{0,e}$ are the transmitted power and the AWGN noise level of the eMBB receiver, respectively. In the scenarios where URLLC and eMBB traffic coexist, the signal of each eMBB user may be superposed with or punctured by the signal of a URLLC user. Let l^n be the number of transmitted URLLC packets at mini-time slot n. These l^n packets belong to the URLLC user set $\{1, \ldots, U\}$. Then, g_l^n represents the channel gain of the lth URLLC packet at mini-time slot n, where $g_l^n \in \{g_1, g_2, \ldots, g_U\}$. Practically, the URLLC packet l can be segmented over multiple eMBB users. Then, for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$, the resulting signal from superposing the data of the eth eMBB user and the signal of lth URLLC packet (or a segment of the lth URLLC packet) at mini-time slot n is expressed as [42]

$$x_{e,l}^n = \sqrt{P_e} \left(\sqrt{\alpha_{e,l}^n} x_l^n + \sqrt{1 - \alpha_{e,l}^n} x_e \right), \tag{2.2}$$

where x_e is the signal of the eth eMBB user, x_l^n is the signal of the lth URLLC packet (or a segment of the lth URLLC packet) at mini-time slot n, $\alpha_{e,l}^n \in [0, 1]$ is the power factor allocated for URLLC packet l at mini-time slot n. We assume that the URLLC receiver does not perform SIC. Accordingly, the power allocation factor is assumed to satisfy $0.5 < \alpha_{e,l}^n \leq 1$ for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$, such that the eth eMBB user can adopt SIC to cancel the interference resulting from the coexistence of URLLC packets and decodes its interference-free signal with a rate of [42]:

$$r_{e,l}^{n}(\alpha_{e,l}^{n}) = \log_2\left(1 + (1-\alpha_{e,l}^{n})\gamma_e\right). \tag{2.3}$$

From (2.2) and (2.3), the average achievable data rate of the e-th eMBB user over one frame (e.g., time slot) can be expressed as[1]

$$R_e = W\left[\frac{1}{N}\sum_{n=1}^{N}\left((\phi_e - \sum_{l=1}^{l^n}\varphi_{e,l}^{n})r_e + \sum_{l=1}^{l^n}\varphi_{e,l}^{n} r_{e,l}^{n}(\alpha_{e,l}^{n})\right)\right], \tag{2.4}$$

where $\varphi_{e,l}^{n} \in \{0,\ldots,\phi_e\}$ denotes the number of frequency resources (blocks) extracted from all frequency resources of the eth eMBB user and allocated to the URLLC packet l, where $0 \leq \sum_{e}^{E}\varphi_{e,l}^{n} \leq B$. In (2.4), the average rate of the e-th eMBB user is expressed in terms of frequency and power resources superimposed by the URLLC traffic. According to (2.4), the achievable data rate of the e-th eMBB user at mini-time slot n can be expressed as:

$$R_e^n = W\left((\phi_e - \sum_{l=1}^{l^n}\varphi_{e,l}^{n})r_e + \sum_{l=1}^{l^n}\varphi_{e,l}^{n} r_{e,l}^{n}(\alpha_{e,l}^{n})\right). \tag{2.5}$$

On the other hand, the URLLC user decodes its signal directly without performing SIC by treating the eMBB signal as noise [42]. Since all URLLC packet segments have small sizes in general, the Shannon capacity is not longer accurate [43]. Hence, the achievable rate of the lth URLLC packet (or segment) can be accurately evaluated through the finite block length regime, and therefore, it can be expressed as [43, 44]

$$\begin{aligned} C_{e,l}^{n}(\varphi_{e,l}^{n}, \alpha_{e,l}^{n}) = \log_2&\left(1 + \frac{\alpha_{e,l}^{n}\gamma_l^n}{(1-\alpha_{e,l}^{n})\gamma_l^n + 1}\right) \\ &-\frac{1}{\ln(2)}\sqrt{\frac{V}{\delta\varphi_{e,l}^{n}W}}Q^{-1}(\epsilon_u), \end{aligned} \quad [\text{bit/s/Hz}], \tag{2.6}$$

[1] According to the 3GPP superposition/puncturing framework, the BS sends an indicator signal to inform the eMBB users of the punctured resources. In this context, we adopted the linear loss function to represent the eMBB rate loss associated with the superposition/puncturing process.

where $\gamma_l^n = \frac{P_e|g_l^n|^2}{\sigma_{0,l}^2}$, in which $\sigma_{0,l}$ is the AWGN noise level of the URLLC receiver, ϵ_u denotes the target URLLC block error rate which is very small for the URLLC traffic, i.e., 1^{-6}. The term $V = 1 - \frac{1}{(1+\frac{\alpha_{e,l}^n \gamma_l^n}{(1-\alpha_{e,l}^n)\gamma_l^n+1})^2}$ denotes the URLLC channel dispersion.

2.3 Problem Formulation

2.3.1 Objective

This work seeks to minimize the rate loss (of all eMBB users, their sum) and the sum of segmentation losses of all URLLC packets while satisfying a certain latency and reliability requirements for all URLLC users. Accordingly, at mini-time slot n, the data rate loss of the eth eMBB user superimposed by the URLLC packet l over shared RBs denoted by $(\varphi_{e,l}^n)$ can be expressed as

$$\widehat{R}_{e,l}^n(\varphi_{e,l}^n, \alpha_{e,l}^n) = W\Bigg[r_e - r_{e,l}^n(\alpha_{e,l}^n)\Bigg]\varphi_{e,l}^n. \tag{2.7}$$

The URLLC payload may be segmented and distributed over the frequency resources of several eMBB users within the same mini-time slot, resulting in a costly signaling overhead. Practically, the signaling can either occur in-band, which directly affects the eMBB rate, or out of band, which may render the utility of the control channel less effective [37]. Therefore, avoiding unnecessary segmentation of the URLLC traffic over multiple eMBB users ought be minimized. Accordingly, the segmentation rate loss can be expressed as

$$\widehat{R}_l^n(\varphi_{e,l}^n) = \widehat{R}_o \sum_{e=1}^{E} \min(\varphi_{e,l}^n, 1), \tag{2.8}$$

where $\widehat{R}_o$ is a fixed rate loss which abstracts the signaling overhead per URLLC packet segment. Accordingly, the objective of minimizing the eMBB rate loss and URLLC packet segmentation loss is formulated as

$$\min_{\boldsymbol{\varphi},\boldsymbol{\alpha}} \sum_{e=1}^{E}\sum_{l=1}^{l^n} \widehat{R}_{e,l}^n(\varphi_{e,l}^n, \alpha_{e,l}^n) + \sum_{l=1}^{l^n} \widehat{R}_l^n(\varphi_{e,l}^n), \tag{2.9}$$

where the vectors $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are the resource allocation and the power allocation decision variables.

2.3.2 eMBB QoS

The superposition of eMBB and URLLC data impacts not only the rate of each involved eMBB user but also its reliability decoding capability. Hence, we consider a rate loss threshold (maximum eMBB rate loss) such that the eMBB rate constraint is satisfied. Then, the eMBB rate loss constraint can be derived as

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{l^n}\widehat{R}_{e,l}^{n}(\varphi_{e,l}^{n},\alpha_{e,l}^{n}) \leq \widehat{R}_e^{th}, \tag{2.10}$$

where $\widehat{R}_e^{th}$ depends on the targeted data rate of the eth eMBB user. Equation (2.10) indicates that the eMBB rate loss should not exceed $\widehat{R}_e^{th}$. In fact, $\widehat{R}_e^{th}$ depends on both the eMBB QoS requirement and the ultimate objective of the eMBB allocation problem. In this work, the eMBB allocation problem is formulated to maximize the eMBB rate utility while guaranteeing fairness between the eMBB users [3], that is,

$$\mathcal{P}_{eMBB} : \max_{\boldsymbol{\phi}} \sum_{e}^{E} \log((1-\Delta)\phi_e W r_e) \tag{2.11a}$$

$$\text{s.t.} \ \sum_{e}^{E} \phi_e \leq B, \tag{2.11b}$$

$$\phi_e \in \mathbb{Z}^{+}. \tag{2.11c}$$

Here, Δ is a predefined sharing factor that represents the amount of shared eMBB resources to accommodate the URLLC traffic [3]. In practice, the sharing factor Δ is obtained at the beginning of each time slot from the arrival rate of the URLLC traffic that is predicted at the beginning of each eMBB time slot based on some online measurements at the BS [45]. Accordingly, the eMBB rate threshold can be defined as $\widehat{R}_e^{th} = W\Delta\phi_e r_e$. Problem (2.11) is an integer non-linear problem, and hence, it is an NP-hard problem. It can be solved using one of the over-the-shelf solvers such as such as MOSEK [46]; however, the associated complexity is exponential. To overcome this issue, we relax the integer variable ϕ to be continuous. Then, problem

(2.11) becomes convex whose solution can be derived by applying the KKT conditions [35]. The obtained solution of the relaxed convex problem is then rounded to get a near-optimal solution for the original integer problem [35].

2.3.3 URLLC QoS

The URLLC traffic requires high reliability and it is subjected to latency constraints which should be satisfied. Actually, several components, including queuing delay and transmission time, can impact the URLLC latency. The queuing delay can be eliminated by transmitting the URLLC packets immediately upon arrivals, i.e., in the next mini-time slot. On the other hand, the transmission delay can be guaranteed by controlling the transmission rate of URLLC packets. In fact, a URLLC packet should be transmitted within one mini-time slot, otherwise the URLLC packet will be dropped. Hence, the reliability of a URLLC packet can be expressed as [26, 47]

$$\sum_{e}^{E} W\delta\varphi_{e,l}^{n} C_{e,l}^{n}(\varphi_{e,l}^{n}, \alpha_{e,l}^{n}) \geq \zeta. \tag{2.12}$$

Based on the URLLC achievable rate expression in (2.12), the reliability requirement of each URLLC packet can be satisfied by guaranteeing a certain minimum achievable rate to transmit the whole URLLC packet of ζ bits. Hence, if constraint (2.12) is not satisfied for a URLLC packet, then it will be dropped. Let $\hat{l}^n$ denote the number of dropped URLLC packets at mini-time slot n. Then, the URLLC packet admission rate, denoted by η, is defined as the ratio between the admitted URLLC packets and the total packets, i.e., [48, 49]

$$\eta = \frac{\sum_{1}^{N} l^n - \sum_{1}^{N} \widehat{l}^n}{\sum_{1}^{N} l^n}. \tag{2.13}$$

2.3.4 Problem Formulation

Based on the above analysis, and at each mini-time slot belonging to slot $n \geq 2$, the final optimization problem of the URLLC scheduler is formulated as

$$\mathcal{P}: \quad \min_{\boldsymbol{\varphi},\boldsymbol{\alpha}} \sum_{e=1}^{E} \sum_{l=1}^{l^n} \widehat{R}_{e,l}^{n}(\varphi_{e,l}^{n}, \alpha_{e,l}^{n}) + \sum_{l=1}^{l^n} \widehat{R}_{l}^{n}(\varphi_{e,l}^{n}), \tag{2.14a}$$

$$\text{s.t.} \sum_{e}^{E} (W\delta)\, \varphi_{e,l}^{n} C_{e,l}^{n}(\varphi_{e,l}^{n}, \alpha_{e,l}^{n}) \geq \zeta, \ \forall l \in \{1, \ldots, l^n\}, \tag{2.14b}$$

$$\sum_{l=1}^{l^n} \widehat{R}_{e,l}^{n}(\varphi_{e,l}^{n}, \alpha_{e,l}^{n}) \leq N\, \widehat{R}_{e}^{th} - \sum_{l=1}^{l^n} \sum_{i=1}^{n-1} \widehat{R}_{e,l}^{i}, \forall e \in \{1, \ldots, E\}, \tag{2.14c}$$

$$\frac{\text{sign}(\varphi_{e,l}^{n})}{2} \leq \alpha_{e,l}^{n} \leq \text{sign}(\varphi_{e,l}^{n}), \forall e \in \{1, \ldots, E\}, \forall l \in \{1, \ldots, l^n\}, \tag{2.14d}$$

$$0 \leq \sum_{l=1}^{l^n} \varphi_{e,l}^{n} \leq \phi_e, \ \forall e \in \{1, \ldots, E\}, \forall l \in \{1, \ldots, l^n\}, \tag{2.14e}$$

$$\varphi_{e,l}^{n} \in \mathbb{Z}^{+}\ , \forall e \in \{1, \ldots, E\}, \forall l \in \{1, \ldots, l^n\}. \tag{2.14f}$$

Problem (2.14) seeks both the optimum resource allocation matrix $\boldsymbol{\varphi}$ and the vector of power allocation fractions $\boldsymbol{\alpha}$ that minimize the sum of losses of the eMBB rates (2.14a). Constraint (2.14b) ensures the URLLC packets reliability while constraint (2.14c) ensures the QoS of each eMBB user. In fact, for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$, if a segment of the lth URLLC packet is superposed over the data of the eth eMBB user, then $\varphi_{e,l}^{n} > 0$, whereas in the opposite case, $\varphi_{e,l}^{n} = 0$. Constraints (2.14d) set the bounds of the power allocation decision variables to be between $[0.5, 1]$ if $\text{sign}(\varphi_{e,l}^{n}) = 1$, and 0 otherwise. Constraints (2.14e) sets the bound on the punctured frequency resources of each eMBB users while (2.14f) indicates that the URLLC packet allocated integer RBs.

Problem (2.14) is a mixed integer non-linear problem (MINLP), which is generally hard to solve. However, understanding the relations between the decision variables $\varphi_{e,l}^{n}$ and $\alpha_{e,l}^{n}$, for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$ can simplify the optimization problem $\mathcal{P}$. In fact, for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$, the decision variables $\varphi_{e,l}^{n}$ $\alpha_{e,l}^{n}$ represent the required resource allocation for superposing the data of the eth eMBB user and the lth URLLC packet. Moreover, $\text{sign}(\varphi_{e,l}^{n})$ indicates the pairing between the two. Precisely, for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$, if $\text{sign}(\varphi_{e,l}^{n}) = 1$, then the lth URLLC packet and the eth eMBB user are paired, and not paired if $\text{sign}(\varphi_{e,l}^n) = 0$. Moreover, for all $e \in \{1, \ldots, E\}$ and $l \in \{1, \ldots, l^n\}$, we can observe that the decision variables $\varphi_{e,l}^n$ and $\alpha_{e,l}^n$ are strongly coupled, since increasing $\alpha_{e,l}^n$ decreases the required number of frequency resources of the lth URLLC packet $\varphi_{e,l}^n$ and the opposite is true. Consequently, if the optimal ${\alpha_{e,l}^n}^*$ is obtained, then the optimal ${\varphi_{e,l}^n}^*$ can be obtained. In such a case, problem $\mathcal{P}$ can be seen as an assignment problem between eMBB users and URLLC packets after evaluating the optimal power and resource allocation problem for each possible pair of eMBB user and URLLC packet (or segment).

2.4 Solution Approach

In this section, we present the proposed solution approach for problem $\mathcal{P}$ in (2.14). First, we investigate and solve problem $\mathcal{P}$ for the special case of one-to-one pairing, in which each URLLC packet is forced to be paired with at most one eMBB user, i.e., no segmentation for the URLLC packets. Then, we discuss the limitations and the impacts of the one-to-one pairing on the URLLC packet admission rate. Finally, based on the obtained results and discussions, we extend the obtained solution to the case of segmentation of the URLLC packet over several eMBB traffic streams, hence, each URLLC segment is paired with at most one eMBB user, which is indeed equivalent to problem $\mathcal{P}$.

2.4.1 One-to-One Pairing

In this subsection, we reformulate problem $\mathcal{P}$ in (2.14) by assuming that each URLLC packet can be superposed to the data of only one eMBB user, i.e., there is no segmentation of the URLLC packets, and that each eMBB user can be only paired with at most one URLLC packet. In this case, problem $\mathcal{P}$ is re-written as

$$\mathcal{P}_1 : \quad \min_{\boldsymbol{\varphi}, \boldsymbol{\alpha}} \sum_{e=1}^{E} \sum_{l=1}^{l^n} \widehat{R}_{e,l}^n(\varphi_{e,l}^n, \alpha_{e,l}^n), \tag{2.15a}$$

$$\text{s.t.} \quad (2.14\text{b}) - (2.14\text{f}), \tag{2.15b}$$

$$\sum_{l=1}^{l^n} sign(\varphi_{e,l}^n) \leq 1, \ \forall e \in \{1, \dots, E\}, \tag{2.15c}$$

$$\sum_{e=1}^{E} sign(\varphi_{e,l}^n) = 1, \ \forall l \in \{1, \dots, l^n\}. \tag{2.15d}$$

Problem $\mathcal{P}_1$ has a nice property which can be exploited to efficiently solve $\mathcal{P}_1$. This property is based on the fact that the *optimal* resource (power and frequency) allocation policy is independent of the pairing policy $\text{sign}(\varphi_{e,l}^n)$. In other words, if (e, l) are paired, i.e., $\text{sign}(\varphi_{e,l}^n) = 1$, then the obtained optimal resource allocations $\varphi_{e,l}^{n^*}$ and $\alpha_{e,l}^{n^*}$ are determined, and if they are not paired, then $\varphi_{e,l}^{n^*} = 0$. Precisely, let $\left\{ \left(\varphi_{e,l}^{n^*}, \alpha_{e,l}^{n^*}\right) | \, e \in \{1, \dots, E\}, l \in \{1, \dots, l^n\} \right\}$ denote the set of optimal user power allocation policy and resource allocation policy, which are the solutions to problem $\mathcal{P}_1$. For all $e \in \{1, \dots, E\}$ and $l \in \{1, \dots, l^n\}$, if $\text{sign}(\varphi_{e,l}^n) = 0$, then $\left(\varphi_{e,l}^{n^*}, \alpha_{e,l}^{n^*}\right) = (0, 0)$. However, if $\text{sign}(\varphi_{e,l}^n) = 1$, then $\left(\varphi_{e,l}^{n^*}, \alpha_{e,l}^{n^*}\right)$ should be the optimal solutions of the resource and power allocation policies of the pair (e, l). In other words, for all $e \in \{1, \dots, E\}$ and $l \in \{1, \dots, l^n\}$ if we assume that the eth eMBB user and the lth URLLC packet are paired together and that we can obtain their optimal resource and power coefficients $\left(\varphi_{e,l}^{n^*}, \alpha_{e,l}^{n^*}\right)$, then problem $\mathcal{P}_1$ becomes a linear assignment problem and it remains to determine the optimal pairing policy. Accordingly, we decompose problem $\mathcal{P}_1$ into two sub-problems. The first is a resource and power allocation problem that minimizes the eMBB rate loss for eMBB-URLLC pairs, whereas the second is a linear assignment problem that minimizes the total eMBB rate loss.

1) Resource allocation problem:

Here, our objective is to obtain, for every (e, l) pair, the optimal resource and power allocation coefficients $\left(\alpha_{e,l}^{n^*}, \varphi_{e,l}^{n^*}\right)$; that is, for each possible pairing of eMBB-URLLC pair (e, l), we aim to minimize the eMBB rate loss. This problem is expressed:

$$\mathcal{P}_{e,l}: \quad \min_{\varphi_{e,l}^n, \alpha_{e,l}^n} \widehat{R}_{e,l}^n(\varphi_{e,l}^n, \alpha_{e,l}^n) \tag{2.16a}$$

$$\text{s.t.} \quad W\delta\varphi_{e,l}^n C_{e,l}^n \geq \zeta, \tag{2.16b}$$

$$\widehat{R}_{e,l}^n(\varphi_{e,l}^n, \alpha_{e,l}^n) \leq N\widehat{R}_e^{th} - \sum_{i=1}^{n-1} \widehat{R}_{e,l}^i, \tag{2.16c}$$

$$0.5 < \alpha_{e,l}^n \leq 1, \tag{2.16d}$$

$$0 \leq \varphi_{e,l}^{n} \leq \phi_e, \tag{2.16e}$$

$$\varphi_{e,l}^{n} \in \mathbb{Z}^{+}. \tag{2.16f}$$

$\mathcal{P}_{e,l}$ is solved for all pairs of eMBB and URLLC users; we however exploit properties of this problem in order to obtain closed form expressions for the optimal solutions. Now, before deriving the optimal solutions, we investigate its feasibility conditions, which are defined in the following theorem.

Theorem 2.1. *Problem $\mathcal{P}_{e,l}$ is feasible if and only if the following conditions hold:*

- *Condition 1:*

$$A_{e,l}^{n}({\varphi_{e,l}^{n}}^{\max}) \leq 1. \tag{2.17}$$

- *Condition 2:*

$$\max(0, G_{e,l}^{n}(1)) \leq \min(\phi_e, \beta_{e,l}^{n}(0.5)), \tag{2.18}$$

where $A_{e,l}^{n}(.)$, $G_{e,l}^{n}(\cdot)$ and $\beta_{e,l}^{n}(\cdot)$, respectively, are

$$A_{e,l}^{n}(\varphi_{e,l}^{n}) = \max\left(0.5,\ \frac{(\gamma_l^n + 1) \times \gamma_1^{th}(\varphi_{e,l}^{n})}{(\gamma_1^{th}(\varphi_{e,l}^{n}) + 1) \times \gamma_l^n},\right), \tag{2.19}$$

$$\begin{aligned} G_{e,l}^{n}\left(\alpha_{e,l}^{n}\right) =& \frac{1}{4}\Bigg(\frac{Q^{-1}(\epsilon_u)}{\ln(2)}\sqrt{\frac{V}{\delta W}}\frac{1}{\log_2(1+\frac{\alpha_{e,l}^{n}\gamma_l^n}{(1-\alpha_{e,l}^{n})+1\gamma_l^n})}+ \\ & \sqrt{(\frac{Q^{-1}(\epsilon_u)}{\ln(2)})^2\frac{V}{\delta W}\frac{1}{\log_2(1+\frac{\alpha_{e,l}^{n}\gamma_l^n}{(1-\alpha_{e,l}^{n})+1\gamma_l^n})^2}+\frac{4\zeta}{\delta W\log_2(1+\frac{\alpha_{e,l}^{n}\gamma_l^n}{(1-\alpha_{e,l}^{n})+1\gamma_l^n})}}\Bigg)^2, \end{aligned} \tag{2.20}$$

$$\beta_{e,l}^{n}\left(\alpha_{e,l}^{n}\right) = \frac{N\widehat{R}_e^{th}(\varphi_{e,l}^{n}, \alpha_{e,l}^{n}) - \sum_{i=1}^{n-1}\widehat{R}_{e,l}^{i}}{W\widehat{R}_{e,l}^{n}}, \tag{2.21}$$

$$\gamma_1^{th}(\varphi_{e,l}^n) = 2^{\frac{\zeta}{W\delta\varphi_{e,l}^n} + \frac{Q^{-1}}{log(2)}\sqrt{\frac{V}{W\delta\varphi_{e,l}^n}}} - 1. \tag{2.22}$$

Proof. First, the channel dispersion $V \approx 1 - \frac{1}{1+\frac{\alpha_{e,l}^n}{1-\alpha_{e,l}^n}}$. In addition, since $0.5 \leq \alpha_{e,l}^n \leq 1$, then $0.5 \leq V \leq 1$. Hence, as an approximation, we will assume that V is a constant that does not depend on $\alpha_{e,l}^n$. The proof of Theorem 2.1 can be easily derived by observing the bounds of the variables $\varphi_{e,l}^n$ and $\alpha_{e,l}^n$ as shown in Fig. 2.2. Let us first investigate constraints (2.16c) and (2.16e), which define together the upper bound of the variable $\varphi_{e,l}^n$. Constraints (2.16c) can be equivalently transformed into the following inequality:

$$\beta_{e,l}^n(\alpha_{e,l}^n) = \varphi_{e,l}^n \leq \frac{\widehat{R}_e^{th} - \sum_{i=1}^{n-1}\widehat{R}_{e,l}^i}{W(r_e - r_{e,l}^n(\alpha_{e,l}^n))}. \tag{2.23}$$

By substituting $\alpha_{e,l}^n = 0.5$, we obtain the upper bound of $\varphi_{e,l}^n$ which is ${\varphi_{e,l}^n}^{max} = \min(\phi_e, \beta_{e,l}^n(0.5))$. After defining the maximum frequency resources that can be superposed, we can obtain the minimum feasible $\alpha_{e,l}^n$ that satisfies URLLC rate constraints in (2.16b). By substituting $\varphi_{e,l}^n = {\varphi_{e,l}^n}^{max}$ one can derive the feasibility on (2.17) for $\alpha_{e,l}^n$. Moreover, constraints (2.16b) define the lower bound of $\varphi_{e,l}^n$, which should be between $[0, \phi_e]$. Hence, if we substitute $y = \sqrt{\varphi_{e,l}^n}$ in (2.16b), we obtain $\varphi_{e,l}^n \geq G_{e,l}^n$. Solving (2.20) at the intersection point $\alpha_{e,l}^n = 1$, we obtain the lower bound on the feasible $\varphi_{e,l}^n$. Now, in order for problem $P_{e,l}$ to be feasible, the feasibility region should be non-empty, i.e., $\varphi_{e,l}^n$ should be greater than ${\varphi_{e,l}^n}^{min} = \max(G_{e,l}^n(1), 0)$, and it should be less than ${\varphi_{e,l}^n}^{max} = \min(\beta_{e,l}^n(0.5), \varphi_e)$. This completes the proof. □

Now, assuming that $P_{e,l}$ is feasible, the optimal resource and power allocation coefficients $\left(\alpha_{e,l}^{n^*}, \varphi_{e,l}^{n^*}\right)$ are presented in the following theorem.

Theorem 2.2. *Let* $\Phi_{e,l}^n \triangleq \{{\varphi_{e,l}^n}^{min}, {\varphi_{e,l}^n}^{max}\}$ *be the integer feasibility region of the resource allocation variable* $\varphi_{e,l}^n$*. Hence, the optimal power allocation coefficient* $\varphi_{e,l}^{n^*}$

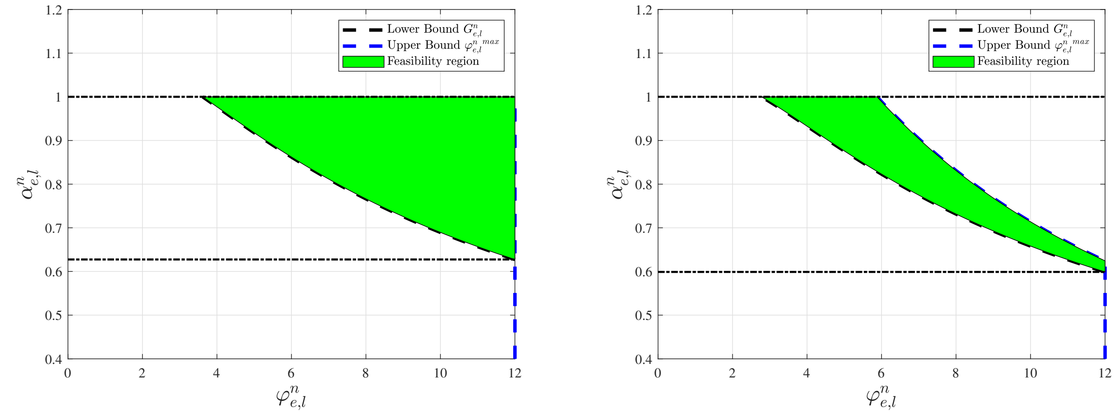


Fig. 2.2: Examples for the feasibility region of the resource allocation problem.

is expressed as

$$\varphi^{n^*}_{e,l} = \arg \min_{\varphi^n_{e,l} \in \Phi^n_{e,l}} A^n_{e,l}(\varphi^n_{e,l}), \tag{2.24}$$

such that

$$A^n_{e,l}(\varphi^n_{e,l}) \leq 1 - \frac{1 + \sqrt{(1+\gamma_e)((1 - A^n_{e,l}(\varphi^n_{e,l} - 1))\gamma_e + 1)}}{\gamma_e}, \tag{2.25}$$

and the optimal power allocation coefficient is given by $\alpha^{n^*}_{e,l} = A^n_{e,l}(\varphi^{n^*}_{e,l})$.

Proof. Let us consider a random number of frequency resources $\varphi \in \{2, \ldots, \phi_e\}$. In addition, let $\alpha_1 = A^n_{e,l}(\varphi - 1)$ and $\alpha_2 = A^n_{e,l}(\varphi)$ be the minimum power allocation factors associated to $\varphi - 1$ and φ_1, respectively (see Fig. 2.2). Now, let us evaluate the eMBB rate loss at $(\varphi - 1, \alpha_1^*)$ and (φ, α_2^*). One can easily verify that the function $\varphi \mapsto A^n_{e,l}(\varphi)$ is a decreasing function. Therefore, we get $\alpha_2 < \alpha_1$. Afterwards, the corresponding eMBB rate losses associated to $(\varphi - 1, \alpha_1)$ and (φ, α_2) are given , respectively, as

$$\begin{aligned} \widehat{R}^1_e &= W\,(\varphi - 1) \log_2 \left(\frac{1 + \gamma_e}{(1 - \alpha_1)\gamma_e + 1} \right), \\ \widehat{R}^2_e &= W \varphi \log_2 \left(\frac{1 + \gamma_e}{(1 - \alpha_2)\gamma_e + 1} \right). \end{aligned} \tag{2.26}$$

Then, the pair (φ, α_2) is an optimal solution over the set $\{(\varphi - 1, \alpha_1), (\varphi, \alpha_2)\}$, if and

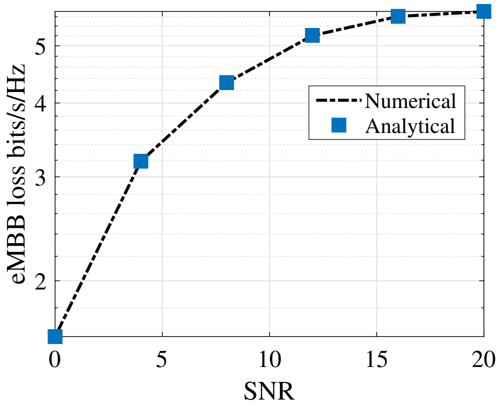


Fig. 2.3: Analytical and numerical average eMBB rate loss versus SNR.

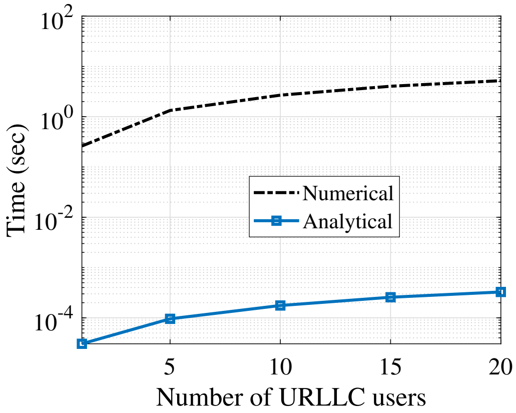


Fig. 2.4: Analytical and numerical time complexity.

only if $\widehat{R}_e^1 \geq \widehat{R}_e^2$. By solving the latter inequality, we obtain

$$A_{e,l}^n(\varphi) < 1 + \frac{1 - \sqrt{(1+\gamma_e)((1 - A_{e,l}^n(\varphi - 1))\gamma_e + 1)}}{\gamma_e}. \tag{2.27}$$

Therefore, we conclude that the optimal resource allocation ${\varphi_{e,l}^n}^*$ is the resource allocation φ that minimizes $A_{e,l}^n(\varphi)$ and satisfies (2.27), which completes the proof. □

Fig. 2.3 illustrates the average analytical and numerical eMBB rate loss versus the transmit SNR from the BS. The analytical results are obtained through the closed-form expression derived in Theorem 2.2. Moreover, the numerical results are obtained by solving problem $\mathcal{P}_{e,l}$ using an off-the-shelf optimization solver.[2] This figure shows that the analytical results match perfectly the numerical results, which validate the optimality of the expressions of the resource and power allocation coefficients derived in Theorem 2.2. On the other hand, Fig. 2.4 presents the computational time of the derived closed-form expressions and of the numerical solutions versus the number of URLLC users. This figure shows that obtaining the numerical solutions is extremely time-consuming, which violates the low-latency constraint of the URLLC traffic. However, the proposed closed-form expressions have a processing time in the order of sub-millisecond, which is suitable for the URLLC traffic.

2) The eMBB-URLLC Assignment Problem:

[2] The used solver is the genetic algorithm, which is a predefined Matlab solver [50].

Using the results of the optimal resource and allocation derived in the previous part (optimal pair $\alpha_{e,l}^{n^*}$ and $\varphi_{e,l}^{n^*}$ for every possible pairing), problem $\mathcal{P}_1$ is reduced to a simple linear assignment problem. Let $\mathbf{I} = \{I_{1,1}, I_{1,2}, ..., I_{E,l^n}\}$ represent the pairing vector, where $I_{e,l} \overset{\text{def}}{=} \text{sign}(\varphi_{e,l})$. Hence, the purpose of this part is to find the optimal pairing policy $\mathbf{I}^*$ that solves problem $\mathcal{P}_1$, i.e., that minimizes the total eMBB rate loss. Hence, problem $\mathcal{P}_1$ is rewritten as

$$\mathcal{P}_1^{outer} : \min_{\mathbf{I}} \sum_{e=1}^{E} \sum_{l=1}^{l^n} I_{e,l}^n \times \left(\widehat{R}_{e,l}^n \left(\alpha_{e,l}^{n^*}, \varphi_{e,l}^{n^*} \right) + \widehat{R}_o^n \right) \tag{2.28a}$$

$$\text{s.t.} \quad \sum_{l}^{l^n} I_{e,l}^n \leq 1, \quad \forall e \in \{1, \ldots, E\}, \tag{2.28b}$$

$$\sum_{e=1}^{E} I_{e,l}^n = 1, \quad \forall l \in \{1, \ldots, l^n\}, \tag{2.28c}$$

$$I_{e,l}^n \in \{0, 1\}, \quad \forall e \in \{1, \ldots, E\}, \forall l \in \{1, \ldots, l^n\}. \tag{2.28d}$$

Problem $\mathcal{P}_1^{outer}$ is a linear assignment problem which can be easily solved using an off-the-shelf optimization solver or the Hungarian method. However, the Hungarian method gives the optimal solution in a polynomial time complexity, the computational time for large number URLLC users is in order of milliseconds which may violate the URLLC latency requirements [51].

3) Limitations:

Several limitations can impact the performance of the URLLC service while considering the above one-to-one pairing policy. These limitations are summarized as follows:

1. A limited number of eMBB users makes the one-to-one pairing scheme insufficient for the case where the URLLC packets are more than the available eMBB transmissions. For example, if $l^n = 3$ and the available eMBB users is $E = 2$, then only two URLLC packets can be served.

2. If the QoS of the eMBB user is strict, i.e., small rate loss threshold R_e^{th}, the pairing possibility decreases, which affects the service of the URLLC packets. Although the URLLC packets segmentation will enhance the admission of URLLC packets, the segmentation loss will dramatically increase. However, when the

Algorithm 2.1: One-to-one pairing Algorithm.

```
- Sort eMBB users based on the channel gain in ascending order. ;
- Sort URLLC users based on the channel gain in descending order.;
for $l \in \{1, \dots, l^n\}$ do
    Boolean=0 allocating indicator variable;
    for $e = 1 \rightarrow E$ do
        if feasibility conditions in (2.17) and (2.18) then
            allocate URLLC packet $l$ on the eMBB user $e$;
            Update $\widehat{R}_e^{th}$ and $\phi_e$;
            Boolean=1;
            break;
        end
    end
    if Boolean==0 then
        break;
    end
end
```

URLLC packet is divided into several segments over multiple eMBB transmissions. For example, consider the case when the minimum number of resources needed to serve one URLLC packet is four frequency resources while each eMBB user can afford only two frequency resources, then the URLLC packet is segmented over two eMBB transmissions. Accordingly, higher packet admission is achieved at the expense of the control signalling (loss) due to segmentation.

4) Proposed One-to-One Pairing Algorithm:

The outer problem $\mathcal{P}_1^{outer}$ is a one-to-one pairing problem, which means that the number of URLLC packets that can be served is limited by the number of available eMBB users. Due to this, a limited number of URLLC packets can be allocated within the same mini-time slot. To overcome this limitation (limitation one in the previous paragraph), we propose a fast greedy algorithm to allocate the URLLC users. The algorithm starts by sorting the users in an ascending order based on a pre-defined URLLC allocation strategy π. The allocation strategy aims at balancing between the URLLC packet admission rate and the eMBB QoS (rate loss threshold). In this context, we have defined the following URLLC allocation strategies:

- Minimum-eMBB loss (MeL): This policy exploits two observations. First, the rate loss of each eMBB user is bounded by R_e^{th}. Therefore, starting the allocation over the weak eMBB user leads to the minimum total loss. Second, starting the allocation with the strong URLLC user increases the admitted URLLC packets rate as it needs less resources than the remaining URLLC users (weaker URLLC users), which increases the probability to allocate more URLLC packets. Accordingly, **Algorithm 1** (which refers to Algorithm 2.1) starts by sorting the eMBB and the URLLC packets such that $|h_1|^2 \leq |h_2|^2 \cdots \leq |h_E|^2$ and $|g_1^n|^2 \geq |g_2^n|^2 \geq \cdots \geq |g_l^n|^2$, respectively.

- Loss-threshold proportional (TP): The main idea is that the eMBB rate loss threshold is already defined to satisfy the QoS requirements of the eMBB traffic. Consequently, the eMBB QoS will not be significantly impacted by superimposing eMBB resources, while the eMBB rate loss is less than the loss threshold. Accordingly, Algorithm 1 starts by sorting the eMBB users in a descending order based on the rate loss threshold.

Then, for each URLLC packet, the BS tests the feasibility conditions between the URLLC packet and the first available eMBB user. If the feasibility conditions hold, the BS allocates the URLLC packet over this eMBB user and updates its rate loss threshold and its associated frequency resources. Otherwise, the BS repeats the same procedure with the next eMBB user. The Boolean variable aims to reduce the time complexity of the proposed algorithm. For clarity, as the URLLC users are sorted in a descending order based on their channel conditions, then for all $l \in \{1, \ldots, l^n\}$, the lth URLLC packet will need the same or higher resource coefficient $\varphi_{e,l}^n$ and power allocation $\alpha_{e,l}^n$ than the previous allocated packet, respectively. Hence, for all $l \in \{1, \ldots, l^n\}$, the lth URLLC packet will not satisfy the feasibility conditions in (2.17) and (2.18) if the previous packet is dropped, i.e., the feasibility conditions are not satisfied.

Algorithm 2.2: Many-to-Many pairing Algorithm.

```
Step 1: One-to-one pairing;
Sort eMBB users based the adopted URLLC allocation policy $\pi$ ;
- Sort URLLC users based on the channel gain in descending order.;
for $l \in \{1, \dots, l^n\}$ do
    Boolean=0 allocating indicator variable;
    for $e = 1 \rightarrow E$ do
        if feasibility conditions in (2.17) and (2.18) then
            if feasibility conditions in (2.29) not hold then
                $\varphi_{e,l}^{n} = \varphi_{e,l}^{n\ min}$
            end
            allocate URLLC packet $l$ on the eMBB user $e$;
            Update $\widehat{R}_e^{th}$ and $\phi_e$;
            Boolean=1;
            break;
        end
    end
    if Boolean==0 then
        $\hat{l}^n = append(l)$ add non allocated URLLC packets.
    end
end
Step 2: URLLC packet allocation with segmentation ;
for $l \in \{1, \dots, \hat{l}^n\}$ do
    if feasibility conditions in (2.29) holds then
        allocate URLLC packet $l$ on the eMBB users $\hat{\mathcal{E}} \subseteq \{1, \dots, E\}$;
        Update $\widehat{R}_e^{th}$ and $\phi_e$ of the eMBB users $\hat{\mathcal{E}} \subseteq \{1, \dots, E\}$;
    else
        break;
    end
end
```

2.4.2 Many-to-Many Pairing

This section addresses the second limitation of the one-to-one pairing approach discussed in the previous subsection, by considering URLLC packet segmentation. Going back to the original problem, the optimization problem $\mathcal{P}$ ends up as a many to many pairing assignment problem, where each URLLC packet can be segmented and the different segments are distributed over multiple eMBB users. In order for problem $\mathcal{P}$ to be solved, the following condition should be satisfied.

Theorem 2.3. *The outer problem $\mathcal{P}_1^{outer}$ is feasible if and only if the following condition holds.*

$$\sum_{l}^{l^n} \varphi_l^n \leq \sum_{e}^{E} \frac{\widehat{R}_e^{th} - \sum_{i=1}^{n-1} \widehat{R}_{e,l}^{i}}{WR_e}, \text{ almost surely,} \tag{2.29}$$

where φ_l^n is the needed frequency resources for allocating the lth URLLC packet when puncturing is used.

Proof. For proving Theorem 2.3, one can compare the minimum needed frequency resources for all URLLC packets with the available resources at the eMBB users. Then, the outer problem $\mathcal{P}_1^{outer}$ is feasible if there is enough frequency resources to serve all the URLLC packets. This completes the proof. □

Using the result of Theorem 2.3, **Algorithm 2** (which refers to algorithm 2.2) is an extension of Algorithm 1 by considering the segmentation of the URLLC packets. Algorithm 2 is composed of two steps. The first step performs pairing between each URLLC packet with one eMBB user (more than one URLLC packet can be allocated onto one eMBB user). The feasibility condition in (2.29) aims to maximize the number of the served URLLC packets by selecting the minimum feasible frequency resources, which enhance the utilization efficiency of the superposed eMBB frequency resources. The second step aims to allocate each URLLC packet that needs frequency resources more than what is available for one eMBB user. This step therefore performs the segmentation of these URLLC packets among several eMBB users. Accordingly, our proposed algorithm attempts to minimize the segmentation loss by giving high priority for one-to-one pairing (i.e., the URLLC packet does not get segmented). Then, it performs segmentation for the remaining URLLC packets. Based on this, we introduce a *Segmentation ratio* (SR) metric, which measures the overhead associated with the URLLC packets segmentation as:

$$SR = \frac{\text{Number of transmitted segments}}{\text{Number of transmitted packets}}. \tag{2.30}$$

This ratio increases when the URLLC packet segmentation increases. When no segmentation is performed, $SR = 1$.

1) **Algorithm complexity**: One can easily notice that the proposed algorithm, Algorithm 2, has a polynomial time complexity of $O(l^n \times E)$, as the algorithm consists of two loops: inner loop of time complexity of $O(1)$ and outer loop of time complexity of $O(l^n)$. Moreover, the algorithm convergence is guaranteed as it has deterministic stopping criteria with maximum number of iterations of $l^n \times E$. [3]

2.4.3 Extension to Multi-Cell Setup

This work considers the single-cell network setting as a proof of concept for the proposed low complexity mechanism that is based on the derived closed-form expression when employing the superposition scheme in the context of the eMBB and URLLC coexistence problem. Specifically, the proposed scheme can be employed for the multi-cell scenario while considering the essential considerations of the multi-cell setup. To elaborate, the eMBB and URLLC users can be clustered into cell-center and cell-edge users, where the association problem should be solved first. The association problem aims to define the cell-center users (eMBB and URLLC) that are served through only one BS, and the cell-edge users (eMBB and URLLC) that can bThise served by multiple adjacent and cooperating cells through the coordinated multipoint (CoMP) technique [52, 53]. For both clusters of users, the proposed scheme can be employed while using an inter-cell interference mitigation mechanism in order to ensure the reliability of the URLLC traffic and the eMBB QoS requirements.

[3] Analyzing the optimality gap for the proposed pairing algorithm through analytical or numerical analysis is difficult, since all the states of the URLLC packets in all mini-time slots must be known at the BS at the beginning of the time slot, which is not a realistic assumption. However, the gaps between the proposed scheme and a lower bound for the eMBB rate loss and an upper bound for the URLLC packet admission rate are around 20% and 0.1%, respectively. These bounds are obtained numerically by relaxing the integer constraints of the URLLC resources allocation and assuming independent mini-time slots in terms of the eMBB rate loss constraints.

Table 2.1: Simulation parameters

Parameter	Symbol	Value
URLLC packet distribution	$\mathcal{D}$	Bernoulli
URLLC packet probability	p	0.04
URLLC packet size	ζ	96, 256 bits
Transmitted signal to Noise Ratio	P/σ_0	20 dB
mini-time slots	N	7
Number of resource blocks	B	100
mini-time slot duration	δ	0.143 ms
Resource block bandwidth	W	180 kHz
URLLC block error probability	ϵ_u	10^{-6}

2.5 Simulation Results

2.5.1 Simulation Settings

We perform various simulations to evaluate the performance of the proposed superposition algorithm. We consider a wireless network which consists of one BS and $E = 8$ eMBB users. The number of URLLC users U is a key parameter in the performance of the proposed superposition scheme, and it is assumed to be in the range $\{20, 110\}$, unless otherwise stated. Moreover, we consider a high URLLC load with packet generation probability $p = 0.04$ for each user. Hence, for the case when the number of URLLC users $U = 60$, the average arrival URLLC packets is $60 * .04 * 7 = 16.8$ packets per millisecond, which is high when compared to the values considered in prior works [41].[4] The wireless channels between the BS and the eMBB and the URLLC users are assumed flat fading Rayleigh distribution with time slot coherence time and a scale parameter equals to one. The parameters used throughout are shown in Table 2.1 and the simulation results are performed over 10^4 independent realizations of the channel gains.

The performance of the proposed superposition algorithm, which is summarized in Algorithm 2 for both MeL and TP allocation strategies, is compared with those of three puncturing baseline schemes proposed in the literature, which are 1) random puncturing [3] 2) user-based puncturing [30], and 3) fairness-based puncturing. These baselines are detailed as follows.

[4] Accordingly, the distribution of arriving URLLC packets at mini-slot n follows the binomial distribution with parameters $\mathbf{Binomial}(p, U)$.

1. *Baseline 1 (random puncturing):* the eMBB user that is going to be punctured is uniformly selected from all existing eMBB users while its rate loss threshold is satisfied. Random puncturing is considered as a baseline due to the optimality of random placement for the linear loss models [3]. This follows from the fact that if the eMBB resources are punctured uniformly, then the punctured resources are proportional to the bandwidth assigned to each eMBB user [3].

2. *Baseline 2 (user-based puncturing):* the eMBB users are sorted in an ascending order based on their channel conditions. Then, the puncturing procedure starts with the user with the lowest (worst) channel gain until its rate loss threshold is reached. The advantage of user-based puncturing scheme is its ability to minimize the cumulative rate loss of all eMBB users by puncturing the users with the lowest achievable rates [30].

3. *Baseline 3 (fairness-based puncturing):* the fairness policies presented in [3, 54] are considered in the coexistence problem due to their simplicity and optimality. In threshold-based fairness, the URLLC traffic is allocated to the eMBB users according to the eMBB users associated loss thresholds. As such, at each mini-time slot, the eMBB user with a higher loss threshold is punctured.

Besides those puncturing baselines, the proposed scheme is compared with the case when the URLLC traffic is perfectly known at the beginning of each eMBB time slot. Although this scheme is not realistic due to the sporadic nature and the latency requirement of the URLLC traffic [45], it can assess the performance of the proposed scheme.

In the following, we first start by showing the advantage of URLLC packet segmentation on the URLLC packet admission rate by comparing the performance of Algorithm 1 and Algorithm 2. Second, we investigate the segmentation loss of the proposed algorithm (Algorithm 2) with the aforementioned puncturing baselines. Third, we investigate the impact of the URLLC rate threshold and URLLC packet size on the admission rate of the URLLC and the eMBB rate loss. Then, we analyze the effect of the eMBB rate threshold on the URLLC packet admission rate. Finally, we evaluate the computational time of the proposed algorithm with those of the considered puncturing baselines.

Fig. 2.5 presents the URLLC packet admission rate of both Algorithm 1 and Al-

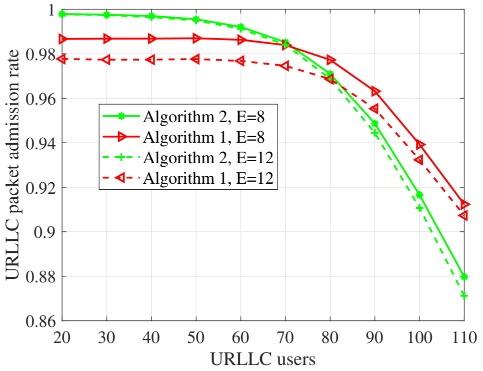


Fig. 2.5: URLLC packet admission rate of Algorithm 1 and Algorithm 2 against the number of URLLC users. The adopted URLLC packet size is $\zeta = 256$ bits and MeL policy is adopted.

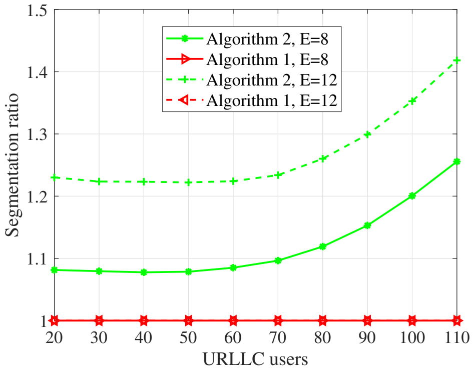


Fig. 2.6: Segmentation ratio of Algorithm 1 and Algorithm 2 against the number of URLLC users. The adopted URLLC packet size is $\zeta = 256$ bits. MeL is adopted.

gorithm 2 as we vary the number of URLLC users (load). The results show that Algorithm 2 exhibits better packet admission rate for the URLLC service than Algorithm 1 when the load is low, and the opposite is true at higher loads. The reason behind that is that Algorithm 2 is able to allocate URLLC packets which experience bad channel conditions by segmenting each packet and allocating it over multiple eMBB users, which leads to admitting URLLC packets requiring more resources, i.e., $\varphi_l^n > \phi_e$. On the other hand, Algorithm 1 drops the URLLC packets with bad channel conditions, i.e., the packets that required more than ϕ_e, which means that Algorithm 1 supports a limited number of URLLC users, i.e., less than U. Hence, at lower URLLC load where the eMBB rate users have rate constraint more than or equal the expected URLLC load, Algorithm 2 can accommodate all URLLC users, even with bad channel conditions, without impacting the URLLC packet admission rate. At higher URLLC load where the eMBB rate users have rate constraint less than the expected URLLC load, Algorithm 2 allocates the packets in the sense of first come first serve, even the packets that have bad channel conditions, which in turn impacts the URLLC performance due to the eMBB QoS. In other words, as the URLLC load is low, the allocation of the URLLC traffic with bad channel conditions does not impact the performance of Algorithm 2 compared to Algorithm 1. At high URLLC load, the allocation of the URLLC traffic with bad channel conditions will

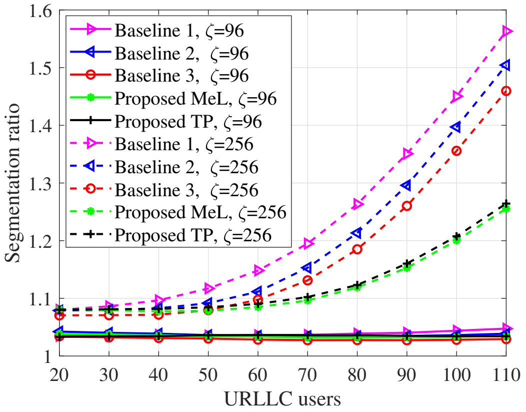


Fig. 2.7: Segmentation ratio vs the number of URLLC users. The adopted URLLC packet size is $\zeta = 256$ bits and MeL policy is adopted.

degrade the performance of Algorithm 2. Noting that the URLLC packet admission rate is considered to be 99.99%, hence Algorithm 2 is proper in the URLLC operating region.

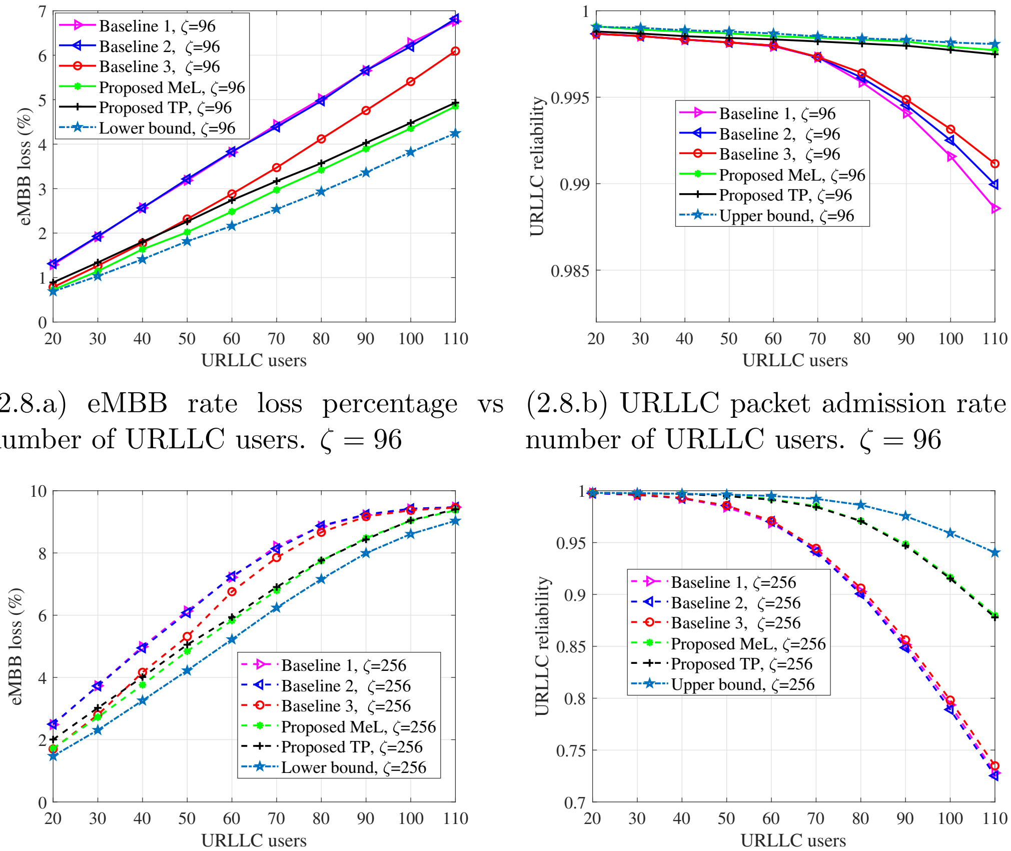
Now, as the number of served eMBB users increases, the URLLC packet admission rate for Algorithm 1 becomes very low. In fact, it is reduced from 0.99 for $E = 8$ to 0.978 for $E = 12$. This is because, for each eMBB user, the pairing possibility and the afforded resources for superposition decreases by increasing the number of served eMBB users. Hence, more segmentation is essential to allocate the URLLC packets (as discussed in the limitations in section 2.4.1). Fig. 2.6 presents the SR against the number of URLLC users. Indeed, as Algorithm 1 does not perform segmentation, i.e., (SR=*1*), Algorithm 1 has a lower SR than Algorithm 2, but at the expense of achieving lower performance in terms of the admission rate of the URLLC traffic as depicted in the same figure.

2.5.2 Segmentation Loss Comparison

Fig. 2.7 presents the segmentation ratio as we vary the number of URLLC users in the network, where packets of different sizes are considered for the URLLC service. As shown in this figure, when the packet size is large, i.e., 256 bits, the proposed superposition scheme has a lower segmentation ratio than those exhibited by the puncturing baseline methods. This is mainly because the rate loss threshold for eMBB users is reached faster by the considered puncturing baselines than the proposed approach, forcing the URLLC service to split among multiple eMBB users. In other words, the eMBB rate loss resulting from puncturing is high and exceeds the loss threshold. Hence, more segmentation is required to allocate the URLLC packets. On the other hand, when the packet is small, the proposed scheme has a similar segmentation ratio to the puncturing schemes. This is because the URLLC packet can occupy a lower number of RBs. Hence the possibility that the packet needs segmentation is very low. Moreover, this figure shows that when the number of URLLC users is low, the proposed algorithm has a segmentation ratio close to that of the considered puncturing baselines. This is because when the number of URLLC users is low, the number of generated URLLC packets is also low. Therefore, the BS has more room to superpose the entire set of URLLC packets over the eMBB data while guaranteeing the eMBB rate loss threshold R_e^{th}, which in turn reduces the need for segmenting the URLLC packets.

2.5.3 Effect of URLLC Packet Size

The impact of the URLLC packet size ζ on both the URLLC packet admission rate and the eMBB rate loss is illustrated in Fig. 2.8. Specifically, Fig. 2.8 presents the percentage of eMBB rate loss and the achieved URLLC service packet admission rate η versus the number of URLLC users in the network, respectively, for different URLLC packet sizes. Fig. 2.8.a and Fig. 2.8.c show that, as we vary the number of URLLC users, the proposed superposition schemes achieves lower eMBB rate loss and higher URLLC packet admission rate compared to the considered puncturing baseline methods. This is because, with puncturing, the eMBB data of punctured resources are entirely lost to carry the URLLC service, while in the proposed superposition method, each eMBB user adopts SIC to cancel the interference and to extract its own data, which enables to achieve higher rates, hence less eMBB rate loss and better URLLC packet admission rate. In fact, Fig. 2.8.b and Fig. 2.8.d show that for larger number of URLLC users, such as $U = 110$ for example, the maximum achievable packet admission rate (0.999) is maintained via the proposed schemes, whereas for the baseline methods, the service packet admission rate starts to gradually decrease when the number of URLLC users U reaches 80. Moreover, Fig. 2.8.b and Fig. 2.8.d show that the gap between the proposed MeL algorithm and the upper bound is extremely low when the URLLC load is moderate, which is the case of



(2.8.a) eMBB rate loss percentage vs number of URLLC users. $\zeta = 96$

(2.8.b) URLLC packet admission rate vs number of URLLC users. $\zeta = 96$

(2.8.c) eMBB rate loss percentage vs number of URLLC users. $\zeta = 256$

(2.8.d) URLLC packet admission rate vs number of URLLC users. $\zeta = 256$

Fig. 2.8: eMBB rate loss and URLLC packet admission rate versus the number of URLLC users U for different URLLC packet sizes.

the URLLC service class. This result demonstrates that the proposed superposition method can accommodate higher URLLC load, around 30% more, compared to the considered puncturing baselines. On the other hand, the suggested superposition method (**Proposed MeL**) has a loss rate of 4% for 110 URLLC users, while the loss for puncturing methods is 6%. Moreover, it is clear that the **Proposed MeL** scheme has slightly better performance, in term of the eMBB rate loss, compared to the **Proposed TP**. This is because the **Proposed MeL** allocates the URLLC load based on the users with higher rate threshold, i.e., eMBB user with high rate will be superimposed instead of the eMBB users with lower data rates. Similarly, puncturing **baseline 3** superiors other puncturing base lines as it allocates the URLLC load on the eMBB user with lower data rates and this gain is vanished by increasing the

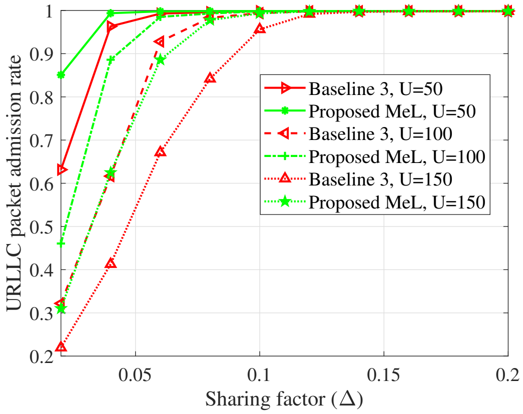


Fig. 2.9: URLLC traffic packet admission rate achieved by the proposed algorithm versus the eMBB rate loss threshold for different numbers of URLLC users.

URLLC load. Another interesting observation from Fig. 2.8 can be remarked; as the URLLC packet size ζ increases, the URLLC packet admission rate decreases as shown in Fig. 2.8.b and Fig. 2.8.d, since more URLLC packets will be dropped due to the limited resources for puncturing (eMBB QoS limitation).

2.5.4 Impact of the Sharing Factor (eMBB Rate Loss Threshold)

The impact of the eMBB rate loss threshold $R_e^{th} = \Delta * R_e$ on the packet admission rate of the URLLC service is illustrated in Fig. 2.9. We compare here the performance of the proposed algorithm and the puncturing baseline while adopting the minimum eMBB rate loss policy. As shown in this figure, when R_e^{th} (sharing factor Δ) increases, the URLLC packet admission rate increases until the maximum achievable packet admission rate. In other words, when the loss threshold R_e^{th} is low (i.e., strict QoS for the eMBB), the BS has less room to explore to allocate the URLLC load. On the other hand, when the eMBB rate loss threshold increases (i.e., less stringent requirement for eMBB and loss can be tolerated), the BS has more room to allocate the URLLC traffic over the eMBB resources, which enhances the URLLC packet admission rate. Similarly, when the URLLC load increases, the possibilities to allocate the URLLC load decreases, hence larger R_e^{th} is essential to achieve the maximum URLLC packet admission rate. This is evident since at 50, 100, 150 URLLC users the maximum URLLC packet admission rate is achieved at 2%, 4%, and 6%, respectively.

In summary, the eMBB QoS is clearly shown to affect the URLLC reliability.

Therefore, as a future work, the eMBB rate loss threshold, the URLLC load and the URLLC QoS requirement will be jointly considered when optimizing the eMBB spectral efficiency. Moreover, to enhance the URLLC packet admission rate, the impact of the quality of the wireless propagation environment should be minimized. Hence, reconfigurable intelligent surfaces [19] will be included to control URLLC channel conditions.

2.5.5 Complexity of the Proposed Algorithm

As the latency is a critical metric for the URLLC service, Fig. 2.10 presents the computational time of the proposed superposition scheme (Algorithm 2). It shows that the proposed MeL allocation policy has a lower computation as compared to the TP policy. This is due to the fact that the MeL leads to fewer segments and hence less computational time. Moreover, increasing the URLLC load, which is equivalent to increasing the URLLC allocated segments, increases the computational time. However, the time complexity of the proposed algorithm is still less than the duration of one mini-time slot, i.e., $\delta = 0.143$ ms, which makes the proposed superposition technique a practical method for real-time and efficient multiplexing between eMBB and URLLC traffic. Moreover, the processing time of the proposed algorithm can be further enhanced when using more computing resources available at the network edge.[2]

2.6 Summary

In this chapter, we proposed a low-complexity resource allocation scheme in a downlink network which consists of a single base station serving simultaneously URLLC and eMBB users. With the objective to minimize concurrently the eMBB rate loss and the overhead due URLLC packet segmentation, we formulated the allocation problem as a MINLP which is generally hard to solve. We derived the feasibility region and the optimal solution for the case of one-to-one pairing. Then, we applied the results for the case of many-to-many pairing. Simulation results showed that the

[2] The algorithm was implemented in Matlab using a machine with the following characteristics: System Type: x64-based PC Processor: Intel(R) i7-8700H CPU @3.20GHz, 16 Gigabyte RAM.

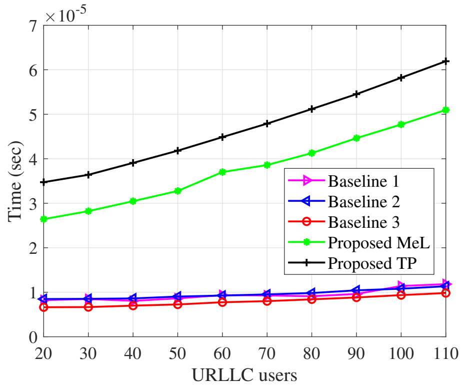


Fig. 2.10: Computational time against the number of URLLC users. The adopted URLLC packet size is $\zeta = 256$ bits

proposed algorithm achieves better URLLC packet admission rate and eMBB rate while satisfying the QoS of the eMBB users compared to state of the art puncturing baselines. Moreover, the proposed algorithm has low time complexity which is in order of sub-millisecond, making it an efficient tool to be used in practice.

Chapter 3

RIS-Aided eMBB/URLLC Traffic Multiplexing

3.1 Background, Related Works, and Contributions

Chapter 2 studied the superposition/puncturing scheme as spectrally efficient multiplexing scheme for the coexisting eMBB and URLLC traffic. However, the improve of the channel conditions of both services can further boost their performance in terms of rate, reliability and latency. For clarity, the performance of the eMBB and URLLC services, which is measured in terms of data rates for the eMBB traffic and the reliability and latency for the URLLC traffic, depends directly on the channel quality of the coexisting eMBB and URLLC users. When the channel conditions are favorable, fewer resources are needed to serve the eMBB and URLLC users simultaneously. Precisely, when the channel gains of the URLLC users are high, the number of frequency resources which are required to achieve the target reliability and latency and will be punctured from the ones of the eMBB users will reduce. Therefore, the losses in the data rates of eMBB users will decrease as well. Moreover, reducing the

The content of this chapter leads to an IEEE published journal, one conference and one IEEE magazine [29, 49, 55].

needed transmission time/frequency resources for the URLLC traffic is equivalent to lower transmission latency, lower outage probability due to the limited resources, and better availability to allocate more URLLC packets. Alternatively, when the channel gains of the eMBB users are high, the resources required to achieve their target data rates will reduce. Hence, more of these resources can be punctured to accommodate the URLLC traffic. Therefore, the number of served URLLC users with the eMBB users will increase. Thus, better eMBB rate, URLLC latency, reliability, and better transmission resources availability are achieved. Subsequently, the following question arises: *how can one increase the channel gains of the coexisting eMBB and URLLC services? In other words, how can one enhance the propagation environment and the channel conditions of the coexisting eMBB and URLLC users?* RIS technology has emerged as a key solution which provides answers to the above questions.

RIS is a promising technology for next generation wireless networks that has been lately receiving a significant interest from both academia and industry due to its capability in controlling and configuring the wireless propagation environment [17]. RIS is a planar array that is composed of a large number of passive and low-cost reflecting elements, where each can be tuned independently to a certain phase-shift [19, 56]. By appropriately tuning the phase-shift of each passive element, the reflected signals by the RIS can be constructively added at the points of interests [19, 57]. Therefore, the channel gains and the received signal strengths at the end users, which are the eMBB and the URLLC users in the context of this chapter, are enhanced [19, 56]. As opposed to traditional relaying techniques (e.g., amplify-and-forward, and decode-and-forward), RIS exhibits a multitude of advantages. In fact, RIS offers a low cost solution that is both energy and spectral efficient [19, 56]. In addition, RIS is capable of passively reflecting the incident signals without additional radio frequency chains, which results in a lower power consumption. Moreover, the radio signals reflected by the RIS are free from noise corruption [19]. Consequently, and motivated by the aforementioned benefits of RIS, this chapter considers an RIS-assisted wireless network and study the problem of coexistence of services with heterogeneous requirements, namely URLLC and eMBB. The chapter explores the added value of this new degree of freedom offered by the RIS technology on the performance of superimposing the URLLC traffic on a network designed to serve the eMBB users, which to the best of our knowledge, has not been explored in earlier work.

Motivated by the great benefits offered by RIS in controlling and configuring the wireless propagation environments, several works studied the configuration of the phase-shifts of the RIS elements, also known as passive beamforming, to enhance the performance of wireless cellular systems [58–61]. The authors in [58] proposed an alternating algorithm for passive and active beamforming design to minimize the total transmit power, where a semi-definite relaxation (SDR) approach was adopted to configure the passive beamforming. Yu et al. exploited in [59] the fixed point iteration and manifold optimization methods to maximize the spectral efficiency in RIS-aided cellular networks. In [60], the RIS-aided NOMA systems was studied with the objective max-min rate problem by jointly optimizing the power allocation and the RIS phase-shift matrix. In [61], multi-user communications aided by single or multiple RISs were studied for multiple-input single-output (MISO) and multiple-input multiple-output (MIMO) systems. From a medium access control (MAC) perspective, the authors also came up with different possible solutions for the considered RIS-aided cellular systems. However, a few works investigated RIS aided URLLC traffic [28, 62–64]. Authors in [62] proposed joint active beamforming and phase-shift optimization to allocate URLLC traffic with the objective of maximizing the URLLC sum rate in MISO system aided by RIS, in which a set of BSs cooperate to serve the URLLC traffic. The integration between the UAVs and the RIS was studied in [63] to support the URLLC traffic. In [64], the coverage and link performance of the RIS-assisted UAV systems was studied by proposing an adaptive RIS-assisted transmission protocol to control the RIS association and the RIS phase-shifts configuration. The authors also proposed a multi-task learning to reduce the time complexity of the proposed transmission protocol. In [28], a grant-free access scheme aided by an RIS was proposed to enhance the URLLC reliability. Although, the works in [28, 62–64] studied the performance of RIS-aided URLLC traffic, they considered a simple scenario where the BS serves only URLLC users. In other words, the coexistence of URLLC and eMBB traffic aided by RIS was not addressed in [28, 62–64]. Similarly, the works in [58–61] considered the RIS-aided wireless networks for only the eMBB service class. Hence, there is a noticeable lack in studying the coexistence problem of eMBB and URLLC traffic in RIS-aided wireless networks, which is the focus of this chapter.

In this chapter, we consider an RIS-assisted cellular network that supports coexistent eMBB and URLLC services while taking into consideration the trade-off between the services regiments. In this context, we first formulate a problem for allocating the eMBB users at the beginning of each time slot. The problem aims at maximizing the eMBB sum rate with resource over-provisioning to accommodate future arrival of URLLC packets while satisfying required QoS of the different eMBB users by jointly optimizing the power allocation policy and the RIS phase-shift matrix. However, due to the coupling between the power allocation at the BS and the passive beamforming at the RIS, the formulated eMBB allocation problem is not convex, and hence, difficult to be solved. To overcome this issue, the problem is solved by applying the alternating optimization (AO) technique. Particularly, the original problem is decomposed into two sub-problems, namely, a power allocation sub-problem and a phase-shift matrix optimization sub-problem. The power allocation sub-problem is a convex problem which is solved by applying the Karush-Kuhn-Tucker (KKT) method. Meanwhile, the SDR and Gauss randomization methods are adopted to solve the phase-shift sub-problem.

At each mini-time slot, the incoming URLLC packets should be served and transmitted simultaneously over the ongoing eMBB data using the puncturing scheme in order to satisfy the URLLC latency requirement. Hence, with the goal of maximizing the number of served URLLC packets at each mini-time slot while satisfying the eMBB and URLLC rate requirements, the frequency and power allocation should be jointly optimized along with the RIS phase-shift matrix. However, the task of optimizing the RIS configuration has a high computational time compared to the mini-time slot duration, which may violate the URLLC latency requirement. To overcome this issue, multiple RIS configurations are proactively designed and communicated to the RIS controller at the beginning of each time slot prior to the arrival of the URLLC load. Then, a control signal, if needed, is sent to the RIS controller to switch between these configurations at each mini-time slot. Once the RIS configuration is fixed, the URLLC allocation problem is formulated as a mixed integer non linear program which is difficult to be solved in a polynomial time within each mini-time slot. Hence, we convexify the problem by applying a variable change approach. Due to the nature of the URLLC service which requires high reliability, the problem is decomposed into two sub-problems 1) the URLLC admission problem and 2) the URLLC allocation problem. The first problem aims at maximizing the admitted URLLC packets while the latter aims at minimizing the eMBB loss according to the URLLC allocation

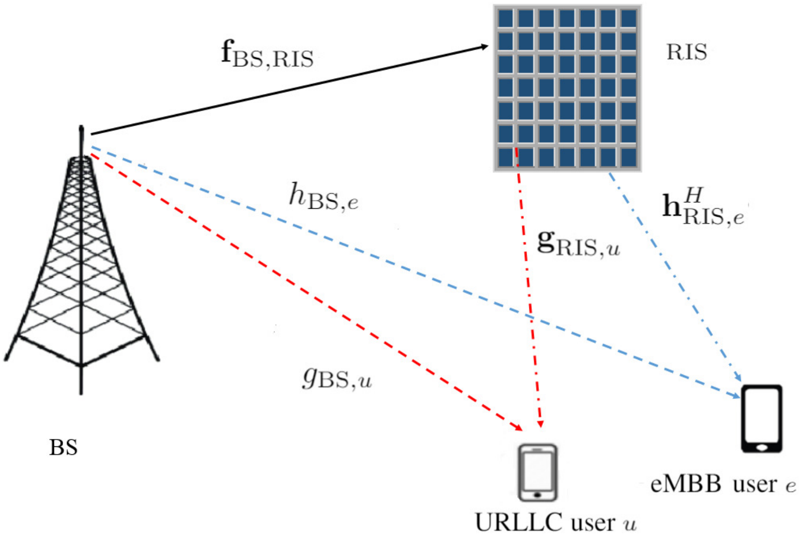

$\mathbf{f}_{\text{BS,RIS}}$
RIS
$h_{\text{BS},e}$
$\mathbf{h}^H_{\text{RIS},e}$
$\mathbf{g}_{\text{RIS},u}$
$g_{\text{BS},u}$
BS
eMBB user e
URLLC user u

Fig. 3.1: System model.

strategies.

The performance of the proposed scheme is illustrated through extensive simulations. The obtained results show that the proposed scheme achieves considerable gain on the URLLC packet admission rate and the eMBB sum rate compared to when no RIS is deployed. Moreover, the proposed switching scheme between the proactively configured RIS phase-shift matrices enhances the performance of both URLLC and eMBB services compared to the case when a fixed phase-shift matrix is used. Finally, the best allocation for the RIS is demonstrated to be close to the BS when the eMBB and URLLC users are distributed randomly on the coverage area of the BS.

The rest of the chapter is organized as follows. Section 3.2 presents the system model. Section 3.3 presents the problem formulations for the eMBB and URLLC allocation. Section 3.4 presents the roadmap for the solution approach. Sections 3.5 presents the proposed solution approach for the URLLC allocation problem. Sections 3.6 and 3.7 present the simulation results and the conclusion, respectively.

3.2 System Model

We consider a downlink radio access network consisting of a single BS, equipped with one single antenna, serving several spatially dispersed eMBB and URLLC users, each equipped with one single antenna.[1] The total bandwidth allocated to the BS is divided into B resource blocks (RBs), each of bandwidth W. The service period of cellular users is divided into equally sized time slots. Each time slot is further divided into a set of M equally sized mini-time slots, denoted by $\mathcal{M} \triangleq \{1, 2, \ldots, M\}$, where the duration of each mini-time slot is denoted by τ. Let $\mathcal{E} \triangleq \{1, 2, \ldots, E\}$ and $\mathcal{U} \triangleq \{1, 2, \ldots, U\}$ denote the sets of the eMBB and URLLC users, respectively, where E and U denote the total numbers of eMBB and URLLC users, respectively, that are simultaneously communicating with the BS within one time slot. The eMBB users are admitted at the beginning of each time-slot and the adopted multiple access technique is the orthogonal frequency-division-multiple-access (OFDMA). In addition, the eMBB users share equally the available frequency resources, i.e., each eMBB user has $b = \frac{B}{E}$ RBs. The URLLC load, on the other hand, can arrive within the serving time slot, i.e., during one mini-time slot within the same time slot, and it should be served immediately to satisfy its latency requirements. Hence, the URLLC packets are immediately transmitted upon arrival in the following mini-time slot by puncturing the frequency resources that are already allocated to the eMBB users at the beginning of the time slot. Moreover, for all $u \in \mathcal{U}$, the arrival process of URLLC packets per mini-time slot of the uth URLLC user is assumed to follow a Poisson distribution with an average arrival rate λ_u [65]. The symbols used throughout the chapter are listed in Table 3.1.

One single RIS, equipped with N reflective elements, is deployed to dynamically control the propagation environment between the BS and the different eMBB and URLLC users. Each reflective element consists of an atom that can adjust the phase of each incident wave. Let $\mathcal{N}$ denote the set of indices between 1 and N, i.e., $\mathcal{N} \triangleq \{1, 2, \ldots, N\}$. Then, we denote by $\mathbf{\Phi} \triangleq \operatorname{diag}\left(e^{j\phi_1}, e^{j\phi_2}, ..., e^{j\phi_N}\right) \in \mathbb{C}^{N \times N}$ the phase-shift matrix of the RIS, where for all $n \in \mathcal{N}$, $\phi_n \in [0, 2\pi)$ denotes the phase-shift of the nth RIS reflective element. The RIS is connected to a control unit that adjusts the phase-shift matrix $\mathbf{\Phi}$. The channel state information (CSI) of all communicating nodes, including the eMBB and URLLC users and the RIS, are assumed

[1] In this work, we consider the case of single antenna BS and cellular users. The main motivation behind such a choice is that a proof of concept for multiplexing URLLC and eMBB traffic with the aid of RIS is aimed to be investigated in a simple setup, so that the fundamental insights and observations can be obtained. Nevertheless, the same analysis can be extended for the case when the BS and/or the cellular users are equipped with multiple antennas.

Table 3.1: Table of symbols used in the chapter.

Symbol	Description
$\mathcal{E}$	Set of eMBB users
$\mathcal{U}$	Set of URLLC users
$\mathcal{N}$	Set of RIS elements
$\mathcal{M}$	Set of mini-time slots
$\mathcal{L}^m$	Set of URLLC packets at mini-time slot m
E	Number of eMBB users
U	Number of URLLC users
N	Number of RIS elements
M	Number of mini-time slots per time slot
L^m	Number of URLLC packets at mini-time slot m
P_{BS}	Power budget at the BS
$\mathbf{\Phi}$	RIS phase-shift matrix
B	Number of resource blocks
τ	Mini-time slot duration
W	Resource block bandwidth
r_{th}	eMBB rate threshold
c_{th}	URLLC rate threshold

to be perfectly known at the BS [58].[2] In this context, let $\mathbf{f}_{\text{BS,RIS}} \in \mathbb{C}^{N\times 1}$ denote the vector that contains the channel coefficients between the BS and the RIS elements. Moreover, for all $e \in \mathcal{E}$, let $h_{\text{BS},e} \in \mathbb{C}$ and $\mathbf{h}_{\text{RIS},e} \in \mathbb{C}^{N\times 1}$ denote the channel coefficients from the BS to the eth eMBB user and from the RIS to the eth eMBB user, respectively. Additionally, for all $u \in \mathcal{U}$, let $g_{\text{BS},u} \in \mathbb{C}$ and $\mathbf{g}_{\text{RIS},u} \in \mathbb{C}^{N\times 1}$ denote the channel coefficients from the BS to the uth URLLC user and from the RIS to the uth URLLC user, respectively. Each communication link in the network is assumed to have a quasi-static flat-fading Rayleigh channel, except the ones between the BS and the RIS elements, which are assumed to have a Rician channel model. This is basically due to the fact that a necessary condition for the deployment of RIS in any cellular system is that the BS has a direct line-of-sight with the RIS.

[2] We assume that the BS knows the locations of the RIS and of all cellular users, which can be used then to perfectly estimate the different communication nodes at the beginning of each time slot.

3.2.1 Signal Model and Rate Analysis

Considering both the direct link and the cascaded link through the RIS between the BS and each eMBB user, the received signal per RB at the eth eMBB user, for all $e \in \mathcal{E}$, can be expressed as [66]

$$y_e = \left(h_{\mathrm{BS},e} + \mathbf{h}_{\mathrm{RIS},e}^{H} \mathbf{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}\right) \sqrt{p_e}\, x_e + z_e, \tag{3.1}$$

where x_e is the signal that contains the data of the eth eMBB user to be transmitted throughout the entire time-slot, p_e is its associated allocated power per RB at the beginning of the time slot and z_e is the additive white Gaussian noise (AWGN) experienced at the eth eMBB user throughout the entire time slot, which is assumed to be $\mathcal{CN}(0, \sigma^2)$ distributed. Accordingly, by considering a target block error rate (BLER) ϵ_{eMBB} for all eMBB users, the data rate in [bits/s] of the eth eMBB user per RB, for all $e \in \mathcal{E}$, can be expressed as

$$r_e\left(\mathbf{\Phi}, p_e\right) = W\, \log_2\left(1 + \frac{p_e |h_{\mathrm{BS},e} + \mathbf{h}_{\mathrm{RIS},e}^{H} \mathbf{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}|^2}{\Gamma_{\mathrm{eMBB}} \sigma^2}\right), \tag{3.2}$$

where $\Gamma_{\mathrm{eMBB}} = \frac{-\ln(5\, \epsilon_{\mathrm{eMBB}})}{0.45}$ represents the SNR gap between the Shannon capacity and the achievable rate of the adopted modulation scheme, when the target BLER is ϵ_{eMBB} [65].

For $m \in \mathcal{M}$, let $\mathcal{L}^m \triangleq \{1, 2, \ldots, L^m\}$ denote the set of the arrived URLLC packets that need to be transmitted at mini-time slot m, where L^m is the number of URLLC packets at mini-time slot m. Then, for all $m \in \mathcal{M}$ and $l \in \mathcal{L}^m$, the received signal at the lth URLLC packet during the mth mini-time slot can be expressed as

$$y_l^m = \left(g_{\mathrm{BS},l} + \mathbf{g}_{\mathrm{RIS},l}^{H} \mathbf{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}\right) \sqrt{p_l^m}\, x_l^m + z_l^m, \tag{3.3}$$

where x_l^m is the signal that contains the data of the lth URLLC packet to be transmitted within the mth mini-time slot, p_l^m is its associated allocated power and z_l^m is the AWGN experienced at the URLLC user associated to the lth URLLC packet within the mth mini-time slot, which is assumed to be $\mathcal{CN}(0, \sigma^2)$ distributed. Moreover, $g_{\mathrm{BS},l} \in \{g_{\mathrm{BS},1}, g_{\mathrm{BS},2}, \ldots, g_{\mathrm{BS},U}\}$ and $\mathbf{g}_{\mathrm{RIS},l} \in \{\mathbf{g}_{\mathrm{RIS},1}, \mathbf{g}_{\mathrm{RIS},2}, \ldots, \mathbf{g}_{\mathrm{RIS},U}\}$ are the channel coefficients from the BS and from the RIS to the URLLC user associated to the lth

URLLC packet, respectively. Consequently, by considering a target BLER ϵ_{URLLC} for all URLLC users, the achievable rate per RB of the lth URLLC packet, for all $l \in \mathcal{L}^m$, within the mth mini-time slot, can be expressed as

$$c_l\left(\mathbf{\Phi}, p_l^m\right) = W \, \log_2\left(1 + \frac{p_l^m |g_{\text{BS},l} + \mathbf{g}_{\text{RIS},l}^H \mathbf{\Phi} \, \mathbf{f}_{\text{BS,RIS}}|^2}{\Gamma_{\text{URLLC}} \sigma^2}\right), \tag{3.4}$$

where $\Gamma_{\text{URLLC}} = \frac{-\ln(5\,\epsilon_{\text{URLLC}})}{1.25}$ represents the SNR gap between the Shannon capacity and the achievable rate of the adopted modulation scheme, when the target BLER is ϵ_{URLLC} [65].

3.3 Problem Formulation and Methodology

3.3.1 eMBB Allocation

We aim to design an effective puncturing scheme to multiplex the URLLC traffic over the existing eMBB traffic in an RIS-aided cellular network. The objective is to enhance the system performance in terms of the eMBB throughput and to guarantee simultaneously the different requirements of both eMBB and URLLC services. At the beginning of each time slot, the BS allocates its budget of power and designs the RIS phase-shift matrix in order to serve the eMBB users.[3] With the objective of maximizing the eMBB sum rate subject to a QoS constraint for each eMBB user and a budget power constraint at the BS, the joint power allocation and RIS phase-shift matrix design can be given by the following optimization problem.

$$\mathcal{P}_1: \quad \max_{\mathbf{\Phi}, \mathbf{p}_\text{e}} \sum_{e=1}^{E} r_e\left(\mathbf{\Phi}, p_e\right) \tag{3.5a}$$

$$\text{s.t.} \quad (1-\delta)\, b\, r_e(\mathbf{\Phi}, p_e) \geq r_{\text{th}}, \quad \forall\; e \in \mathcal{E}, \tag{3.5b}$$

$$\sum_{e=1}^{E} b\, p_e \leq P_{\text{BS}}, \tag{3.5c}$$

$$0 \leq \phi_n < 2\pi, \quad \forall\; n \in \mathcal{N}, \tag{3.5d}$$

[3] The eMBB power allocation and RIS phase-shift matrix optimization can be done for one or more eMBB time-slots, depending on the variations of the CSI of the eMBB and URLLC users [67].

where $\mathbf{p}_{\mathrm{e}} = [p_1, p_2, ..., p_E]^T$, P_{BS} is the total power of the BS and $\delta \in [0,1]$ is a predefined rate margin factor. Constraint (3.5b) can be rewritten, for all $e \in \mathcal{E}$, as

$$r_e(\mathbf{\Phi}, p_e) \geq \frac{r_{\mathrm{th}}}{b} + \delta r_e(\mathbf{\Phi}, p_e), \tag{3.6}$$

where $r_e(\mathbf{\Phi}, p_e)$ is the achievable rate of the eth eMBB user, which should be guaranteed even when some of its allocated RBs are punctured within the time-slot, r_{th} is the minimum required data rate per eMBB user, and $\delta r_e(\mathbf{\Phi}, p_e)$ depicts a rate surplus as a result of over provisioning of resources for the URLLC load at the beginning of the time slot. This over-provisioned resources will be of utility for URLLC users to use at the time of arrival of URLLC packets. In other words, some RBs of the eth eMBB user will be punctured within the time slot as long as the resulting rate-loss does not exceed $\delta r_e(\mathbf{\Phi}, p_e)$. Additionally, constraint (3.5c) guarantees that the total transmit power by the BS does not exceed its power budget. Problem $\mathcal{P}_1$ will be solved at the beginning of the time slot within which the BS will transmit the data of the E eMBB users. Afterwards, the BS will keep employing the obtained optimal power allocation scheme $\mathbf{p}_{\mathrm{e}}^*$ and the optimal RIS phase-shift matrix $\mathbf{\Phi}_{\mathrm{e}}^*$ over the following time slots as long as the number of eMBB users E and the CSI of all eMBB users do not change.

We note that, owing to constraint (3.5b), problem $\mathcal{P}_1$ may not be always feasible. Hence, if the problem is not feasible for the eMBB users, we solve an eMBB admission problem to guarantee the feasibility conditions of problem $\mathcal{P}_1$. The admission problem aims to select a set of eMBB users $\mathcal{E}_f \triangleq \{1, 2 \dots, E_f\}$, where $E_f \leq E$, based on their contribution on the eMBB sum rate such that $\mathcal{P}_1$ is feasible. To do this, we first solve problem $\mathcal{P}_1$. If the problem is not feasible, we resolve the same problem while setting $r_{\mathrm{th}} = 0$. Then, we remove the eMBB user with the lowest contribution on the eMBB sum rate. These steps are repeated until a set of eMBB users $\mathcal{E}_f \subseteq \mathcal{E}$ can be admitted.

3.3.2 URLLC Allocation

Within one time-slot, the URLLC packets are allocated within any mini-time slot upon arrival in order to satisfy their latency requirement. Such allocation is performed by puncturing the frequency RBs of the eMBB users and distributing them over the different URLLC packets to be transmitted. Within this scheme, the reliability of the

URLLC load needs also to be satisfied. In fact, let us assume that, for all $m \in \mathcal{M}$, the incoming URLLC packets have the same size, which is denoted by ζ [bits]. In this case, for all $m \in \mathcal{M}$ and $l \in \mathcal{L}^m$, the lth URLLC packet arriving at the mth mini-time slot must be entirely and successfully transmitted to the relative URLLC users. Such QoS constraint can be expressed, for all $m \in \mathcal{M}$, as

$$I_l^m c_l(p_l^m, \mathbf{\Phi}^m) \geq c_{\text{th}}, \tag{3.7}$$

where $I_l^m = \sum_{e=1}^{E} I_{e,l}^m$, in which for all $e \in \mathcal{E}_f$, $I_{e,l}^m \in \{0, 1, \ldots, b\}$ is the number of RBs punctured for the eth eMBB user and allocated to the lth URLLC packet, $\mathbf{\Phi}^m \triangleq \text{diag}\left(e^{j\phi_1^m}, e^{j\phi_2^m}, ..., e^{j\phi_N^m}\right) \in \mathbb{C}^{N \times N}$ is the RIS phase-shift matrix at mini-time slot m and $c_{\text{th}} = \frac{\zeta}{\tau}$. Furthermore, puncturing the RBs of any eMBB user may violate its QoS requirement. In fact, according to the linear rate loss model [3], the instantaneous achievable rate of the eth eMBB, for all $e \in \mathcal{E}_f$, at mini-time slot m, for all $m \in \mathcal{M}$, is given by

$$R_e\left(\mathbf{\Phi}^m, p_e, I_e^m\right) = \left(1 - \frac{I_e^m}{b}\right) r_e\left(\mathbf{\Phi}^m, p_e\right), \tag{3.8}$$

where $I_e^m = \sum_{l=1}^{L^m} I_{e,l}^m$ is the total number of punctured RBs from the eth eMBB user at mini-time slot m, which must verify $I_e^m \in \{0, 1, \ldots, b\}$. Afterwards, at each mini-time slot, we should ensure that puncturing the eMBB resources while admitting the URLLC packets does not impact adversely the eMBB QoS, which is represented by the minimum data rate r_{th} for the entire M mini-time slots. To guarantee this, we assume that the URLLC scheduler at the BS is causal so it only knows the current and the past states of the URLLC load. As a result, at each mini-time slot m, the BS assumes that no URLLC packets will arrive at the subsequent mini-time slots $\{m+1, \ldots, M\}$, i.e. $\sum_{i=m+1}^{M} L^i = 0$. Then, for all $m \in \mathcal{M}$ and $e \in \mathcal{E}_f$, the QoS constraint of the eth eMBB user at mini-time slot m that ensures the eMBB QoS rate constraint for the entire time slot can be expressed as

$$\frac{\sum_{i=1}^{m} R_e\left(\mathbf{\Phi}^i, {p_e}^i, I_e^i\right) + (M-m)\, r_e(\mathbf{\Phi}, p_e)}{M} \geq \frac{r_{\text{th}}}{b}, \tag{3.9}$$

which can be simplified as

$$R_e\left(\mathbf{\Phi}^m, {p_e}^m, I_e^m\right) \geq r^{'}_{e,\text{th}}, \tag{3.10}$$

where

$$r^{'}_{e,\text{th}} = \frac{M\, r_{\text{th}}}{b} - \left(\sum_{i=1}^{m-1} R_e\left(\mathbf{\Phi}^i, {p_e}^i, I_e^i\right) + (M-m)\, r_e(\mathbf{\Phi}, p_e)\right). \tag{3.11}$$

The inequality in (3.10) guarantees that the eMBB loss at each mini-time slot, resulting from puncturing the eMBB RBs and adjusting the RIS phase-shift matrix, does not impact the eMBB rate requirements r_{th} for the entire M mini-time slots.

The URLLC allocation problem aims at maximizing the admitted URLLC packets while minimizing the eMBB rate loss at each mini-time slot while guaranteeing the target QoS of the eMBB users and the rate requirements of the URLLC packets. For all $m \in \mathcal{M}$, the number of the admitted URLLC packets at the mth mini-time slot (which we aim to maximize) can be expressed as

$$f_1(\mathbf{k}^m) = \sum_{l=1}^{L^m} k_l^m, \tag{3.12}$$

where $\mathbf{k}^m = [k_1^m, k_2^m, ..., k_{L^m}^m]^T$ is a $L^m \times 1$ binary vector that represents the admission of the URLLC packets, i.e., for all $l \in \{1, 2, \ldots, L^m\}$, if $k_l^m = 1$, then the lth URLLC packet is admitted, and $k_l^m = 0$, then it is not admitted. Based on this, the admission rate of the URLLC load, denoted by η, is defined as the total number of URLLC packets that are successfully served at each time slot divided by the total number of arrived URLLC packets at the same time slot, i.e., $\eta = \frac{\sum_m^M \hat{L}^m}{\sum_m^M L^m}$, where, for all $m \in \mathcal{M}$, $\hat{L}^m$ is the number of served URLLC packets at mini-time slot m.

On the other hand, for all $m \in \mathcal{M}$, the overall eMBB traffic rate loss (which we aim to minimize) can be expressed as

$$f_2(\mathbf{I}^m) = \sum_{e=1}^{E_f} I_e^m \beta_e^{\pi,m}, \tag{3.13}$$

where for all $e \in \mathcal{E}$, $\beta_e^{\pi,m} \in [0, 1]$ is the weight of allocating the URLLC load on the eth eMBB user at mini-time slot m, which depends on the URLLC allocation strategy, denoted by π. The objective function f_2 is a weighted sum of the number of frequency resources to be punctured $(I_e^m)_{1\leq e\leq E_f}$, and hence, it is a convex function. Therefore, the optimal minimization strategy of the objective function f_2 is the one that punctures a higher number of frequency resources from the eMBB users with lower weights, i.e., for all $(i,j)\in\mathcal{E}_f$, if $\beta_i^{\pi,m}\leq\beta_j^{\pi,m}$, then $I_i^m\geq I_j^m$. At this stage, one might think on how to efficiently design weight of allocating the URLLC load at each mini-time slot m, for all $m\in\mathcal{M}$. In practice, several URLLC allocation strategies, such as random allocation, minimum eMBB rate loss and proportional fairness can be adopted to distribute/control the eMBB loss [3]. For all $m\in\mathcal{M}$, let $\boldsymbol{\beta}^{m,\pi}=\left[\beta_1^{m,\pi},\beta_2^{m,\pi},\ldots,\beta_{E_f}^{m,\pi}\right]^T$ denote the $E_f\times 1$ vector of puncturing weights of the allocation strategy π, which can be illustrated, when the URLLC allocation strategy π is the minimum eMBB rate loss and the proportional fairness URLLC, as follows:

1. *Minimum eMBB Rate Loss (MeRL):* This strategy aims at minimizing the eMBB rate loss. Based on the observation above, the BS allocates lower weight to the eMBB users that are susceptible to have low rate losses, i.e., the users with low achievable data rates compared to the eMBB users with high achievable data rates. Hence, this strategy aims to minimize the eMBB rate loss by puncturing the eMBB users with low achievable rates. Thus, for all $e\in\mathcal{E}_f$, $\beta_e^{m,\pi}$ is expressed as
$$\beta_e^{m,\pi}=\frac{\hat{R}_e^m}{\sum_{e=1}^{E}\hat{R}_e^m},\tag{3.14}$$
where $\hat{R}_e^m$ is the maximum allowed rate loss for the eth eMBB user at mini-time slot m, and it is expressed as
$$\begin{aligned}\hat{R}_e^m&=r_e\left(\boldsymbol{\Phi}^m,p_e\right)-r_{e,\text{th}}^{'}\\&=\sum_{i=1}^{m-1}R_e\left(\boldsymbol{\Phi}^i,p_e,I_e^i\right)+\left(M-m+1\right)r_e\left(\boldsymbol{\Phi},p_e\right)\\&\quad-\frac{M\times r_{th}}{b}.\end{aligned}\tag{3.15}$$

2. *Proportional Fairness (PF):* This strategy aims at ensuring certain fairness between the eMBB users. Based on the observation above, the BS allocates lower weight to the eMBB users with high data rates compared to the ones with lower data rates. This scheme encourages the puncturing of eMBB users

with higher rates. Consequently, for all $e \in \mathcal{E}_f$, the weight $\beta_e^{m,\pi}$ is expressed as

$$\beta_e^{m,\pi} = 1 - \frac{\hat{R}_e^m}{\sum_{e=1}^{E} \hat{R}_e^m}. \tag{3.16}$$

Based on the above discussion, and for all $m \in \mathcal{M}$, the URLLC allocation problem is formulated at the mth mini-time slot as

$$\begin{aligned}
\mathcal{P}_2^m : \quad & \max_{\mathbf{\Phi}^m, \mathbf{p}_{\mathrm{L}}^m, \mathbf{k}^m, \mathbf{I}^m} [f_1(\mathbf{k}^m),\ -f_2(\mathbf{I}^m)] && \text{(3.17a)}\\
& \text{s.t. } I_l^m \, c_l(p_l^m, \mathbf{\Phi}^m) \geq k_l^m c_{th},\ \forall l \in \mathcal{L}^m, && \text{(3.17b)}\\
& R_e\left(\mathbf{\Phi}^m, p_e, I_e^m\right) \geq r'_{e,\mathrm{th}}, \qquad \forall e \in \mathcal{E}_f, && \text{(3.17c)}\\
& \sum_{e=1}^{E_f} (b - I_e^m) p_e + \sum_{l=1}^{L^m} p_l^m I_l^m \leq P_{\mathrm{BS}}, && \text{(3.17d)}\\
& k_l^m \in \{0, 1\}, \qquad \forall l \in \mathcal{L}^m, && \text{(3.17e)}\\
& \frac{I_l^m}{B} \leq k_l^m \leq I_l^m, \qquad \forall l \in \mathcal{L}^m, && \text{(3.17f)}\\
& I_{e,l}^m \in \{0, 1, \ldots, b\}, \qquad \forall e \in \mathcal{E}_f, \forall l \in \mathcal{L}^m, && \text{(3.17g)}\\
& 0 \leq \phi_n < 2\pi, \qquad \forall n \in \mathcal{N}, && \text{(3.17h)}
\end{aligned}$$

where $\mathbf{I}^m = (I_{e,l})_{\substack{1 \leq e \leq E_f \\ 1 \leq l \leq L^m}}$ and $\mathbf{p}_{\mathrm{L}}^m = [p_1^m, p_2^m, ..., p_{L^m}^m]^T$. Based on this formulation, and for all $m \in \mathcal{M}$, the URLLC allocation problem at mini-time slot m consists of obtaining the optimal power and frequency resource allocation for the URLLC packets, which are represented by $\mathbf{p}_{\mathrm{L}}^m$ and $\mathbf{I}^m$, respectively, along with the optimal RIS phase-shift matrix $\mathbf{\Phi}^m$ at each mini-time slot m. Constraint (3.17b) represents the QoS requirement of the URLLC packets. The eMBB QoS constraints are guaranteed in (3.17c). Constraint (3.17d) indicates that the allocated eMBB and URLLC power should not exceed the total BS power. For all $l \in \mathcal{L}^m$, constraint (3.17e) indicates that the admission variable k_l is binary, whereas constraint (3.17f) guarantees that, if the lth URLLC packet is allocated, i.e., $k_l^m = 1$, then its allocated resources must be greater than zero, i.e., $I_l^m > 0$. Constraint(3.17g) indicates that, for all $e \in \mathcal{E}_f$, the punctured resources of the eth eMBB user should not exceed its total frequency resources b.

3.4 Solution Roadmap

As illustrated in the previous section, the eMBB allocation problem $\mathcal{P}_1$ aims at maximizing the eMBB sum rate over an entire time slot, whereas for all $m \in \mathcal{M}$ the URLLC allocation problem $\mathcal{P}_2^m$ aims at jointly maximizing the number of admitted URLLC packets and minimizing the eMBB rate loss at each mini-time slot m, while maintaining the QoS of the URLLC and eMBB services. For all $m \in \mathcal{M}$, the URLLC allocation problem $\mathcal{P}_2^m$ is a mini-time slot basis problem which should be solved whenever the URLLC load exists. Although optimizing the RIS phase-shift matrix at each mini-time slot is going to provide optimal solution for both eMBB and URLLC traffic, this approach may impact the URLLC latency requirements. Specifically, for all $m \in \mathcal{M}$, problem $\mathcal{P}_2^m$ is a very complex problem, and it should be solved within less than one millisecond. Now, although sub-optimal solutions for problem $\mathcal{P}_2^m$ may be attained using iterative methods for all $m \in \mathcal{M}$, the iterative methods usually need high computational time which could exceed the URLLC latency constraint.

3.4.1 Methodology

In this section, we propose an efficient approach to alleviate the high computational complexity of optimizing the RIS phase-shift matrix per mini-time slot. The main idea is to move the RIS phase-shift matrix optimization, that will be used to jointly serve the eMBB and URLLC services at each mini-time slot, at the beginning of the associated time-slot. In other words, as shown in Fig. 3.2, the RIS phase-shift matrix will be proactively designed at the beginning of the time-slot in a way that can possibly satisfy the requirements of both the existing eMBB traffic and the upcoming URLLC traffic at each mini-time slot.[4] Hence, the pre-configured RIS phase-shift matrix can be employed directly at each mini-time slot whenever a URLLC traffic is present. Accordingly, the BS just sends control signals to the RIS to switch between the pre-computed phase-shift matrices [68–70].[5] [6] Now, the question that arises here is the following. *"Without prior knowledge of the upcoming URLLC traffic, based on*

[4] The pre-configured phase shift matrices are optimized in parallel, hence there is no extra delay effect the real-time configuration of the RIS.

[5] Recent work proposed multiple prototypes of reconfigurable meta-surfaces, and they showed that the RIS could be configured in real-time.

[6] The BS sends a small control signal to switch between the configuration. Hence, the overhead associated with the switching mechanism is low.

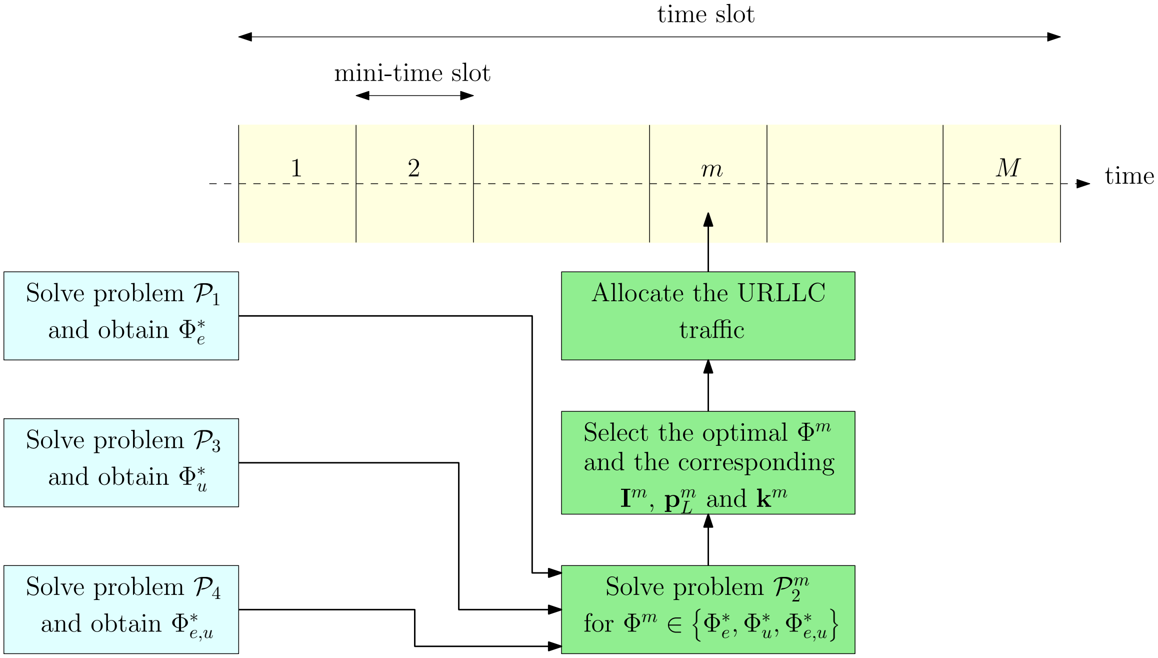

time slot
mini-time slot
1
2
m
M
time
Solve problem $\mathcal{P}_1$ and obtain Φ_e^*
Solve problem $\mathcal{P}_3$ and obtain Φ_u^*
Solve problem $\mathcal{P}_4$ and obtain $\Phi_{e,u}^*$
Allocate the URLLC traffic
Select the optimal Φ^m and the corresponding $\mathbf{I}^m$, $\mathbf{p}_L^m$ and $\mathbf{k}^m$
Solve problem $\mathcal{P}_2^m$ for $\Phi^m \in \{\Phi_e^*, \Phi_u^*, \Phi_{e,u}^*\}$

Fig. 3.2: Proposed methodology

which criteria the RIS phase-shift matrix will be optimized?" To answer this question, we propose in the following three different approaches that can be used.

3.4.2 eMBB RIS Phase-Shift Matrix $\mathbf{\Phi}_{\mathrm{e}}^*$

This approach consists of using the optimal phase-shift matrix $\mathbf{\Phi}_{\mathrm{e}}^*$ that is obtained from solving the eMBB allocation problem $\mathcal{P}_1$ at each mini-time slot, even when the URLLC traffic does exist. However, it is not straightforward to solve problem $\mathcal{P}_1$ directly due to the non-convexity of its objective and and constraints, as well as the high coupling between the transmit power and the phase-shift matrix. With the aid of alternating optimization (AO), problem $\mathcal{P}_1$ is decomposed into two sub-problems, a power allocation sub-problem and an RIS phase-shift matrix optimization sub-problem, which are solved alternately [71]. These two sub-problems are detailed next.

For a fixed RIS phase-shift matrix, problem $\mathcal{P}_1$ is reduced to a power allocation problem that can be formulated as

$$\mathcal{P}_{1,1}: \quad \max_{\mathbf{p}_{\mathrm{e}}} \sum_{e=1}^{E_f} r_e\left(\mathbf{\Phi}, p_e\right) \tag{3.18a}$$

$$\text{s.t.} \quad (1-\delta)\, b\, r_e(\mathbf{\Phi}, p_e) \geq r_{\text{th}}, \quad \forall\, e \in \mathcal{E}_f, \tag{3.18b}$$

$$\sum_{e=1}^{E_f} b\, p_e \leq P_{\text{BS}}. \tag{3.18c}$$

Based on this formulation, problem $\mathcal{P}_{1,1}$ is a convex optimization problem that can be easily solved by applying the KKT condition. On the other hand, for a feasible power solution $\mathbf{p}_{\text{e}}$, the phase-shift optimization sub-problem can be written as

$$\mathcal{P}_{1,2}: \quad \max_{\mathbf{\Phi}} \sum_{e=1}^{E_f} r_e\left(\mathbf{\Phi}, p_e\right) \tag{3.19a}$$

$$\text{s.t.} \quad (1-\delta)\, b\, r_e(\mathbf{\Phi}, p_e) \geq r_{\text{th}}, \quad \forall\, e \in \mathcal{E}_f, \tag{3.19b}$$

$$0 \leq \phi_n < 2\pi, \qquad \forall n \in \mathcal{N}. \tag{3.19c}$$

Due to non convexity of its objective function and its constraints, problem $\mathcal{P}_{1,2}$ is a non-convex problem. In order to tackle this challenge, the SDR technique along with the Gaussian randomization are applied [71]. The details of the solution approach of problem $\mathcal{P}_{1,2}$ are provided in Appendix A.1.

3.4.3 URLLC RIS Phase-Shift Matrix $\mathbf{\Phi}_{\text{u}}^*$

As discussed above, the goal of configuring the RIS is to add a degree of freedom for the BS to improve the URLLC reliability during each mini-time slot. Accordingly, this approach consists of exploiting the CSI of the coexisting URLLC users in designing the RIS phase-shift matrix. As such, the RIS phase-shift matrix $\mathbf{\Phi}_{\text{u}}^*$ is designed at the beginning of the time slot with the aim of enhancing the channel gains of all URLLC users in the network. Therefore, the problem of designing the RIS phase-shift matrix $\mathbf{\Phi}_{\text{u}}^*$ has the objective of maximizing the minimum URLLC channel gain, which can be formulated as

$$\mathcal{P}_3: \max_{\mathbf{\Phi}_{\text{u}}} \min_{1 \leq u \leq U} \left| g_{\text{BS},u} + \mathbf{g}_{\text{RIS},u}^H \mathbf{\Phi}_{\text{u}}\, \mathbf{f}_{\text{BS,RIS}} \right|^2 \tag{3.20a}$$

$$\text{s.t.} \quad 0 \leq \phi_n < 2\pi, \quad \forall n \in \mathcal{N}. \tag{3.20b}$$

The solution of $\mathcal{P}_3$ is detailed in Appendix A.2.

3.4.4 Joint URLLC-eMBB RIS Phase-Shift Matrix $\mathbf{\Phi}^*_{\text{e,u}}$

The phase-shift matrix $\mathbf{\Phi}^*_{\text{u}}$ obtained by solving problem $\mathcal{P}_3$ enhances the performance of the URLLC traffic. However, this may highly impact the performance of the eMBB traffic. Alternatively, the use of a unified RIS phase-shift matrix that can improve the performance of the URLLC load while reducing the degradation on the performance of the eMBB traffic is highly desired. Similar to problem $\mathcal{P}_3$, an RIS configuration, represented by the phase-shift matrix configuration $\mathbf{\Phi}^*_{\text{e,u}}$ is designed at the beginning of the time slot with the aim of enhancing the channel gains of all coexisting eMBB and URLLC users in the network. Accordingly, the problem is formulated as maximizing the minimum channel gain of all coexisting eMBB and URLLC users, which is formulated as

$$\mathcal{P}_4 : \max_{\mathbf{\Phi}_{\text{e,u}}} \min_{1 \leq x \leq E_f + U} |t_{\text{BS},x} + \mathbf{t}^H_{\text{RIS},x} \mathbf{\Phi}_{\text{e,u}} \, \mathbf{f}_{\text{BS,RIS}}|^2 \tag{3.21a}$$

$$\text{s.t} \quad 0 \leq \phi_n < 2\pi, \quad \forall n \in \mathcal{N}. \tag{3.21b}$$

where $t_{\text{BS},x} = h_{\text{BS},x}$, if $x \in \mathcal{E}_f$ and $t_{\text{BS},x} = g_{\text{BS},x-E_f}$ if $x \in \{E_f + 1, \ldots, E_f + U\}$, and $\mathbf{t}_{\text{RIS},x} = \mathbf{h}_{\text{RIS},x}$, if $x \in \mathcal{E}_f$ and $\mathbf{t}_{\text{RIS},x} = \mathbf{g}_{\text{RIS},x-E_f}$ if $x \in \{E_f + 1, \ldots, E_f + U\}$. The solution of problem $\mathcal{P}_4$ is detailed in Appendix A.2.

3.4.5 URLLC Allocation

The optimization problems $\mathcal{P}_1$, $\mathcal{P}_3$ and $\mathcal{P}_4$ will be solved in parallel at the beginning of each time slot. Once the three RIS configurations $\left\{\mathbf{\Phi}^*_{\text{e}}, \mathbf{\Phi}^*_{\text{u}}, \mathbf{\Phi}^*_{\text{e,u}}\right\}$ are obtained at the beginning of each time-slot, they will be communicated to the RIS controller. Afterwards, within each mini-time slot, a joint power/resource allocation problems for URLLC packets is solved for each candidate RIS phase-shift matrix in the set $\left\{\mathbf{\Phi}^*_{\text{e}}, \mathbf{\Phi}^*_{\text{u}}, \mathbf{\Phi}^*_{\text{e,u}}\right\}$. Specifically, for all $m \in \mathcal{M}$, the URLLC allocation problem $\mathcal{P}^m_2$ will be solved when the RIS phase-shift matrix $\mathbf{\Phi}^m$ is one of the pre-computed RIS phase-shift matrices, i.e., $\mathbf{\Phi}^m \in \left\{\mathbf{\Phi}^*_{\text{e}}, \mathbf{\Phi}^*_{\text{u}}, \mathbf{\Phi}^*_{\text{e,u}}\right\}$. For each of these three cases, the URLLC allocation problem is reduced to a simple joint power/frequency allocation problem and the resulting three optimization problems can solved in a parallel. After doing so, the BS selects the best RIS configuration $\mathbf{\Phi}^{m^*} \in \left\{\mathbf{\Phi}^*_{\text{e}}, \mathbf{\Phi}^*_{\text{u}}, \mathbf{\Phi}^*_{\text{e,u}}\right\}$, along with the corresponding optimal power and frequency allocation policies for the URLLC

Algorithm 3.1: Proposed Algorithm

```
for $\mathbf{\Phi}^m \in \left\{\mathbf{\Phi}^*_{\rm e}, \mathbf{\Phi}^*_{\rm u}, \mathbf{\Phi}^*_{\rm e,u}\right\}$ do
    Solve problem $\mathcal{P}_2^m$ and get the optimal $\mathbf{p}_{\rm L}^{m*}(\mathbf{\Phi}^m), \mathbf{k}^{m*}(\mathbf{\Phi}^m), \mathbf{I}^{m*}(\mathbf{\Phi}^m)$;
end
- $\mathbf{\Phi}^{m*} = \arg\max_{\mathbf{\Phi}^m} \Big\{\mathbf{k}^{m*}(\mathbf{\Phi}^m) |\, \mathbf{\Phi}^m \in \left\{\mathbf{\Phi}^*_{\rm e}, \mathbf{\Phi}^*_{\rm u}, \mathbf{\Phi}^*_{\rm e,u}\right\} \Big\}$;
- $\mathbf{p}_{\rm L}^{m*} = \mathbf{p}_{\rm L}^{m*}(\mathbf{\Phi}^{m*})$, $\mathbf{k}^{m*} = \mathbf{k}^{m*}(\mathbf{\Phi}^{m*})$, and $\mathbf{I}^{m*} = \mathbf{I}^{m*}(\mathbf{\Phi}^{m*})$;
```

traffic, such that the admitted URLLC packets is maximized. Afterwards, at each mini-time slot m, for all $m \in \mathcal{M}$, the BS sends a control signal to the RIS in order to switch the RIS configuration to the best obtained configuration $\mathbf{\Phi}^{m^*}$. The overall URLLC allocation procedure in each mini-time slot is presented in Algorithm 3.1 in this chapter. The remaining now is how to obtain the optimal frequency and power allocation policies for the URLLC traffic at each mini-time slot when the RIS phase-shift matrix is fixed, which is detailed in the following section.

3.5 URLLC Allocation Problem for Fixed RIS Phase-Shift Configuration

3.5.1 URLLC Resource Allocation

In this section, the RIS phase-shift matrix $\mathbf{\Phi}^m$, for all $m \in \mathcal{M}$, is fixed, i.e., $\mathbf{\Phi}^m \in \left\{\mathbf{\Phi}^*_{\rm e}, \mathbf{\Phi}^*_{\rm u}, \mathbf{\Phi}^*_{\rm e,u}\right\}$. In this case, for all $m \in \mathcal{M}$, the URLLC allocation problem $\mathcal{P}_2^m$ is reduced to a joint power/frequency allocation problem. To simplify problem $\mathcal{P}_2^m$, we consider disjoint optimization integer vectors $\mathbf{I}^m_{E_f} = \left[I_1^m, I_2^m, \dots, I^m_{E_f}\right]$ and $\mathbf{I}^m_L = [I_1^m, I_2^m, \dots, I^m_{L^m}]$ for the punctured eMBB RBs and the RBs allocated to the URLLC packets, respectively, i.e, using the change of variables $I_l^m = \sum_{e=1}^{E_f} I^m_{e,l}$ and $I_e^m = \sum_{l=1}^{L^m} I^m_{e,l}$, for all $e \in \mathcal{E}_f$ and $l \in \mathcal{L}^m$. Then, problem $\mathcal{P}_2^m$ can be reformulated as

$$\mathcal{P}^m_{2,1}: \quad \max_{\mathbf{p}_{\rm L}^m, \mathbf{I}^m_{E_f}, \mathbf{I}^m_L, \mathbf{k}^m} \; [f_1(\mathbf{k}^m), \, -f_2(\mathbf{I}^m_{E_f})], \tag{3.22a}$$

$$\text{s.t.} \quad (3.17\text{b}) - (3.17\text{f}), \tag{3.22b}$$

$$I_e^m \in \{0, 1, \dots, b\}, \; \forall e \in \mathcal{E}_f, \tag{3.22c}$$

$$I_l^m \in \{0,1,\ldots,B\},\ \forall l \in \mathcal{L}^m, \tag{3.22d}$$

$$\sum_{l=1}^{L^m} I_l^m = \sum_{e=1}^{E_f} I_e^m. \tag{3.22e}$$

Constraint (3.22e) guarantees that the number of punctured eMBB resources is equal to the number the resources allocated to the URLLC traffic. Problem $\mathcal{P}_{2,1}^m$ is a mixed integer non-linear problem (MINLP), and it is a non-convex problem due to the non-convex constraints (3.17b) and (3.17c). Using the fact that the power allocated to the eMBB users are already optimized at the beginning of the time slot, then $\sum_{e=1}^{E_f} b\,p_e = P_{\text{BS}}$. Hence, for all $m \in \mathcal{M}$ and $e \in \mathcal{E}_f$, by manipulating constraints (3.17c) and (3.22c), one can reach that the maximum available RBs for allocating URLLC load over eMBB user e at mini-time slot m is

$$I_e^{\max,m} = \min\left(\left\lfloor b\left(1 - \frac{r_{e,\text{th}}^{'}}{r_e\left(\mathbf{\Phi}^m, p_e\right)}\right)\right\rfloor, b\right), \tag{3.23}$$

where $\forall x \in \mathcal{R}$, $\lfloor x \rfloor$ represents the greatest integer less than or equal x. From constraint (3.17b), one can see that, the optimal power allocation for the lth URLLC packet given a fixed number of allocated RBs I_l, for all $l \in \mathcal{L}^m$ and $m \in \mathcal{M}$, is given by

$$p_l^{m*}(I_l^m) = \frac{\left(e^{\frac{k_l^m c_{\text{th}}}{\log(2) I_l^m}} - 1\right)}{\alpha_l^m}, \tag{3.24}$$

where $\alpha_l^m = \frac{|g_{\text{BS},l} + \mathbf{g}_{\text{RIS},l}^H \mathbf{\Phi}^m\, \mathbf{f}_{\text{BS,RIS}}|^2}{\sigma^2 \Gamma_{\text{URLLC}}}$. Based on this, for all $m \in \mathcal{M}$, problem $\mathcal{P}_{2,1}^m$ can be equivalently transformed to

$$\mathcal{P}_{2,2}^m: \quad \max_{\mathbf{I}_{E_f}^m, \mathbf{I}_L^m, \mathbf{k}^m} \ [f_1(\mathbf{k}^m),\ -f_2(\mathbf{I}_{E_f}^m)], \tag{3.25a}$$

$$\text{s.t.} \quad (3.17\text{e}) - (3.17\text{f}), (3.22\text{c}) - (3.22\text{e}) \tag{3.25b}$$

$$\sum_{l=1}^{L^m} I_l^m p_l^{m*}(I_l^m) \le \sum_{e=1}^{E_f} I_e^m p_e,\ \forall l \in \mathcal{L}^m. \tag{3.25c}$$

Algorithm 3.2: Optimization-based URLLC allocation algorithm

```
- **Solve** problem $\mathcal{P}_{2,4}^{m}$and get the optimal $\mathbf{k}^{m*}$ and the corresponding
  $\hat{\mathcal{L}}^{m} \subset \mathcal{L}^{m}$ ;
- **Solve** problem $\mathcal{P}_{2,5}^{m}$ for and get the optimal $\mathbf{I}_{L}^{m*}$, $\mathbf{I}_{E_f}^{m}{}^{*}$, ;
- **Evaluate** the optimal $\mathbf{p}_{L}^{m*}$ from (3.24);
- **Evaluate** $\mathbf{I}^{m*}$ by distributing $\mathbf{I}_{L}^{m*}$ over $\mathbf{I}_{E_f}^{m}{}^{*}$;
```

The proposed solutions of $\mathcal{P}_{2,2}^{m}$ are provided in the following subsection. Specifically, we proposed two algorithms for the URLLC allocation problem: an optimization-based algorithm and a Heuristic algorithm. The Heuristic algorithm aims to overcome the computational complexity of the optimization-based algorithm such the URLLC latency requirements are met.

3.5.2 Proposed Algorithms

3.5.2.1 Optimization-based URLLC allocation

For all $m \in \mathcal{M}$, problem $\mathcal{P}_{2,2}^{m}$ is a multi objective problem which aims at concurrently maximizing the number of admitted URLLC packets and minimizing the eMBB rate loss at mini-time slot m. Due to its latency and reliability constraints, it is important to mention here that the objective of maximizing the number of admitted URLLC packets has a higher priority than the one of minimizing the eMBB loss. Hence, an efficient solution is hard to be obtained by solving problem $\mathcal{P}_{2,2}^{m}$ directly. Alternatively, for all $m \in \mathcal{M}$, we decompose problem $\mathcal{P}_{2,2}^{m}$ into two sub-problems, namely, a URLLC admission problem and a URLLC allocation problem. The first sub-problem aims at maximizing the admission of URLLC packets $\sum_{l=1}^{L^m} k_l^m$, and it is expressed as

$$\mathcal{P}_{2,3}^{m}: \quad \max_{\mathbf{I}_{E_f}^{m}, \mathbf{I}_{L}^{m}, \mathbf{k}^{m}} \quad f_1(\mathbf{k}^{m}) \tag{3.26a}$$

$$\text{s.t.} \quad (3.17\text{e}) - (3.17\text{f}), (3.22\text{c}) - (3.22\text{e}), (3.25\text{c}). \tag{3.26b}$$

Once the optimal number of admitted URLLC packets $\mathbf{k}^{m*}$ is obtained at each mini-time slot m, for all $m \in \mathcal{M}$, then the set of the URLLC packets $\hat{\mathcal{L}}^{m} \subseteq \mathcal{L}^{m}$ that can be allocated is defined, i.e., $\hat{\mathcal{L}}^m = \{l \in \mathcal{L}^m | k_l^m = 1\}$. Then, the optimal frequency resources for admitting the URLLC packets, i.e., the optimal frequency resources to be punctured from the eMBB users, are determined based on the chosen allocation strategy π. The associated optimization problem can be written as

$$\mathcal{P}_{2,4}^m : \quad \min_{\mathbf{I}_{E_f}^m, \mathbf{I}_L^m} \quad f_2(\mathbf{I}_{E_f}^m) \tag{3.27a}$$

$$\text{s.t.} \quad (3.22\text{c}) - (3.22\text{e}), (3.25\text{c}). \tag{3.27b}$$

Although problems $\mathcal{P}_{2,3}^m$ and $\mathcal{P}_{2,4}^m$ can be solved optimally using an integer optimization solver [46], the computational time is very high. Hence, we can solve these problems by relaxing the integer variables $\mathbf{I}_L^m$ and $\mathbf{I}_{E_f}^m$ to be continuous. Then, the resulting solutions are rounded to get the optimal integer values $\mathbf{I}_L^{m*}$ and ${\mathbf{I}_{E_f}^m}^*$. Hence, the optimal $\mathbf{p}_L^{m*}$ is evaluated from (3.24). Based on the above, the solution approach of problem $\mathcal{P}_{2,1}^m$ is summarized in **Algorithm 2**, which refers to Algorithm 3.2. Finally, by distributing the URLLC packets over the punctured eMBB RBs ${\mathbf{I}_{E_f}^m}^*$ based on their required RBs $\mathbf{I}_L^{m*}$, the solution of the original integer problem $\mathcal{P}_2^m$ can be obtained.

3.5.2.2 Heuristic URLLC allocation algorithm

The URLLC traffic requires strict latency requirements which is less than 1 msec, and hence, a low complexity algorithm is essential to allocate the incoming URLLC packets. Accordingly, in this part, we develop a low complexity Heuristic algorithm that exploits two observations. First, because the channel gains of the URLLC users are known for fixed RIS phase-shift matrix $\mathbf{\Phi}^m$, the URLLC packets with good channel conditions need less frequency and power resources than those with bad channel conditions. Accordingly, allocating the URLLC packets with good channel conditions before those with bad channel conditions will enhance the URLLC admission rate. Second, as was shown in section 3.3, for all $m \in \mathcal{M}$, the eMBB users with low allocation weights have a higher chance to be punctured than those with high allocation weights. Accordingly, the proposed approach starts by sorting the URLLC packets in descending order based on their channel conditions. Specifically, at each mini-time slot m, for all $m \in \mathcal{M}$, and when the RIS phase-shift matrix $\mathbf{\Phi}^m$

```
Algorithm 3.3: Heuristic URLLC allocation algorithm
Initiate $\mathbf{I}^m$ and $\mathbf{p}_L^m$;
Evaluate $\mathbf{I}^{\max,m}$ and $\boldsymbol{\beta}^{m,\pi}$;
Sort URLLC users based on the channel gain in descending order;
Sort eMBB users based on $\boldsymbol{\beta}^{m,\pi}$in ascending order ;
for  $l = 1 \rightarrow L^m$ do
    $c_l^{temp} = 0$;
    Boolean=0,  URLLC binary variable variable;
    for $e = 1 \rightarrow E_f$ do
        $I_{e,l}^{temp} = \lceil \frac{c_{th} - c_l^{temp}}{c_l(p_e)} \rceil$;
        if $I_{e,l}^{temp} \leq I_e^{\max,m}$ then
            $I_{e,l}^m = I_{e,l}^{temp}$;
            Boolean=1; break;
        else
            $I_{e,l}^m = I_e^{\max,m}$;
            $c_l^{temp} = c_l^{temp} + I_{e,l}^m * c_l(p_e)$;
    if boolean=1 then
        $p_l = \frac{\sum_{e=1}^{E_f} I_{e,l} p_e}{\sum_{e=1}^{E_f} I_{e,l}}$;
        update $\mathbf{I}^{\max,m}$;
    else
        break;
```

is fixed, let $g_l^m = |g_{\text{BS},l} + \mathbf{g}_{\text{RIS},l}^H \boldsymbol{\Phi}^m \, \mathbf{f}_{\text{BS,RIS}}|^2$ be the channel gain of the URLLC user associated to the lth URLLC packet, for $l \in \mathcal{L}^m$. Then, the sorted channel gain of the URLLC users can be expressed as $\mathbf{g}^m = \{g^m_{(1)}, g^m_{(2)}, \dots, g^m_{(L^m)}\}$ where, for all $l \in \mathcal{L}^m$, $g^m_{(l)}$ is the lth highest channel gain. Similarly, the eMBB users are sorted in a ascending order based on their allocation weights of the URLLC load allocation as $\hat{\boldsymbol{\beta}}^{m,\pi} = \{\beta^{m,\pi}_{(1)}, \beta^{m,\pi}_{(2)}, \dots, \beta^{m,\pi}_{(E_f)}\}$ where, for all $e \in \mathcal{E}_f$, $\beta^{m,\pi}_{(l)}$ is the lth lowest allocation weight. Afterwards, for all $l \in \mathcal{L}^m$, the algorithm assumes zero data rate c_l for the lth URLLC packet. Then, the proposed approach allocates resources to the lth URLLC packet by iterating over the ordered eMBB users and considering that the power allocated to lth URLLC packet on the punctured RB is equal to that of the eMBB user within each iteration. The algorithm continues the iterative procedure until the rate requirement of the lth URLLC packet is satisfied at a given eMBB user. Precisely, for all $l \in \mathcal{L}^m$, the algorithm iterates over the eMBB users and checks if the cumulative data rate of the lth URLLC packet at each eMBB user is higher than the rate threshold c_{th}, i.e.,

$$\sum_{i=1}^{e} I_{i,l}^{m} c_l(p_i) = \sum_{i=1}^{e} I_{i,l}^{m} \times \log(1 + g_l\, p_i I_{i,l}^{m}) \geq c_{\text{th}}, \tag{3.28}$$

where e is the index of the eMBB user under checking. Once the frequency resources $\left(I_{i,l}^{m}\right)_{1 \leq i \leq e}$ are allocated to $l$the URLLC packet, the power that will be allocated to the lth URLLC packet per RB is given by $p_l^m = \frac{\sum_{i=1}^{e} I_{i,l}^{m} p_e}{\sum_{i=1}^{e} I_{i,l}^{m}}$. This is basically due to the facts that the condition in (3.28) guarantees that the URLLC rate requirement is satisfied and that the rate function is concave, i.e.,

$$\begin{aligned} \log\left(1 + g_l p_l^m\right) \sum_{i=1}^{e} I_{i,l} &= \log\left(1 + g_l \frac{\sum_{i=1}^{e} p_i I_{i,l}^{m}}{\sum_{i=1}^{e} I_{i,l}^{m}}\right) \\ \sum_{i=1}^{e} I_{i,l} \sum_{i=1}^{e} I_{i,l}^{m} &\times \log(1 + g_l\, p_i I_{i,l}) \geq c_{\text{th}}. \end{aligned} \tag{3.29}$$

These steps are repeated for all URLLC packets while satisfying the eMBB QoS. Otherwise, the URLLC packet is dropped. Finally, **Algorithm 3** (which refers to Algorithm 3.3) presents the details steps of the proposed approach, where it consists of two loops, an inner loop and an outer loop of L^m and E_f iterations, respectively. This algorithm has a polynomial time complexity with respect to the number of URLLC packet L^m and the worst-case time complexity of the algorithm is $O(L^m \times E_f)$.

3.6 Simulation Results

3.6.1 Simulation Settings

In this section, we perform various simulations to evaluate the performance of the proposed scheme. In this simulation environment, the wireless network consists of one

BS and one RIS. The coverage area of the BS is assumed to be 110 meters. The BS serves $E = 8$ eMBB users and U (in the range $[5, 80]$) URLLC users are distributed uniformly at random over the coverage area of the BS. In order to increase the BS coverage, the RIS is located 20 meters away from it. The number of RIS reflecting elements N is a key parameter in the performance of the proposed scheme. Hence, we vary N in the range $[10, 50]$. Each time slot consists of $M = 7$ mini-time slots. We consider both the large-scale fading and the small-scale fading for all communication links. Particularly, the large scale fading for the direct and the cascaded links are modeled as $P_0 = \alpha_0(d_{\text{BS,e/u}})^{-\varrho_0}$ and $P_1 = \alpha_1(d_{\text{BS,RIS}})^{-\varrho_1}(d_{\text{BS,e/u}})^{-\varrho_2}$, respectively, where $d_{\text{BS,e/u}}$, $d_{\text{BS,RIS}}$ and $d_{\text{BS,RIS}}$ are the distances of BS-users, BS-RIS, and RIS-users links, respectively. $\varrho_0 = 3.5$, $\varrho_1 = 2.2$ and $\varrho_2 = 2.8$ are the path loss exponents of BS-users, BS-RIS, and RIS-users links, respectively. Also, $\alpha_0 = -30$ dB and $\alpha_1 = -40$ dB are the path loss at the reference distance for the direct links and the cascaded links, respectively. On the other hand, the small-scale fading for all channels is modeled as $f = \sqrt{\frac{\kappa}{1+\kappa}} f^{\text{LoS}} + \sqrt{\frac{1}{1+\kappa}} f^{\text{NLoS}}$, where κ is the Rician factor, f^{LoS} is the line-of-sight component and f^{NLoS} is non-line-of-sight component, which follows a Rayleigh distribution with a scale parameter equals to one [72]. The communication links between the BS and the cellular users and between the RIS and the cellular users are assumed to have a quasi-static flat-fading Rayleigh channel. The links between the BS and the RIS elements are assumed to have a Rician channel model with a Rician factor $\kappa = 10$ [72]. The remaining system parameters are summarized in Table 3.2 [62, 72–74]. The simulation results are performed over $2 * 10^3$ independent Monte-Carlo realizations on the channel gains of all cellular users.

In order to show the performance of the different possible RIS configurations on the performance of both the URLLC and the eMBB traffic, we introduce the following four schemes:

- Proposed Scheme-1: The RIS phase-shift matrix $\mathbf{\Phi}_{\text{e}}^*$ is used to serve the eMBB and the URLLC during the entire time slot, i.e., at each mini-time slot.
- Proposed Scheme-2: The RIS phase-shift matrix $\mathbf{\Phi}_{\text{u}}^*$ is used whenever a URLLC packet has arrived. Accordingly, at each mini-time slot, if a URLLC packet exists, the RIS configuration $\mathbf{\Phi}_{\text{u}}^*$ is used. Otherwise, the configuration $\mathbf{\Phi}_{\text{e}}^*$ is used.

Table 3.2: Simulation parameters

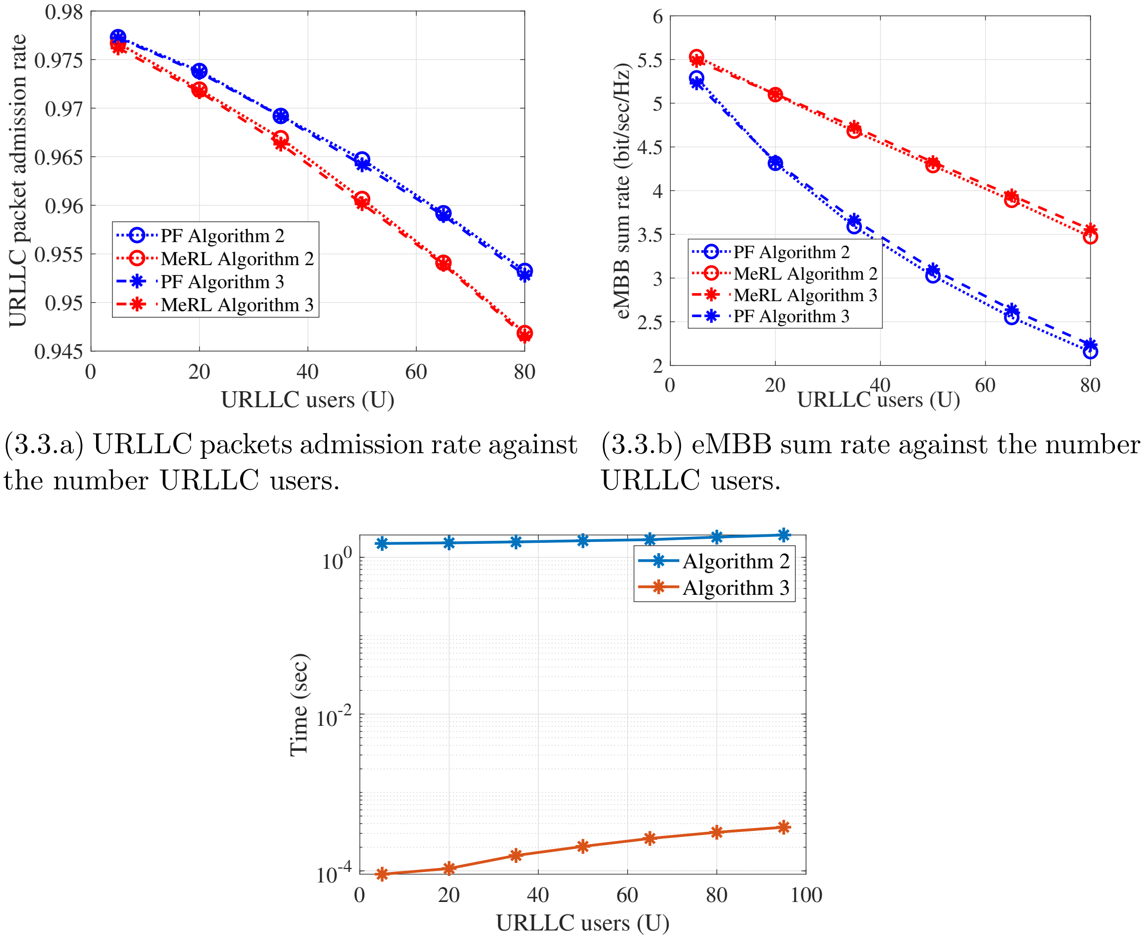
Parameter	Symbol	Value
Power budget at the BS	P_{BS}	33 dBm
Noise power	σ^2	-97.5 dBm
URLLC packet arrival rate	λ_u	0.7 packet/msec
Mini-slot duration	τ	0.143 ms
Number of resource blocks	B	96
Resource block bandwidth	W	180 kHz
eMBB block error probability	ϵ_{eMBB}	10^{-1}
eMBB rate threshold	r_{th}	1 Mb/sec
URLLC block error probability	$\epsilon_{\mathrm{URLLC}}$	10^{-6}
URLLC packet size	ζ	256 bits

- Proposed Scheme-3: The RIS phase-shift matrix $\mathbf{\Phi}^*_{\mathrm{e,u}}$ is used whenever a URLLC packet has arrived. Accordingly, at each mini-time slot, if a URLLC packet exists, the RIS configuration $\mathbf{\Phi}^*_{\mathrm{e,u}}$ is used. Otherwise, the configuration $\mathbf{\Phi}^*_{\mathrm{e}}$ is used.

- Selected RIS configuration: In each mini-time slot, the BS computes the maximum URLLC admitted packets and the lowest eMBB rate loss for the RIS configurations $\mathbf{\Phi}^*_{\mathrm{e}}$, $\mathbf{\Phi}^*_{\mathrm{u}}$ and $\mathbf{\Phi}^*_{\mathrm{e,u}}$. Then, as shown in Algorithm 1, the BS will select the phase-shift matrix that maximizes the number of admitted URLLC packets.

3.6.2 Performance Evaluation of the URLLC Allocation Strategies

Fig.3.3.a and Fig.3.3.b illustrate the performance of Algorithm 2 and Algorithm 3 in terms of both the URLLC packets admission rates and the eMBB sum rate, respectively, while varying the number of URLLC users (URLLC load). We observe that as the URLLC load increases, the URLLC admission rate starts to reduce, which is due to the following. First, the BS has limited power and frequency resources and is required to protect the QoS of the already admitted eMBB; therefore, at each mini slot, it decides which resources can be punctured to admit the incoming URLLC packets. As more packets arrive, the available resources which can be punctured without impacting the eMBB service may not suffice to admit all the URLLC load and



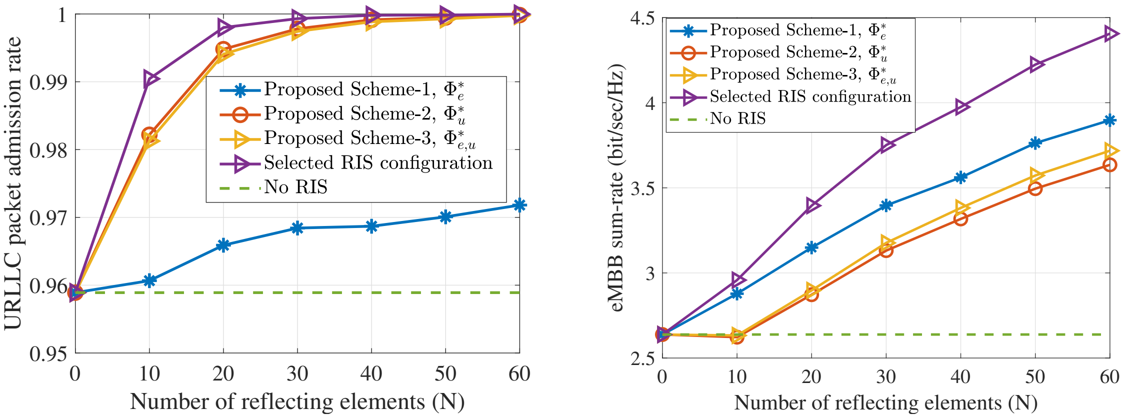
(3.3.a) URLLC packets admission rate against the number URLLC users.

(3.3.b) eMBB sum rate against the number URLLC users.

(3.3.c) Algorithm time complexity vs number of URLLC users.

Fig. 3.3: Performance comparison between Algorithm 2 and Algorithm 3.

therefore some packets get rejected; indeed, both algorithms (and policies) exhibits similar performance decay as observed from the figures. Furthermore, it can be seen from Fig. 3.3.a that the PF URLLC allocation strategy achieves better URLLC packets admission rate than the MeRL allocation strategy. The reason being that by puncturing the eMBB users with low data rates, the eMBB resources available for puncturing become limited to accommodate the high URLLC load. Furthermore, Fig. 3.3.b shows that the MeRL URLLC allocation strategy achieves better eMBB sum rate than that of the PF URLLC allocation. The reason behind that is the punctured resources resulting from the MeRL approach are belonging to the eMBB users with low data rates rather than those with high data rates. Since, the PF achieves better



(3.4.a) URLLC packet admission rate against the number of RIS elements.

(3.4.b) eMBB sum rate against the number of RIS elements.

Fig. 3.4: Performance of the proposed algorithm against the number of RIS elements. $\delta = 0.1$ and $U = 65$

URLLC packets admission rate than the MeRL, we consider the PF for the rest of simulations. On the other hand, Fig. 3.3.a and Fig. 3.3.b show that Algorithm 2 and Algorithm 3 have broadly the same performance in terms of both the URLLC packets admission rates and the eMBB sum rate, respectively. However, Fig. 3.3.c illustrates the run time complexity of Algorithm 2 and Algorithm 3. It can be shown that Algorithm 2 computation time is in the order of one second which can violate the URLLC latency requirements. However, Algorithm 3 has a lower compute time which is around one mini-time slot, 0.143 millisecond. This low operating time makes Algorithm 3 favorable in practice. Moreover, the processing time of the Algorithm 3 can be further improved when using more computing resources available at the network edge.[7] Since, Algorithm 3 has very low time complexity and almost similar performance to Algorithm 3, we consider Algorithm 3 for the rest of our simulations.

3.6.3 Impact of the Number of RIS Elements

Fig. 3.4 depicts the impact of the size of the RIS (N) and the main observations are summarized below.

- It can be shown that the proposed schemes outperform the baseline with no RIS

[7] The algorithm was implemented in Matlab using a machine with the following characteristics: System Type: x64-based PC Processor: Intel(R) i7-4510U CPU @2GHz, 8 Gigabyte RAM.

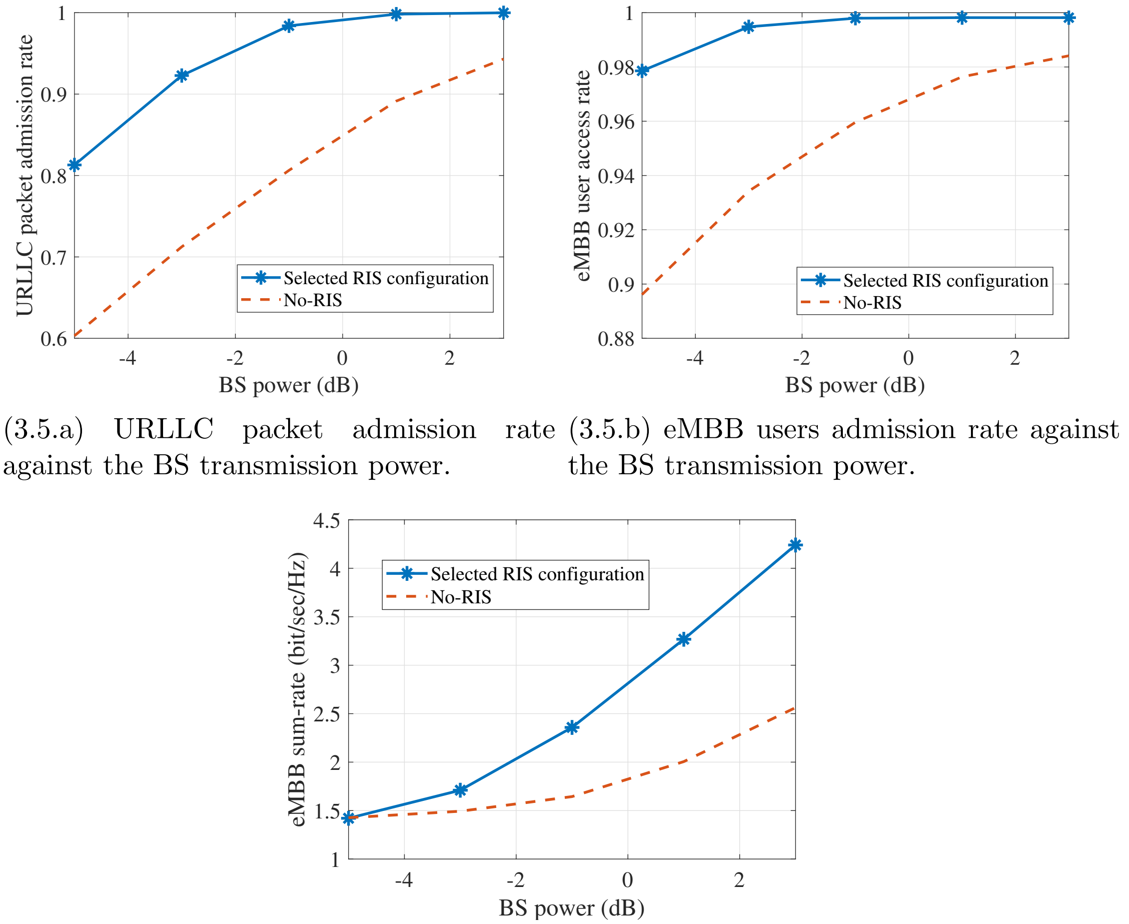
in both the URLLC packets admission rate and the eMBB sum rate. Particularly, as shown in Fig. 3.4.a, the proposed schemes achieve 96.87%, 99.9%, 99.9, and 99.98% URLLC packets admission rate at $N = 40$ compared to 95.6% when the RIS is not deployed. For the same number of reflecting elements $N = 40$, Fig. 3.4.b shows that the proposed schemes achieve enhancement on the eMBB rate around 52%, 37%, 30% and 27% when the RIS is not deployed. The reason behind that is the ability of the BS to select a phase-shift matrix, depending on the number of the URLLC packets and their channel conditions, from the set $\{\mathbf{\Phi}^*_{\mathrm{e}}, \mathbf{\Phi}^*_{\mathrm{u}}, \mathbf{\Phi}^*_{\mathrm{e,u}}\}$ to configure the RIS. Particularly, the selected phase-shift matrix should achieve the best performance in terms of the URLLC packet admission rate and the minimum eMBB loss.[8]

- We can also see that only $N = 60$ RIS elements are enough to achieve 99.99% URLLC packets admission rate along with around 70% enhancement of the eMBB sum rate compared to the case when no RIS is deployed.

- The trade-off between the URLLC packet admission rate and the eMBB sum rate is clear in the behavior of scheme-1, scheme-2 and scheme-3. By enhancing the channels condition of the URLLC traffic, scheme-2 and scheme-3 give the URLLC traffic a higher priority over the eMBB traffic which means better URLLC packet admission rate and more eMBB rate-loss. Conversely, scheme-1 gives the eMBB traffic a higher priority over the URLLC counterpart by using the eMBB phase-shift matrix that was optimized to enhance the eMBB rate. This leads to a better eMBB sum rate and a lower URLLC admission.

- Increasing the number of RIS elements enhances both the URLLC packet admission rate and the eMBB sum rate. This is evident since at higher N, the URLLC channel conditions are enhanced which means less frequency resources are needed to accommodate the URLLC packets, which means better URLLC reliability and less eMBB rate loss. Moreover, the eMBB rate is improved by enhancing the eMBB channel conditions, which means a higher availability for eMBB resources to allocate the URLLC packets. While increasing N, scheme-1 exhibits lower enhancement in URLLC service admission. In other words,

[8] It is important to note here that the achieved gain of the proposed scheme in terms of the URLLC reliability and the eMBB sum-rate comes at the expense of an extra overhead to transmit the different RIS configurations and to switch between them.



(3.5.a) URLLC packet admission rate against the BS transmission power.

(3.5.b) eMBB users admission rate against the BS transmission power.

(3.5.c) eMBB sum rate against the BS transmission power.

Fig. 3.5: Performance of the proposed algorithm against the BS transmission power. $N = 50$ and $U = 65$.

unlike $\mathbf{\Phi}^*_{\mathrm{u}}$ and $\mathbf{\Phi}^*_{\mathrm{e,u}}$, $\mathbf{\Phi}^*_{\mathrm{e}}$ acts as a random phase-shift matrix for the URLLC traffic, i.e., the improvement on the URLLC channel conditions is moderate, and hence more transmission resources are need to accommodate the URLLC packets which means lower reliability [62].

3.6.4 Effect of the BS Transmission Power

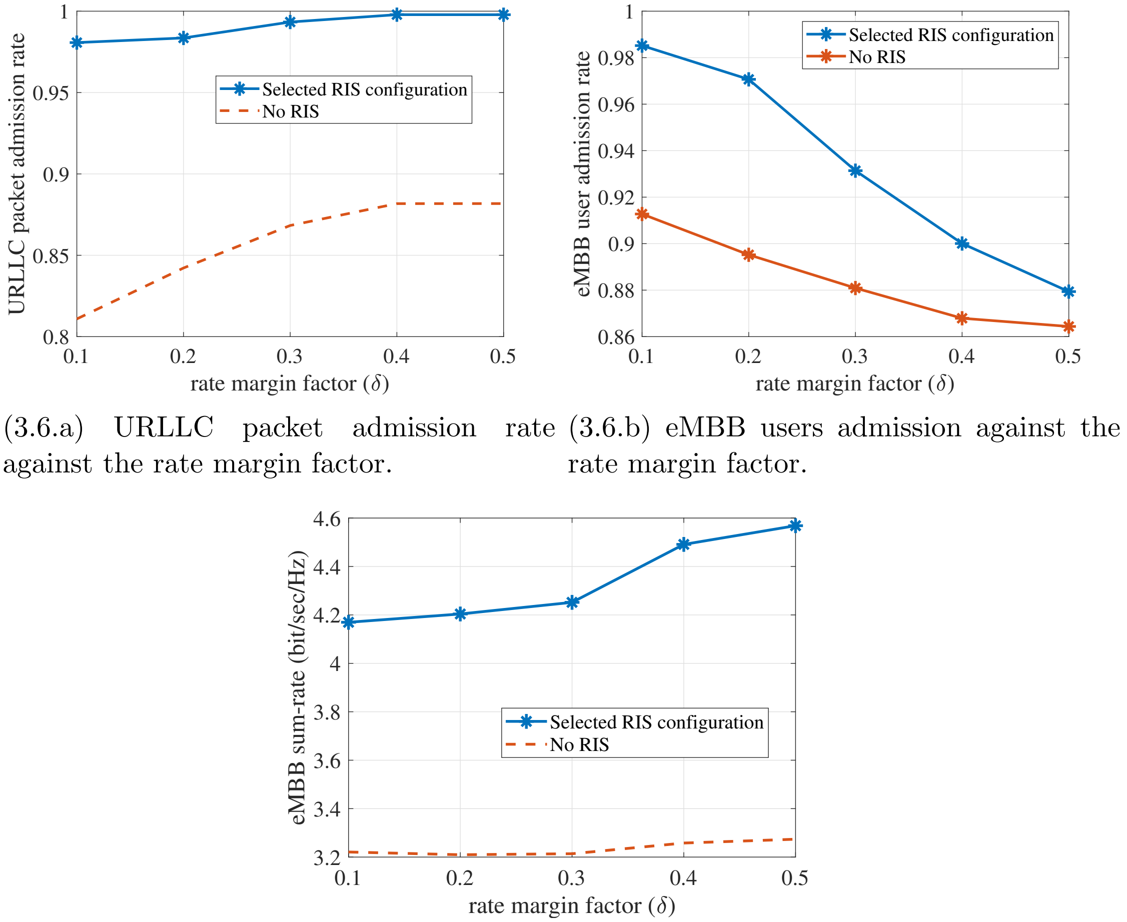
Fig. 3.5 depicts the impact of of the BS transmission power P_{BS}; the observations on Fig. 3.5 are summarized as follows.

- As shown in Fig. 3.5.a and Fig. 3.5.b, the selected RIS configuration can achieve better URLLC packet admission rate and better eMBB users admission rate compared to the case of no RIS is deployed. As whown in Fig. 3.5.a, the achieved gain of using the RIS is equivalent to 4 dB on the transmission power at 90% URLLC packet admission rate. This is attributed to the benefits of RIS in enhancing the channel conditions of the URLLC users such that the URLLC load can be admitted with limited frequency and power resources. On the other hand, the gain is equivalent to 4.5 dB at eMBB users admission rate of 95%. Moreover, by increasing the transmission power, the URLLC packet admission rate and the eMBB users admission rate enhance because more power resources are available to guarantee the URLLC and the eMBB rates requirements. By keeping increasing the transmission power, the gain achieved by the proposed scheme is reduced on the URLLC packet admission rate and the eMBB users admission rate. However this decreasing is translated to better gain on the eMBB sum rate as shown in Fig. 3.5.c.

- Fig. 3.5.c shows that at low transmission power the selected RIS configuration has a quite similar performance to the baseline in terms of the eMBB sum rate, whereas a better behaviour can be seen at high transmission power. This is attributed to the benefits of RIS in enhancing the URLLC admission which means less eMBB resources are punctured to allocate higher URLLC load.

3.6.5 Effect of the Rate Margin Factor δ

Fig. 3.6 illustrates the impact of the rate margin factor δ on the URLLC packets admission rate, the eMBB users admission rate, and the eMBB sum rate. This figure shows that increasing the margin factor enhances the URLLC packets admission and reduces the eMBB users admission. This is because increasing δ means more over-provisioned resources for the URLLC traffic, i.e., higher rate per eMBB user than the requested, which makes them later available to allocate the URLLC load. Hence, the enhanced URLLC packet admission rate. However, when higher rates are attained per eMBB user, this implies lower eMBB admission rate allowing only eMBB users with good channel conditions to be admitted, which will increase the sum rate, as shown in Fig. 3.6.c. Further, as shown in Fig. 3.6.b, increasing δ causes drastic reduction in the eMBB admission, since the network will over-provision much of its



(3.6.a) URLLC packet admission rate against the rate margin factor.

(3.6.b) eMBB users admission against the rate margin factor.

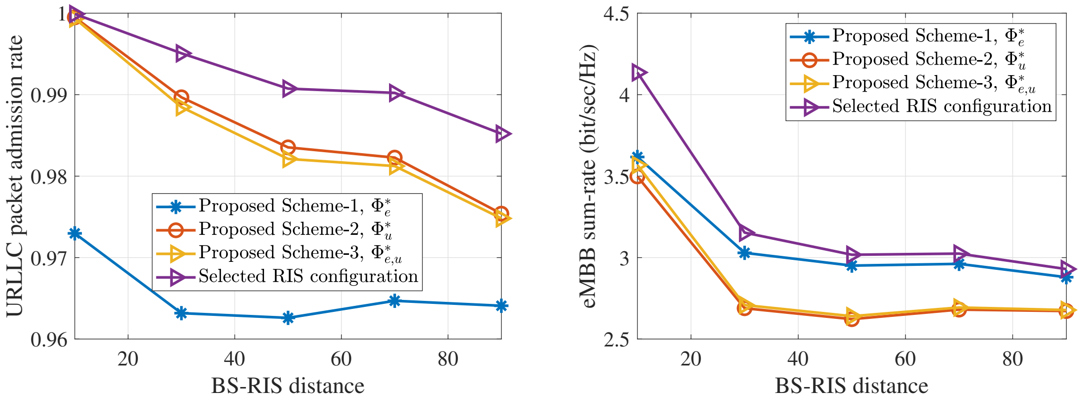
(3.6.c) eMBB sum rate against the rate margin factor.

Fig. 3.6: Performance of the proposed algorithm against rate margin. $N = 50$, $U = 65$ and $r_{\text{th}} = 7$ Mbps.

resources to satisfy the rate constraint per user and the RIS will be optimized to serve only such users with good channel conditions, leaving many of the eMBB users not admitted. This however helpsthe URLLC service to attain a better QoS.

3.6.6 Effect of the RIS Location

Fig. 3.7 illustrates the impact of varying the distance between the BS and the RIS on the URLLC packet admission rate and the eMBB sum rate. This figure shows that locating the RIS away from the BS impacts the performance of both URLLC and eMBB services in terms of the URLLC admission and the sum rate of eMBB,



(3.7.a) URLLC packets admission rate versus the distance between the BS and the RIS.

(3.7.b) eMBB sum rate versus the distance between the BS and the RIS. eMBB sum rate

Fig. 3.7: Performance of the proposed algorithm versus the distance between the BS and the RIS, where $N = 20$ and $U = 65$.

respectively. Indeed, this is consistent with other findings in the literature about the RIS deployment being most beneficial when it is located near the BS or close to the user [75]. However, the users are located randomly in the network. Hence, the best location of the RIS in this context is to be closer to the BS, which represents the optimal RIS placement for all users coexisting in the network.

3.7 Summary

In this chapter, we studied the RIS technology for enabling the coexistence of eMBB and URLLC services in a wireless networks. The RIS is deployed to improve the performance of the URLLC and eMBB users by controlling their channel conditions. Two optimization problems are formulated for multiplexing the eMBB and URLLC traffic, i.e., the time-slot basis eMBB allocation problem and the mini-time slot URLLC allocation problem. The eMBB allocation problem has the objective of maximizing the eMBB sum rate while satisfying the eMBB QoS requirements. Meanwhile, the URLLC allocation problem aims at maximizing the admitted URLLC packets and minimizing the eMBB rate loss while satisfying the QoS of eMBB and URLLC. To overcome the high computational complexity of optimizing the RIS phase-shift matrix per mini-time slot, we proposed a proactively designed RIS phase-shift matrices that are optimized at the beginning of the time slot. Simulation results show that the proposed algorithm has low time complexity which makes it a practical scheme to delay-sensitive URLLC traffic. It is also shown that the proposed RIS scheme achieves 99.99% URLLC packet admission rate using only 60 RIS elements. Moreover, the proposed model can achieve up to 70% enhancement on the eMBB rate compared to no-RIS is deployed.

Chapter 4

Low Complexity Passive Beamforming Designs

4.1 Introduction

4.1.1 Motivation

As discussed in Chapter 3, RIS has been introduced as a low-cost and an energy-efficient technology for controlling the wireless propagation environment [18, 19, 61]. Despite all the works reported in the context of RIS-assisted wireless networks, several challenges, such as the complexity of optimizing the RIS configuration and CSI estimation, need to be alleviated in order to harness the full proclaimed potential of integrating RIS in 6G wireless networks. In terms of complexity, solution optimality, and scalability, however, optimizing the RIS phase shifts remains a challenge, specially since the RIS is typically large. As a result, more effort should be put into lowering the complexity of optimizing RIS phase shifts and improving solution optimality.

The contents of this chapter has been submitted to IEEE Transactions on Vehicular Technology [76].

4.1.2 Background and Related Works

RIS has gained considerable attention from researchers in academia and industry as a promising enabling technology for 6G networks, owing to its capability of controlling the propagation environment with a low cost a low energy consumption [18, 77, 78]. In this context, the RIS-assisted UAV networks was investigated in [57, 79, 80]. The performance of NOMA aided by RIS was studied in [81–83]. The works in [84, 85] investigated integrating RIS into wireless networks with RSMA. The benefits of integrating RIS in the uplink and downlink transmissions of URLLC networks were studied in [28, 63, 86, 87]. The interplay between the URLLC and eMBB requirements in RIS-assisted wireless networks has been explored in [49, 55].

Motivated by the diverse benefits that RIS provides, several studies examined possible optimization strategies to achieve good performance in terms of accuracy, i.e., the optimally, and the complexity of the RIS configuration [58, 59, 88–92]. In [58], the transmit power minimization problem was formulated by jointly optimizing the active and the passive beamforming at the BS and the RIS, respectively. In this work, the SDR approach was leveraged to configure the passive beamforming. For the same objective under multi-cluster MISO-NOMA scenario, the authors of [88] proposed a method based on the second-order cone programming (SOCP) and alternating direction method of multipliers (ADMM) algorithm. Similarly, in [91], the power consumption minimization problem was formulated as a difference-of-convex problem, and the successive convex approximation (SCA) technique was used to obtain a sub-optimal solution. The authors of [89] adopted a modified version of the vector approximate message passing (VAMP) to optimize the RIS configuration in a way that maximizes the spectral efficiency. In [90], the authors adopted the fractional programming (FP) technique and the non-convex block coordinate descent (BCD) for joint active and passive beamforming with the objective of maximizing the weighted sum-rate. Yu et al. in [59] used fixed-point iteration and manifold optimization methods to enhance the spectral efficiency in RIS-aided cellular networks. Authors of[91] investigated the max-min rate problem in RIS-aided NOMA systems by concurrently optimizing the power allocation and the RIS phase-shift matrix. For the single-user scenario, the authors in [92] proposed a deep reinforcement learning (DRL) scheme for designing the passive beamforming in order to maximize the received SNR. Similarly, the DRL framework was proposed in [93] to optimize the power allocation and RIS phase-shifts configuration that maximize the average energy efficiency. In the context of URLLC service class, the work in [86] studied the optimization of the RIS configuration by leveraging a novel DRL to maximize the total achievable finite blocklength rate.

4.1.3 Contributions and Outcomes

This work aims at investigating low-complexity solutions for the design of RIS phase shifts in RIS-assisted downlink cellular networks consisting of single BS and a set of connecting U users. We first formulate an optimization problem for power allocation and passive beamforming where the objective it to minimize the total transmit power while satisfying the SNR requirements of individual users. However, due to the coupling between power allocation at the BS and passive beamforming at the RIS, the proposed problem is not convex, making it difficult to solve. Our main contribution is to tackle this difficulty by proposing two solutions which are illustrated as follows:

- For the case of a BS serving a single user, we graphically derive a closed-form expression for the RIS phase-shifts and the formulated problem. We also show that the derived closed-form expression for the RIS phase-shifts is similar to the one obtained in [58, 75].

- Motivated by the results of the single-user scenario and unlike the classical methods of optimizing the RIS which are based on optimizing the phase shift of all passive elements, we propose a new framework for reducing the optimization variables of the passive beamforming. Precisely, the variables optimizing the RIS phase shifts are scaled down from the number of RIS elements N to the number of served users U, which generally satisfies $U \ll N$, by applying a linear transformation. In this regard, we leveraged the optimal phase-shifts configurations obtained by assuming each user is independently served. Then, for each passive element, we combined linearly the optimal solutions of all users. To the best of our knowledge, this is the first work that proposes the transformation method in the context of RIS.

- We derive a sub-optimal solution for the RIS phase-shifts for the multi-users case by leveraging element-wise KKT optimal conditions of the original problem. Based on the obtained solution, we propose an iterative low complexity algorithm for optimizing power allocation and RIS-phase shifts. Although the KKT method is a traditional optimization technique, this is the first work that considers the proposed methodology to obtain a closed-form expression for the RIS configuration in multi-users systems.

- We then extend the proposed solution for MISO systems, where we show that the proposed scheme could be directly employed while adopting zero-forcing beamforming (ZFBF).
- Simulation results show that the proposed approaches have better performance than the baselines in terms of optimality and time complexity. Also, the proposed algorithm based on the closed-form expression has performance that is closed to the numerical (near to optimal). Moreover, it has a computational time of the order of milliseconds, which makes the proposed framework preferred in practical scenarios.

The rest of the Chapter is organized as follows. Section 4.2 presents the system model and problem formulation. Sections 4.3 and 4.4 present the proposed approach for single user and multi-users scenarios, respectively. Sections 4.5 and 4.6 present the simulation results and the conclusion, respectively.

4.2 System Model and Problem Formulation

4.2.1 System Model

We consider the downlink radio access network shown in Fig. 4.1, which consists of a single BS equipped with a single antenna that serves several single-antenna users.[1] Let U denotes the total number of active users in the network and $\mathcal{U} \triangleq \{1, 2, \dots, U\}$ denotes the set of served users. We employ the orthogonal frequency-division-multiple-access (OFDMA) to multiplex the active users. A single RIS is

[1] In this work, we consider the case of single antenna BS and cellular users. The primary motivation behind such a choice is a proof of concept for the proposed scheme and to obtain the main benefits and insights. Nevertheless, we show later that the same technique can be extended for the case of multiple antennas as.

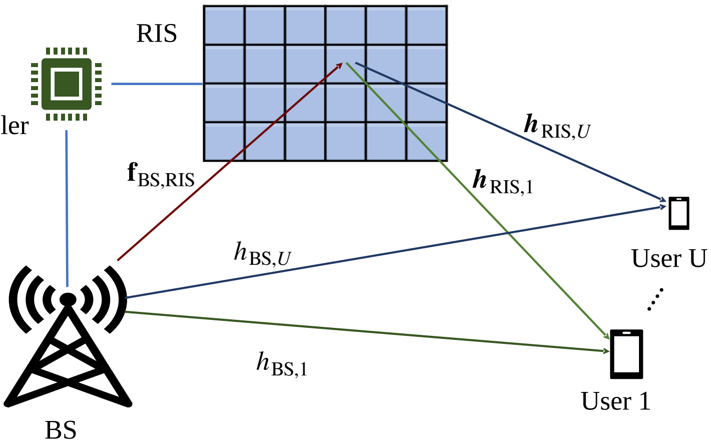

RIS
Controller
$\mathbf{f}_{\text{BS,RIS}}$
$\boldsymbol{h}_{\text{RIS},U}$
$\boldsymbol{h}_{\text{RIS},1}$
$h_{\text{BS},U}$
User U
$h_{\text{BS},1}$
User 1
BS

Fig. 4.1: System model

deployed in the network to improve the BS and users' communication links by dynamically controlling the propagation environment. Each reflective element of the RIS is connected to an atom that can adjust the phase of each incident wave. Let N denotes the total number of passive elements of the RIS and $\mathcal{N} \triangleq \{1, 2, \dots, N\}$ denotes the set of indices of the RIS elements. In this context, let $\mathbf{\Phi} \triangleq \text{diag}\left(e^{j\phi^1}, e^{j\phi^2}, ..., e^{j\phi^N}\right) \in \mathbb{C}^{N\times N}$ denote the phase-shift matrix of the RIS, where for all $n \in \mathcal{N}$, ϕ^n denotes the phase shift coefficient of the nth RIS reflective element. Accordingly, let $\boldsymbol{\vartheta} = [\vartheta^1, \vartheta^2, \dots, \vartheta^N]$ be the vector representing the phase shift coefficients of the RIS, where for all $n \in \mathcal{N}$, $\vartheta^n = e^{j\phi^n}$.

The CSI of all communicating links is assumed to be perfectly known at the BS [58]. Let $\mathbf{f}_{\text{BS,RIS}} \in \mathbb{C}^{N\times 1}$ be the vector that represents the channel coefficients between the BS and the RIS elements. Similarly, for all $u \in \mathcal{U}$, let $h_{\text{BS},u} \in \mathbb{C}$ and $\mathbf{h}_{\text{RIS},u} \in \mathbb{C}^{N\times 1}$ represent the channel coefficients between the BS and the uth active user and between the RIS and the uth active user, respectively. The communication links between the BS and the users and between the RIS and the users are assumed to have a quasi-static flat-fading Rayleigh channel. In contrast, the communication links between the BS and the RIS elements are assumed to have a Rician channel model. This assumption based on the fact that one of the requirements for RIS deployment in any cellular system is that the BS has to have a direct line of sight with the RIS. Accordingly, the small-scale fading of the BS-RIS link is modeled as $f = \sqrt{\frac{\kappa}{1+\kappa}}\, f^{\text{LoS}} + \sqrt{\frac{1}{1+\kappa}}\, f^{\text{NLoS}}$, where κ is the Rician factor, f^{LoS} is the line-of-sight component and f^{NLoS} is non-line-of-sight component, which follows a Rayleigh distribution with a scale parameter equals to one [72]. On the other hand, the large scale fading of the BS-users, BS-RIS, and RIS-users links is modelled as $P_0 = \alpha_0(d_{\mathrm{BS,u}})^{-\varrho_0}$, $P_1 = \alpha_1(d_{\mathrm{BS,RIS}})^{-\varrho_1}$ and $P_2 = \alpha_2(d_{\mathrm{RIS,u}})^{-\varrho_2}$, respectively, where $d_{\mathrm{BS,u}}$, $d_{\mathrm{RIS,u}}$ and $d_{\mathrm{BS,RIS}}$ are the distances of BS-users, BS-RIS, and RIS-users links, respectively. ϱ_0, ϱ_1 and ϱ_2 are the path loss exponents of BS-users, BS-RIS, and RIS-users links, respectively. Also, α_0 dB, α_1 and α_2 dB are path losses at the reference distance for the direct and cascaded links, respectively.

4.2.2 Signal Model and Rate Analysis

Taking into account both direct and cascaded links between the BS and each user, the received signal at the uth user, for all $u \in \mathcal{U}$, can be expressed as [66]

$$y_u = \left(h_{\mathrm{BS},u} + \mathbf{h}_{\mathrm{RIS},u}^{H} \mathbf{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}\right) \sqrt{p_u}\, x_u + z_u, \tag{4.1}$$

where x_u and p_u denote the transmit data symbol and the allocated power to user u, respectively, and z_u is the additive white Gaussian noise (AWGN) experienced at the uth user, which is assumed to be $\mathcal{CN}(0, \sigma^2)$ distributed. Based on this, the received SNR at the uth user, for all $u \in \mathcal{U}$, is expressed as

$$\gamma_u\left(\mathbf{\Phi}, p_u\right) = \frac{p_u |h_{\mathrm{BS},u} + \mathbf{h}_{\mathrm{RIS},u}^{H} \mathbf{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}|^2}{\sigma^2}. \tag{4.2}$$

4.2.3 Problem Formulation

In this chapter, our objective is to minimize the total transmit power of the BS by jointly designing the power allocated to the users and the phase shift matrix of the RIS, subject to the users QoS constraint represented by a minimum SNR threshold at each user. Such an objective can be attained by solving the following optimization problem.

$$\mathcal{P}: \min_{\mathbf{\Phi}, \mathbf{p}_{\mathrm{u}}} \sum_{u=1}^{U} p_u \tag{4.3a}$$

$$\text{s.t.} \ \frac{p_u |h_{\mathrm{BS},u} + \mathbf{h}_{\mathrm{RIS},u}^{H} \mathbf{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}|^2}{\sigma^2} \geq \gamma_u^{\mathrm{th}}, \quad \forall\, u \in \mathcal{U}, \tag{4.3b}$$

$$\phi^n \in [0, 2\pi], \qquad \forall\, n \in \mathcal{N}, \tag{4.3c}$$

where, for all $u \in \mathcal{U}$, γ_u^{th} denotes the minimum SNR requirement of the uth user and $\mathbf{p}_{\text{u}} = [p_1, p_2, ..., p_U]^T$ denotes the vector of the power allocated to the users by the BS. The objective in (4.3a) aims at minimizing the total transmit power $P = \sum_u^U p_u$. Constraint (4.3b) is the QoS of users which is represented by the minimum SNR requirement. Although, the objective is an linear function which is convex, the problem is still not convex due to the quadratic constraints in (4.3b).

4.3 Proposed Scheme for an RIS-aided Single User

Our objective is to propose a low complexity and efficient solution for problem $\mathcal{P}$. To get there, let us start with the optimal RIS configuration assuming only one user in the network is active in order to gain meaningful insights for the optimal joint power allocation and RIS-phase shifts design. In this case, problem $\mathcal{P}$ can be formulated as

$$\mathcal{P}_1 : \min_{\mathbf{\Phi}, p_1} \; p_1 \tag{4.4a}$$

$$\text{s.t.} \; \frac{p_1 |h_{\text{BS},1} + \mathbf{h}_{\text{RIS},1}^H \mathbf{\Phi}\, \mathbf{f}_{\text{BS,RIS}}|^2}{\sigma^2} \geq \gamma_1^{\text{th}}, \tag{4.4b}$$

$$\phi^n \in [0, 2\pi], \; \forall\, n \in \mathcal{N}, \tag{4.4c}$$

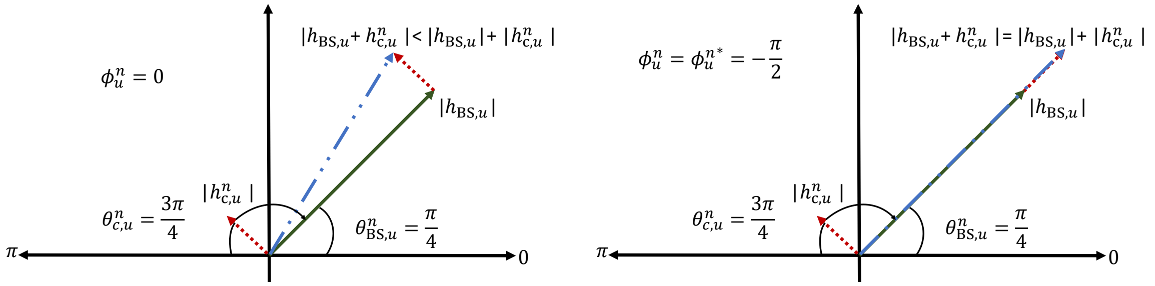
$$|\vartheta^n| = 1, \; \forall\, n \in \mathcal{N}. \tag{4.4d}$$

Problem $\mathcal{P}_1$ is the conventional power allocation minimization problem for the single SISO downlink system, and it is not difficult to verify that the optimal power allocation is $p_1 = \frac{\gamma_1^{\text{th}} \sigma^2}{|h_{\text{BS},1} + \mathbf{h}_{\text{RIS},1}^H \mathbf{\Phi}\, \mathbf{f}_{\text{BS,RIS}}|^2}$. Accordingly, problem $\mathcal{P}_1$ of minimizing the total transmit power can be equivalently transformed to the problem of maximizing the combined channel's gain as follows.

$$\mathcal{P}_2 : \max_{\mathbf{\Phi}} \; |h_{\text{BS},1} + \mathbf{h}_{\text{RIS},1}^H \mathbf{\Phi}\, \mathbf{f}_{\text{BS,RIS}}|^2 \tag{4.5a}$$

$$\text{s.t.} \quad \phi^n \in [0, 2\pi], \; \forall\, n \in \mathcal{N} \tag{4.5b}$$

$$|\vartheta^n| = 1, \; \forall\, n \in \mathcal{N}. \tag{4.5c}$$



$|h_{\mathrm{BS},u}+h^n_{\mathrm{c},u}| < |h_{\mathrm{BS},u}|+|h^n_{\mathrm{c},u}|$
$\phi^n_u = 0$
$|h_{\mathrm{BS},u}|$
$|h^n_{\mathrm{c},u}|$
$\theta^n_{c,u} = \frac{3\pi}{4}$
$\theta^n_{\mathrm{BS},u} = \frac{\pi}{4}$
π
0
$|h_{\mathrm{BS},u}+h^n_{\mathrm{c},u}| = |h_{\mathrm{BS},u}|+|h^n_{\mathrm{c},u}|$
$\phi^n_u = \phi^{n*}_u = -\frac{\pi}{2}$
$|h_{\mathrm{BS},u}|$
$|h^n_{\mathrm{c},u}|$
$\theta^n_{c,u} = \frac{3\pi}{4}$
$\theta^n_{\mathrm{BS},u} = \frac{\pi}{4}$
π
0

(4.2.a) $\theta^n_u = 0$.

(4.2.b) $\theta^{n*}_u = \frac{-\pi}{2}$

Fig. 4.2: Polar representation of the RIS phase shift adjustment assuming only user $u \in \mathcal{U}$ is active. The optimal RIS phase shift is satisfied when both direct and cascaded links are in the same direction.

Let the channel coefficients of both direct and cascaded links, for each $n \in \mathcal{N}$, be $h_{\mathrm{BS},u} = |h_{\mathrm{BS},u}|\, e^{j\theta_{\mathrm{BS},u}}$ and $h^n_{\mathrm{c},u} = |h^n_{\mathrm{c},u}|\, e^{j\theta^n_{\mathrm{c},u}} = h^n_{\mathrm{RIS},u}\, f^n_{\mathrm{BS,RIS}} = |h^n_{\mathrm{RIS},u}|\, |f^n_{\mathrm{BS,RIS}}|\, e^{j(\theta^n_{\mathrm{RIS},u}+\theta^n_{\mathrm{BS,RIS}})}$, respectively, where $|h_{\mathrm{BS},u}|$, $|h^n_{\mathrm{RIS},u}|$ and $|f^n_{\mathrm{BS,RIS}}|$ are the amplitude of $h_{\mathrm{BS},u}$, $h^n_{\mathrm{RIS},u}$ and $f^n_{\mathrm{BS,RIS}}$, respectively. Similarly, $\theta_{\mathrm{BS},u}$, $\theta^n_{\mathrm{RIS},u}$ and $\theta^n_{\mathrm{BS,RIS}}$ are the angles of the complex channels $h_{\mathrm{BS},u}$, $h^n_{\mathrm{RIS},u}$ and $f^n_{\mathrm{BS,RIS}}$, respectively. Then, the solution of problem $\mathcal{P}_2$ can be found in the following proposition.

Proposition 4.1. *Assuming only user $u \in \mathcal{U}$ is active, the optimal RIS phase shift design ϕ^{n*}_u, for all $n \in \mathcal{N}$, is obtained if and only if the channel coefficients of both direct and cascaded links are in the same direction, i.e.,*

$$\phi^{n*}_u = \theta_{\mathrm{BS},u} - \left(\theta^n_{\mathrm{RIS},u} - \theta^n_{\mathrm{BS,RIS}}\right). \tag{4.6}$$

Proof. We formulate the Lagrangian function and apply the KKT optimality conditions on $\mathcal{P}2$. After manipulating the obtained KKT conditions, an element-wise solution is adopted to solve each independent phase-shift constraint which directly updated in the Lagrangian multiplier. Detailed KKT analysis is omit for brevity. □

To elaborate more on **Proposition 4.1**, Fig. 4.2 provides a graphical depiction for (4.6). The figure shows that, the sum of the channel coefficients $|h_{\mathrm{BS},u}|\, e^{j\theta_{\mathrm{BS},u}}$ and $|h^n_{\mathrm{c},u}|\, e^{j\theta^n_{\mathrm{c},u}}$ is maximized when both vectors are in the same direction i.e., when $\theta_{\mathrm{BS},u} = \phi^{n*}_u + \left(\theta^n_{\mathrm{RIS},u} - \theta^n_{\mathrm{BS,RIS}}\right)$. This means the cascaded link channel should be rotated by $\phi^{n*}_u = \theta_{\mathrm{BS},u} - \left(\theta^n_{\mathrm{RIS},u} - \theta^n_{\mathrm{BS,RIS}}\right)$. This provides a graphical proof to **Proposition 4.1**.

4.4 Extension to the RIS-Aided Multi-User Case

This section extends the previous by considering a multiuser setting. We propose two efficient techniques for solving problem $\mathcal{P}$ sub-optimally by generalising the two techniques used in the single user scenario, i.e., the graphical and the element-wise KKT solutions.

4.4.1 A Linear Transformation Approach

We present an effective strategy to overcome the complexity of the RIS phase shift matrix optimization. To do this, with the aid of alternating optimization, we decompose problem $\mathcal{P}$ into a power control and a phase shift optimization sub-problems. For each iteration i, these sub-problems are solved iteratively until convergence or maximum number iteration reaches. In this regard, for the phase shift optimization sub-problem, we perform a linear transformation on the phase shift matrix based on each user's optimal phase shift matrix. Accordingly, the number of optimization variables is reduced from the number of passive elements N to the number of served users U, which typically satisfies $U \ll N$.

4.4.1.1 RIS Phase Shifts Matrix

For a fixed $\mathbf{p}_{\mathrm{u}}^{i}$, the RIS phase shift matrix can be obtained by solving the following optimization problem.

$$\mathcal{P}_4^{(i)} : \textbf{Find } \mathbf{\Phi}^i \tag{4.7a}$$

$$\text{s.t. } \frac{p_u^i |h_{\mathrm{BS},u} + \mathbf{h}_{\mathrm{RIS},u}^H \mathbf{\Phi}^i \, \mathbf{f}_{\mathrm{BS,RIS}}|^2}{\sigma^2} \geq \gamma_u^{\mathrm{th}}, \ \forall \ u \in \mathcal{U}, \tag{4.7b}$$

$$\phi^n \in [0, 2\pi], \ \forall \ n \in \mathcal{N}, \tag{4.7c}$$

$$|\vartheta^n| = 1, \ \forall \ n \in \mathcal{N}. \tag{4.7d}$$

Problem $\mathcal{P}_4^{(i)}$ is a non-convex quadratic optimization problem. In fact, recent works have difficulty in optimizing such RIS phase shift problem. Specifically, the RIS size should practically be sufficiently large in order to attain the performance gain of RIS [94, 95]. Consequently, optimizing the RIS phase shifts is very complicated and the methods used do not scale when the number of passive elements is significantly

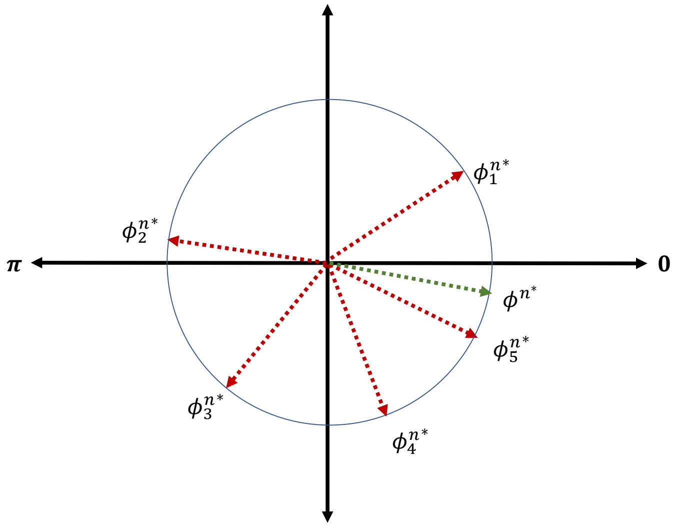


Fig. 4.3: An example for the optimal phase shift for one passive element. Red-doted lines are the optimal passive element configuration for users. Green-doted line is the global optimal configuration for passive element n.

large, i.e., N decision variables and N decision constraints [58, 88–90, 92]. To tackle this problem, our approach is to decrease the optimization variables of the RIS phase shift problem to reduce the complexity of problem $\mathcal{P}_4^{(i)}$ based on the following two facts. First, the number of users associated with the RIS needs to be much smaller than the RIS passive elements, i.e., $U \ll N$ [58, 88–90, 92]. Second, the optimal phase shifts design for the single-user case can be easily obtained via closed-form expression with a complexity of $O(1)$ as illustrated in Proposition 1.

The question which arises then is ”*How can we leverage the optimal RIS for each user independently to obtain a sub optimal solution for problem $\mathcal{P}_4^{(i)}$?*”. To answer this question, let us first define $\boldsymbol{\vartheta}_u^* \triangleq \left[\vartheta_u^{1*}, \vartheta_u^{2*}, ..., \vartheta_u^{N*}\right] \in \mathbb{C}^{N\times 1}$ as the vector corresponding to the optimal RIS configuration of the uth user, for all $u \in \mathcal{U}$, which is obtained using the results of Proposition 1. Also, we define the matrix of optimal configurations all users as $\boldsymbol{\Theta}^* \triangleq [\boldsymbol{\vartheta}_1^*, \boldsymbol{\vartheta}_2^*, ..., \boldsymbol{\vartheta}_U^*] \in \mathbb{C}^{N\times U}$. Then, we want to investigate the relation between the optimal RIS configuration $\boldsymbol{\Phi}^*$, which is the solution for problem $\mathcal{P}_4^{(i)}$, and between the RIS configuration $\boldsymbol{\Theta}^*$. To do this, we graphically depicts in Fig. 4.3 an example of the relation between $\boldsymbol{\Phi}^*$ and $\boldsymbol{\Theta}^*$ for a single passive element. It is clear from Fig. 4.3, if ϕ_2^{n*} is used to configure passive element $n \in \mathcal{N}$, then the reflected signals will constructively add at the receivers of users 2 and 3 only, while they destructively add for the remaining users. Alternatively, if ϕ_5^{n*} is adopted to configure passive element $n \in \mathcal{N}$, then the reflected signal will constructively add (with different levels) at the receivers of all users except user 2. Hence, one can observe that there is a strong correlation between the individual optimal configurations of users and the optimal RIS configuration. In other words, the configuration of the RIS, $\mathbf{\Phi}$, can be expressed as a function of $\mathbf{\Theta}^*$ as

$$\mathbf{\Phi}(\boldsymbol{\vartheta}) \xrightarrow{\mathcal{F}(\mathbf{\Theta}^*,\cdot)} \mathbf{\Phi}(\mathbf{y}), \tag{4.8}$$

where $\mathbf{y} \in \mathbb{C}^{U\times 1}$ is the vector of the new optimization variables. The function $\mathcal{F}$ maps the high dimensional $\mathbf{\Phi}(\boldsymbol{\vartheta})$ to a more convenient matrix $\mathbf{y}$ with less optimization variables. In fact, the transformation (or mapping) function $\mathcal{F}$ can be any general function, such as linear, logarithmic, etc. However, obtaining the exact relation function, i.e., $\mathcal{F}$, is a challenging task.[2] [3] Thus, we consider the following linear transformation due to its simplicity.

$$\mathbf{\Phi}(\mathbf{y}) \triangleq \operatorname{diag}\left(\mathbf{\Theta}^* \times \mathbf{y}\right), \tag{4.9}$$

where

$$|\mathbf{\Theta}^*\,\mathbf{y}| = \mathbf{1}, \tag{4.10}$$

$\mathbf{1} \in \mathbb{R}^{N\times 1}$ is the all one vector. Thus, the phase shift coefficient of RIS element n is defined as $\vartheta^n = (\vartheta_1^{n*} \times y_1 + \vartheta_2^{n*} \times y_2 + \cdots + \vartheta_U^{n*} \times y_U)$. Then, the optimal RIS configuration $\mathbf{\Phi}^* = \operatorname{diag}\left(\mathbf{\Theta}^* \times \mathbf{y}^*\right)$, where $|\mathbf{\Theta}^*\,\mathbf{y}^*| = \mathbf{1}$ and $\mathbf{y}^*$ is the optimal vector of $\mathbf{y}$. For more insights on the linear transformation, let $\mathbf{y}^* = [0, 0, 0, 0, 1]$ be the optimal value. Then, as shown in Fig. 4.3, only the configuration of user 5 is used to configure the RIS. In fact, more complex combinations will result from optimizing $\mathbf{y}$ to converge to a local optimal solution. In fact, this linear transformation has been widely used in the context of machine learning [96]. Moreover, some recent works considered dimensionality reduction scheme where we search for the optimal solution in small subset called the effective subset [97, 98]. Meanwhile, we can rewrite the

[2] Due to the optimization simplicity, we consider a linear transformation of a family of possible transformation functions.

[3] The following subsection provides an approximated closed-form expression by applying the KKT conditions. This expression extracts an effective phase shift for each passive element that minimize the consumed power.

optimization problem $\mathcal{P}_4^{(i)}$ as

$$\mathcal{P}_{4,1}^{(i)} : \textbf{Find } \mathbf{y}^i \tag{4.11a}$$

$$\text{s.t. } \frac{p_u^i |h_{\text{BS},u} + \mathbf{h}_{\text{RIS},e}^H \mathbf{\Phi}(\mathbf{y}^i)\, \mathbf{f}_{\text{BS,RIS}}|^2}{\sigma^2} \geq \gamma_u^{\text{th}},\ \forall\, u \in \mathcal{U}, \tag{4.11b}$$

$$\phi^n \in [0, 2\pi],\ \forall\, n \in \mathcal{N}, \tag{4.11c}$$

$$|\mathbf{\Theta}^* \, \mathbf{y}^i| = \mathbf{1}. \tag{4.11d}$$

It is clear from problems $\mathcal{P}_4^{(i)}$ and $\mathcal{P}_{4,1}^{(i)}$ that the number of optimization variables is reduced from N to U, which generally satisfies $U \ll N$. Hence, the complexity of optimizing the RIS phase shift is significantly reduced. Then, we apply the SDR approach with some modifications [58]. [4] The details are omitted here for brevity.

[4] In this work, we focus on minimizing the power consumption as a proof of concept of the proposed linear transformation. However, the same methodology is applicable with different objectives, such as maximizing the sum rate.

Algorithm 4.1: Proposed SDR-based Alternating Algorithm

```
Evaluate $\boldsymbol{\Theta}^*$
Initialize $\mathbf{y}^0$ and the iteration number $i = 0$.
repeat
    $i \leftarrow i + 1$
    Solve $\mathcal{P}_3^i$ for a given $\mathbf{y}^{i-1}$, and denote the optimal solution as $\mathbf{p}_{\mathrm{u}}^i$.
    Solve $\mathcal{P}_{4,1}^i$ for a given $\mathbf{p}_{\mathrm{u}}^i$, and denote the solution after Gaussian
    randomization as $\mathbf{y}^i$.
until Convergence or maximum iterations reached
```

4.4.1.2 Power Control Problem

Now, for fixed $\boldsymbol{\Phi}^i$, problem $\mathcal{P}$ is reduced to a power control problem as

$$\mathcal{P}_3^i : \min_{\mathbf{p}_{\mathrm{u}}^i} \sum_{u=1}^{U} p_u^i \tag{4.12a}$$

$$\text{s.t.} \quad \frac{p_u^i |h_{\mathrm{BS},u} + \mathbf{h}_{\mathrm{RIS},u}^H \boldsymbol{\Phi}^i \, \mathbf{f}_{\mathrm{BS,RIS}}|^2}{\sigma^2} \geq \gamma_u^{\mathrm{th}}, \ \forall \ u \in \mathcal{U}. \tag{4.12b}$$

Similar to the single user case, $\mathcal{P}_2^{(i)}$ is the conventional power control minimization problem in the multiuser SISO downlink system where the optimal power allocation is

$$p_u^i = \frac{\gamma_u^{\mathrm{th}} \, \sigma^2}{|h_{\mathrm{BS},u} + \mathbf{h}_{\mathrm{RIS},u}^H \boldsymbol{\Phi}^i \, \mathbf{f}_{\mathrm{BS,RIS}}|^2}, \quad \forall u \in \mathcal{U} \tag{4.13}$$

4.4.1.3 Proposed SDR-Based Algorithm

The SDR approach with the Gaussian randomization are used to solve problem $\mathcal{P}_{4,1}^{(i)}$. As illustrated in **Algorithm 1** (refers to Algorithm 4.1), the algorithm starts by calculating the optimal phase shifts independently for all served users. At each iteration, the algorithm solves problem $\mathcal{P}_3^{(i)}$ and the resulting solution, denoted by $\mathbf{p}_{\mathrm{u}}^i$, is fed into $\mathcal{P}_{4,1}^{(i)}$. The resulting $\mathbf{y}^i$ at iteration i is used as the initial point of the next iteration, i.e., $i + 1$. Steps 4 to 6 are attentively repeated until convergence or a maximum number of iterations is reached.

4.4.2 Element-Wise KKT Approach

Motivated by the optimality and the low complexity of the closed-form expression for the RIS-assisted single-user case, we aim to explore the likelihood of obtaining a similar analytical expression for the multi-user scenario. Specifically, we intend to adopt the KKT method to design the RIS-phase shift design. To do so, the optimal power allocation (4.13) is substituted in the optimization problem $\mathcal{P}$. Hence, the problem $\mathcal{P}$ can be equivalently transformed, by optimizing the RIS phase-shifts only, as

$$\mathcal{P}_5 : \min_{\mathbf{\Phi}} \sum_{u=1}^{U} \frac{\sigma^2 \gamma_u^{\text{th}}}{|h_{\text{BS},u} + \mathbf{h}_{\text{RIS},u}^H \mathbf{\Phi}\, \mathbf{f}_{\text{BS,RIS}}|^2} \tag{4.14a}$$

$$\text{s.t. } \phi^n \in [0, 2\pi],\ \forall\, n \in \mathcal{N}, \tag{4.14b}$$

$$|\vartheta^n| = 1,\ \forall\, n \in \mathcal{N}. \tag{4.14c}$$

Problem, $\mathcal{P}_5$ is a non-convex problem. However, $\mathcal{P}_5$ is differentiable and monotonically non-decreasing. Hence, the KKT optimality conditions can be applied. By doing this, the sub-optimal solution of problem $\mathcal{P}_5$ is given by Theorem 4.1.

Theorem 4.1. *The sub-optimal RIS phase shift configuration of problem $\mathcal{P}_5$ can be approximately expressed as:*

$$\phi^{n*} \approx \arctan\left(\frac{\sum_{u=1}^{U} C_u^n}{\sum_{u=1}^{U} D_u^n}\right),\ \phi^{n*} \in \{0, 2\pi\}, \tag{4.15}$$

where

$$C_u^n = \frac{2\,\sigma^2\,\gamma_u^{\text{th}}\,|h_{\text{c,u}}^n|\,\sin(\theta_{d,u}^n - \theta_{\text{c},u}^n)}{|h_{d,u}^n|^3}, \tag{4.16a}$$

$$D_u^n = \frac{2\,\sigma^2\,\gamma_u^{\text{th}}\,|h_{\text{c,u}}^n|\,\cos(\theta_{d,u}^n - \theta_{\text{c},u}^n)}{|h_{d,u}^n|^3}, \tag{4.16b}$$

$$h_{d,u}^n = h_{\text{BS},u} + \sum_{k \neq n}^{N} h_{\text{c},u}^k\, e^{j\theta^k} = |h_{d,u}^n| e^{j\theta_{d,u}^n}. \tag{4.16c}$$

Proof. Detailed proof is provided in Appendix B.1. □

Algorithm 4.2: One iteration Algorithm based on closed-form expression.

```
1 Initialize $\mathbf{\Phi}^0$
2 for all $n \in \mathcal{N}$ do
3     update $\phi^n$ according to (4.15)
4 end for Compute $p_u$, $\forall u \in \mathcal{U}$ according to (4.13)
```

The proposed algorithm based on the closed-form expression is illustrated in **Algorithm 2** (refers to Algorithm 4.2). The algorithm starts by initializing the RIS phase shift at step 1. Then, the RIS phase shifts are updated iteratively. Finally, the algorithm computes the power allocation based on the optimized RIS phase shift matrix. Noting that the initialize the RIS phase shifts is as follows. We first sort the users in ascending order based on their channel conditions, i.e., $|h_{\mathrm{BS},1}| + \sum_{n=1}^{N} |h_{c,1}^{n}| \leq |h_{\mathrm{BS},2}| + \sum_{n=1}^{N} |h_{c,2}^{n}| \leq \cdots \leq |h_{\mathrm{BS},u}| + \sum_{n=1}^{N} |h_{c,U}^{n}|$. Then, we set $\boldsymbol{\phi}^0 = \boldsymbol{\vartheta}_1^*$. This initialization strategy is motivated to start with a possible point biased to the user with the worst channel gain, i.e., users who need maximum power to satisfy the QoS requirements.

Since Algorithm 2 directly applies the results of Theorem 4.1, which lies on some approximations, we propose Algorithm 3 (refers to Algorithm 4.3) to guarantee the improvement on the power consumption at the BS. Inspired by that the power allocation is obtained through closed-form expression, which generally has a complexity of $O(1)$, so we update the RIS phase shifts for all $n \in \mathcal{N}$ if it decreases the BS power consumption, otherwise it remains fixed. This action is performed in steps 6 to 11. In fact, Algorithm 3 updates the phase shift matrix, i.e., steps 3 to 12, in an iterative manner until convergence or maximum iterations is reached. As a result, Algorithm 3 have higher complexity than Algorithm 2.

4.4.3 Complexity and Convergence Analysis

In this section, we provide the complexity and the convergence analysis of Algorithm 2 and Algorithm 3. It is clear, Algorithm 2 consists only of one iteration, i.e., one-shot algorithm. Accordingly, it has deterministic convergence characteristic of one iteration. On the other hand, Algorithm 3 has higher complexity than Algorithm

Algorithm 4.3: Iterative Algorithm based on closed-form expression with power allocation update

```
1  Initialize the iteration number $i = 0$.
2  Initialize $\mathbf{\Phi}^0$, compute $\mathbf{p}_u^0$ and $P^0$
3  repeat
4      $i \leftarrow i + 1$
5      for all$n \in \mathcal{N}$ do
6          Compute $\phi^n$ according to (4.15) and $P^n$ using (4.13)
7          if $P^n \leq P^{n-1}$then
8              update $\phi^n$ and $P^n$
9          else
10             $P^n = P^{n-1}$
11         end if
12     end for
13     Compute $p_u^i$, $\forall u \in \mathcal{U}$ and $P^i$ according to (4.13)
14 until Convergence or maximum iterations reached
```

2 since it is an iterative algorithm. Moreover, Algorithm 3 computes P^n for each iteration over $n \in \mathcal{N}$. Nevertheless, the complexity of Algorithm 3 still low. Precisely, the complexity of computing the P^n is of order $O(U)$ since p_u is computed through closed-form expression of complexity of $O(1)$. Similarly, the complexity of computing ϕ^n is computed through closed-form expression with complexity of $O(1)$. Accordingly, the complexity of Algorithm 2 and Algorithm 3 are only $O(N)$ and $O(U \times N \times \mathcal{I})$, respectively, where $\mathcal{I}$ is the number of iterations. Now, it remains to prove that Algorithm 3 is going to converge. In fact, since steps 6 to 11 guarantee that the BS transmit power is monotonically decreasing, then Algorithm 3 is going to converge to a local optimal solution.

4.4.4 Extension to the Multi-Antenna Case

This section presents possible extensions for the proposed element-wise KKT scheme. Precisely, in the previous section, we have derived a closed-form expression for the case of a single antenna BS serving multiple users. However, this expression can also be extended for a BS equipped with multiple antennas when ZFBF is applied [99]. For clarity, let $\mathbf{G}_{\text{BS,RIS}} \in \mathbb{C}^{N\times M}$ be the matrix that represents the channel coefficients between the BS and the RIS elements, where $M \geq U$ is the number of antennas at the BS. Moreover, for all $u \in \mathcal{U}$, let $\mathbf{h}_{\text{BS},u} \in \mathbb{C}^{1\times M}$ and $\mathbf{h}_{\text{RIS},u} \in \mathbb{C}^{N\times 1}$ represent the channel vectors between the BS and the uth active user and between the RIS and the uth active user, respectively. In this case, problem $\mathcal{P}$ can be written for the case of multiple antenna BS as [58]:

$$\mathcal{P}_6 : \min_{\mathbf{\Phi},\mathbf{W}} \sum_{u=1}^{U} ||\mathbf{w}_u||^2 \tag{4.17a}$$

$$\text{s.t.} \frac{|(\mathbf{h}_{\text{BS},u} + \mathbf{h}^H_{\text{RIS},u}\mathbf{\Phi}\,\mathbf{G}_{\text{BS,RIS}})\mathbf{w}_u|^2}{\sum\limits_{e\neq u}^{U} |(\mathbf{h}_{\text{BS},u} + \mathbf{h}^H_{\text{RIS},u}\mathbf{\Phi}\mathbf{G}_{\text{BS,RIS}})\mathbf{w}_e|^2 + \sigma^2} \geq \gamma_u^{\text{th}}, \forall\, u \in \mathcal{U}, \tag{4.17b}$$

$$\phi^n \in [0, 2\pi], \qquad \forall\, n \in \mathcal{N}, \tag{4.17c}$$

where $\mathbf{w}_u \in \mathbb{C}^{M\times 1}$ is the precoding vector corresponding to the uth user. In addition, the term

$$\frac{|(\mathbf{h}_{\text{BS},u} + \mathbf{h}^H_{\text{RIS},u}\mathbf{\Phi}\,\mathbf{G}_{\text{BS,RIS}})\mathbf{w}_u|^2}{\sum\limits_{e\neq u}^{U} |(\mathbf{h}_{\text{BS},u} + \mathbf{h}^H_{\text{RIS},u}\mathbf{\Phi}\,\mathbf{G}_{\text{BS,RIS}})\mathbf{w}_e|^2 + \sigma^2} \tag{4.18}$$

is the signal-to-interference-plus-noise ratio (SINR) at the uth user. Similar to $\mathcal{P}$, the objective of $\mathcal{P}_6$ is to minimize the power consumption at the BS while satisfying the SINR constraints in (4.17b). Then, the power control problem can be written as

$$\mathcal{P}_7 : \min_{\mathbf{W}} \sum_{u=1}^{U} ||\mathbf{w}_u||^2 \tag{4.19a}$$

$$\text{s.t.} \frac{|\mathbf{h}^{eff}_u \mathbf{w}_u|^2}{\sum_{e\neq u}^{U} |\mathbf{h}^{eff}_e \mathbf{w}_e|^2 + \sigma^2} \geq \gamma_u^{\text{th}}, \quad \forall\, u \in \mathcal{U}, \tag{4.19b}$$

where $\mathbf{h}^{eff}_u = \mathbf{h}_{\text{BS},u} + \mathbf{h}^H_{\text{RIS},u}\mathbf{\Phi}\,\mathbf{G}_{\text{BS,RIS}}$. By adopting ZFBF to nullify the interference, one can easily obtain the optimal value of $\mathbf{W}$ as [99]

$$\mathbf{W}^* = \boldsymbol{\mathcal{H}}^H\,(\boldsymbol{\mathcal{H}}\,\boldsymbol{\mathcal{H}}^H)^{-1}\,\mathcal{Q} \tag{4.20}$$

where

$$\mathcal{Q} = \begin{bmatrix} \sqrt{\sigma^2\,\gamma_1^{\text{th}}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma^2\,\gamma_1^{\text{th}}} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma^2\,\gamma_1^{\text{th}}} \end{bmatrix} \tag{4.21}$$

and $\boldsymbol{\mathcal{H}} = [\mathbf{h}_1^{eff}, \mathbf{h}_2^{eff}, \dots, \mathbf{h}_U^{eff}] \in \mathbb{C}^{U\times M}$ and $(.)^{-1}$ is the matrix inverse.

The resulting signal after adopting ZFBF is free of interference. Hence, we can apply the results of Theorem 4.1, where $h_{\text{BS},u}$ and $h_{\text{c},u}^n$ are given by $\mathbf{h}_{\text{BS},u}\,\frac{\mathbf{w}_u}{||\mathbf{w}_u||}$ and $h_{\text{RIS},u}^n\,\mathbf{g}^n\,\frac{\mathbf{w}_u}{||\mathbf{w}_u||}$, respectively, such that $\mathbf{g}^n \in \mathbb{C}^{1\times M}$. Consequently, we can use directly Algorithm 3 with minor modifications on steps 6 and 13. Specifically, we update $\mathbf{W}$ using the expression in (4.20) instead of updating $\mathbf{p}_u$. The main idea is to ensure that the signal is free of interference (using ZFBF) at each modification on the RIS phase shift matrix. Accordingly, the complexity of Algorithm 3 becomes $O([U\times M^2]\times N\times\mathcal{I})$ where $[U\times M^2]$ is the complexity of computing the ZFBF beamforming. However, we are able to extend the proposed element-wise KKT approach for problem $\mathcal{P}$ for multiple antenna BS; sophisticated expressions are needed based on the precoder and multiplexing scheme such as NOMA. Similarly, refined expression needs further derived under different objectives, such as maximizing the sum rate. In future work, we plan to apply the proposed techniques on such scenarios and objectives which are beyond the scope of this work.

4.5 Performance Evaluation

This section presents extensive simulations to evaluate the performance of our proposed schemes. The simulation environment consists of one BS and one RIS. The BS is located in the 3-dimensional Cartesian coordinates systems (X, Y, Z) at $(0\,\text{m}, 0\,\text{m}, 20\,\text{m})$ meters while the RIS is located at $(0\,\text{m}, 100\,\text{m}, 10\,\text{m})$ metres. Users are randomly located around the RIS with a maximum radius of 20m. The number of RIS reflecting elements N and the number of users U are key parameters in the performance of the proposed scheme. Hence, we vary N in the range $[10, 50]$. Moreover, we test the performance of the proposed scheme for different number of users, i.e., $U = [2, 10]$. The remaining system parameters are summarized in Table 4.1 [62, 72, 74, 100]. For the sake of comparison, we consider two baselines besides the proposed algorithms.[5] Here, we summarize the baselines and the proposed algorithms.

- **Conventional SDR**: The conventional SDR scheme proposed in [58] is considered as a baseline. This algorithm is adopted as a lower bound of the solution as it guarantees only $\pi/4$ of the optimal solution.

- **Numerical**: The solution of problem $\mathcal{P}$ is obtained numerically using the Matlab *FMINCON()* solver. Among 100 independent iterations, we select the solution that minimizes the transmit power. This solution is considered as a close-optimal or upper bound for this problem in [58].

- **Proposed SDR**: By adopting the SDR solution scheme on the proposed problem formation $\mathcal{P}_{4,1}^{(i)}$.

- **Algorithm 2**: This solution is obtained for problem $\mathcal{P}_3$ by using the one-shot Algorithm.

- **Algorithm 3**: This solution is obtained for problem $\mathcal{P}_3$ by iteratively solving the power allocation problem and using the results in **Theorem 4.1** for solving the phase shift optimization.

To validate the efficiency of the proposed algorithms, we first present the convergence behaviour of the proposed SDR and Algorithm 3. Then, we study the performance of the proposed algorithms in terms of *optimality*, *complexity* and *scalability* which are critical metrics to judge any algorithm's applicability in real applications. In *optimality*, we study how much the performance of the proposed algorithms is close to the near-optimal solution obtained numerically. The *complexity* is illustrated through the computational time of the proposed algorithms.[6] The *scalability* measures if we can apply the proposed schemes with large size RIS given that the RIS is generally composed of a large number of passive elements.

[5] Simulation results are performed over 1000 independent realizations of channel gains and users' locations.

[6] The algorithm was implemented in Matlab using a machine with the following characteristics: System Type: x64-based PC Processor: Intel(R) i7-8700 CPU @3.2GHz, 16 Gigabyte RAM.

Table 4.1: Simulation parameters

Parameter	Symbol	Value
Noise power	σ^2	-90 dBm
path loss exponents	ϱ_0, ϱ_1 and ϱ_2	3.5, $2, 2$ and 3.5
Rician factor	κ	3
Minimum SNR requirement	γ_u^{th}	4.8 dB
Maximum number of iterations	$\mathcal{I}$	50
Convergence error	ϵ	1×10^{-2}

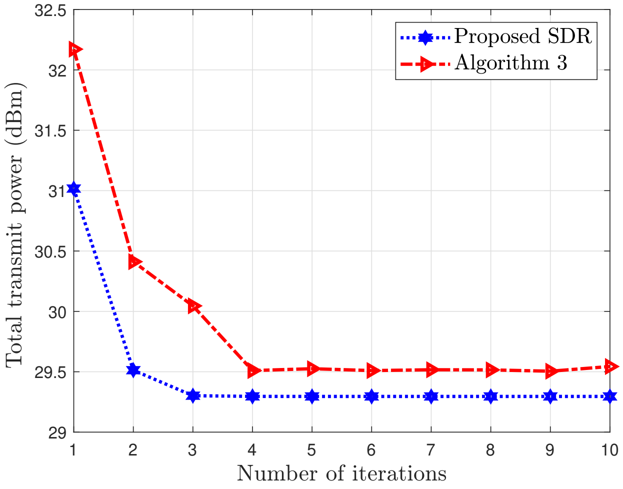
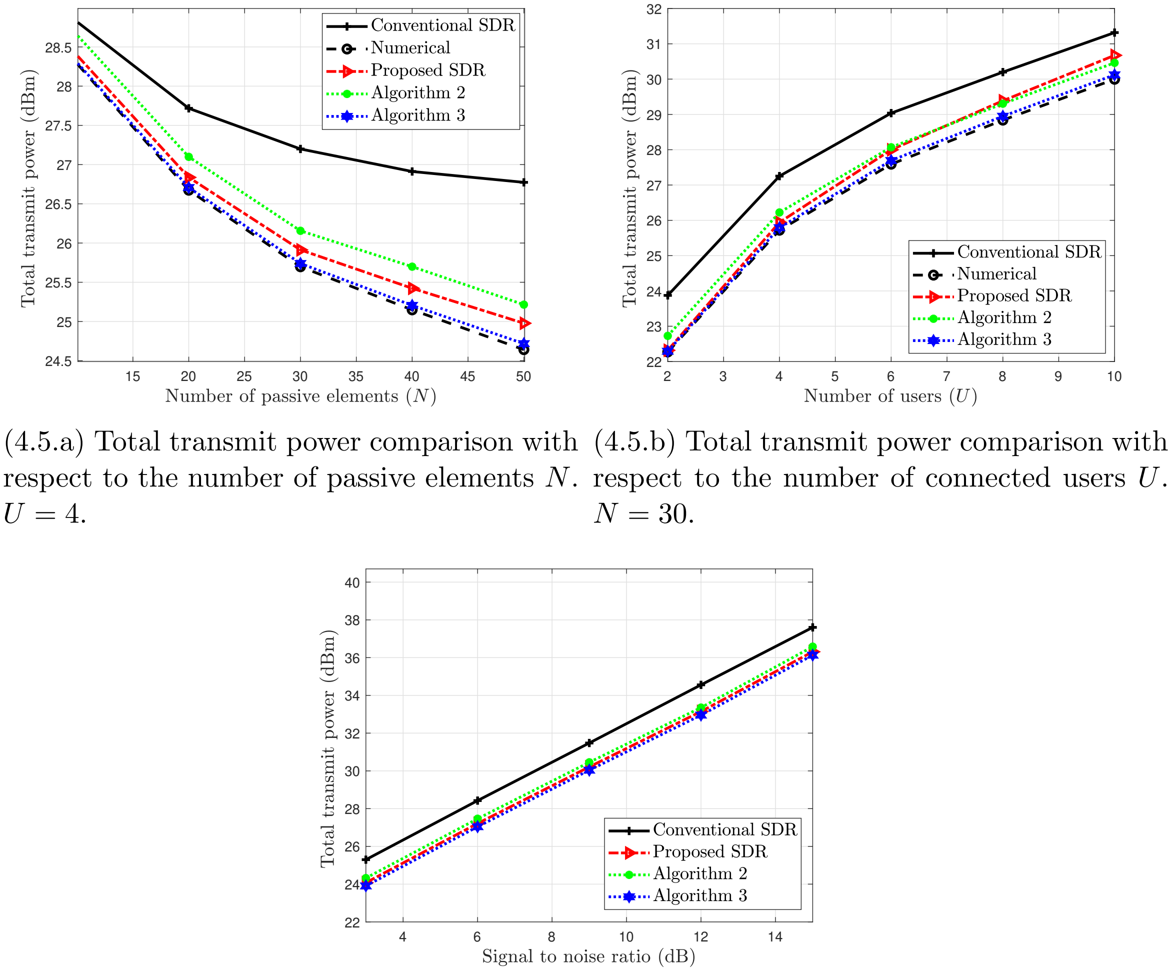


Fig. 4.4: Convergence of the proposed Algorithms.

4.5.1 Convergence of the Proposed Algorithms

Fig. 4.4 depicts the convergence behaviours of the proposed SDR algorithm and the element-wise KKT method (Algorithm 3) when $U = 4$, and $N = 20$. It is clear that the transmit power obtained using the proposed SDR method is higher than that obtained by the proposed element-wise KKT in Algorithm 3. It is also shown that the proposed Algorithm 3 converges to an optimal solution in fewer iterations than the proposed SDR method. Specifically, the proposed Algorithm 3 reaches the solution in 3 iterations while the proposed SDR needs about 4 iterations. This is because the SDR scheme with Gaussian randomization often needs more iterations to obtain a feasible solution close to the dropped rank-one constraint.



(4.5.a) Total transmit power comparison with respect to the number of passive elements N. $U = 4$.

(4.5.b) Total transmit power comparison with respect to the number of connected users U. $N = 30$.

(4.5.c) Total transmit power comparison with respect to SNR requirements γ_u^{th}. $U = 4$ and $N = 30$.

Fig. 4.5: Performance of the proposed schemes in terms of the power consumption dBm. a) We vary the number of passive elements N. b) We vary the number connected users U. c) We vary the SNR requirement threshold γ_u^{th}.

4.5.2 Optimality of Proposed Algorithms

Fig. 4.5 shows a performance comparison between the proposed solutions and the baselines in terms of the power consumption in dBm while varying the size of the RIS (N), the number of connected users (U) and the SNR requirements (γ_u^{th}). Based on the results, we make the following observations.

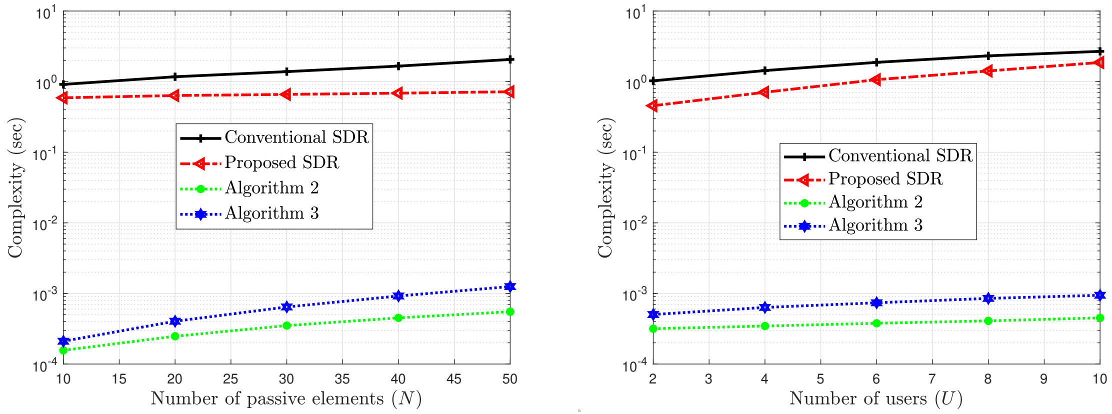
- Intuitively, Fig. 4.5.a shows that the power consumption experienced at the BS is inversely proportional to the number of passive elements, N, equipped at the RIS; as N increases, reflected signals through the RIS will be added constructively at the users and that means less power is required to satisfy the SNR requirements. Similarly, the amount of power consumed at the BS is proportional to the number of connected users, U, and the SNR requirements threshold, γ_u^{th}, as shown in Fig. 4.5.b and Fig. 4.5.c, respectively.

- As shown in 4.5.a, 4.5.b and Fig. 4.5.c, The proposed algorithms have better competitive performance than that achieved by the conventional SDR baseline, and their performances are close to the near-optimal solution obtained numerically. The reason behind this can be explained as follows. Algorithm 2 and Algorithm 3 are obtained through the efficient derived closed-form expression in Theorem 4.1. On the other hand, the proposed linear transformation, in Proposed SDR, improves the efficiency of the Conventional SDR due to the fact that the SDR technique is more efficient in extracting (reaching) rank one constraint when the number of optimization variables decreases.

- Fig. 4.5 also shows that the performance of Algorithm 2 is relatively worse than the other proposed scheme, i.e., Proposed SDR and Algorithm 3, but continues to achieve lower power consumption compared to Conventional SDR. This is because the derived closed-form expression in Theorem 4.1 is approximated and needs to be enhanced iteratively, which is performed with Algorithm 3. Moreover, Algorithm 2 provides a clear intuition that the derived closed-form expression is a good approximation for the optimal solution.

- Fig. 4.5.a shows that by increasing the number of passive elements, the optimality gap between the proposed schemes and the numerical baseline is slightly impacted (increased). Although this gap is still small for the proposed schemes compared to the Conventional SDR, it is crucial to study the performance for large-size RIS, which will be considered next.

- Fig. 4.5.b shows that the number of connected users does not impact the efficiency of the proposed methods Algorithm 2 and Algorithm 3. Specifically, Algorithm 3 still achieves almost the same performance of the near-optimal obtained numerically, i.e., Numerical, for both small and large set of connected users. Similarly, the performance gap between Algorithm 2 and the near-optimal solution obtained numerically, which is still better than that of the Conventional



(4.6.a) Computational time against the number of passive elements N.

(4.6.b) Computational time against the number of connected users U.

Fig. 4.6: Computational time comparison between the proposed schemes and conventional SDR. a) We vary the number of passive elements N with fixed number of users, $U = 4$. b) We vary the number connected users U with fixed number of connected elements $N = 30$.

SDR, is almost the same for both small and large set of connected users. On the other hand, the performance gap of the Proposed SDR and the near-optimal solution increases by increasing U, i.e. after $U = 6$, due to its dependencies on the number of users. Precisely, when increasing the number of users, U, the number of optimization variables increases which reduces the efficient of extracting (reaching) rank one solution using SDR.

4.5.3 Complexity of Proposed Algorithms

Fig. 4.6 depicts the computational time of the proposed schemes Proposed SDR, Algorithm 2, and Algorithm 3 compared to the conventional SDR while varying the number of passive elements and the number of connected users. The figure illustrates that the proposed schemes have lower computational time than the Conventional SDR. The relevant observations on Fig. 4.6 are summarized in the following.

- Fig. 4.6 shows that the computational times of Algorithm 2 and Algorithm 3 are in the order of millisecond which are much lower than the SDR based approach where the computational times are in the order of seconds. For instance, as shown in Fig. 4.6.a, the computation times of Algorithm 2 and Algorithm 3, at $N = 50$, are 12×10^{-4} and 6×10^{-4} seconds, respectively, compared to 2.1 seconds for the Conventional SDR. Similarly, as shown in Fig. 4.6.b, the computation times of Algorithm 2 and Algorithm 3 at $U = 10$, are 10×10^{-4} and 5×10^{-4} seconds, respectively, compared to 2.7 seconds for the Conventional SDR. On other words, the proposed algorithms (Algorithm 2 and Algorithm 3) have a complexity, at least, of ×1000 lower than the Conventional SDR. This observation is expected since Algorithm 2 and Algorithm 3 are all based on the derived closed-form expression which has a complexity of $O(1)$.

- Although the computation time of the Proposed SDR is still higher than that of the proposed Algorithm 2 and Algorithm 3, it remains lower than that of the Conventional SDR. For example, as shown in Fig. 4.6.b, the computation time of the Proposed SDR at $U = 10$ is $2.1 * 10^{-4}$ seconds, respectively, compared to 8 seconds for the Conventional SDR. This means that the Proposed SDR enhances the performance of the SDR approach in terms of the computational time. This is due to the fact that the number of optimization variables is reduced, which implies a lower complexity. Moreover, due to the dependencies of the Proposed SDR on the number of users, increasing the number of users noticeably increases the computing complexity.

- Fig. 4.6.a and Fig. 4.6.b show that increasing both N or U reflects proportionally on the computing, i.e., the computing time at large N or U is higher than smaller large N or U. For clarity, the computing time of Conventional SDR and Algorithm 3 increase from 0.9 seconds at $N = 10$ to 2.1 at $N = 50$. On the other hand, Fig. 4.6.b shows that increasing the number of passive elements has a very limited impact on the computing time of the Proposed SDR as the complexity is, somehow, independent of the number of passive elements, i.e., depend on the number of users.

4.5.4 Scalability

Table 4.2 depicts the performance of the proposed algorithms in terms of power consumption, running time for large size RIS, i.e., $N \in [100, 600]$. For more comparison of the performance of the proposed algorithm and since the computational complexity of the numerical solution is extremely high, we conduct the DC method

Table 4.2: Performance comparison between the proposed schemes and the baselines for large size RIS. $U = 4$ and $\gamma_u^{\text{th}} = 15$ dB and 500 channel realizations.

	Conventional	**Proposed**	**Algorithm**	**Algorithm**	**DC**
	SDR	**SDR**	**2**	**3**	
N	dBm (sec)	dBm (sec)	dBm (sec)	dBm (sec)	dBm (sec)
100	30.98 (6.84)	27.91 (1.73)	28.29 (**0.0014**)	**27.44** (0.0015)	28.07 (15.59)
200	30.15 (40.96)	25.36 (2.27)	25.98 (**0.0030**)	**24.76** (0.0031)	26.34 (78.43)
300	29.61 (131.67)	23.56 (9.59)	24.31 (**0.0085**)	**22.84** (0.0087)	25.47 (273.96)
400	−(−)	22.04 (11.61)	22.91 (**0.0138**)	**21.27** (0.0139)	−
500	−(−)	20.83 (16.31)	21.84 (**0.0191**)	**20.028** (0.0193)	−
600	−(−)	19.71 (20.52)	20.83 (**0.0250**)	**18.88** (0.0254)	−

[91]. It is clear from the results in Table 4.2 that the DC method achieves lower consumption power compared to the Conventional SDR. However, its performance is still slightly worst than the proposed schemes in terms of power consumption and complexity. Moreover, the results observed in Table 4.2 support that the optimality performance of the proposed algorithms in terms of power consumption and computational time. Moreover, Table 4.2 shows that Conventional SDR and DC method are not scalalbe for large size RIS, hence it is not a practical solution. In contrast, the proposed schemes are shown to be reliable and scalable for large size RIS. The table shows that the optimality gap is not much impacted by applying the proposed linear transformation scheme while guaranteeing low complexity of optimizing the RIS phase shifts. Furthermore, the table shows that the proposed linear transformation approach scales down the computational time of optimizing the RIS phase shifts (different transformation functions may be applied), for SDR based solution. Consequently, the proposed transformation scheme may be a base idea which may lead to more efficient strategies to facilitate the RIS optimization in practical settings. On the other hand, the running time of the proposed algorithms based on the derived closed-form expression is still in the order of milliseconds, making them preferred for low-latency services, such as (URLLC). In practice, the processing time of proposed schemes can be even further reduced by utilizing more computing resources available at the network edge. In summary, we have presented different low complexity RIS

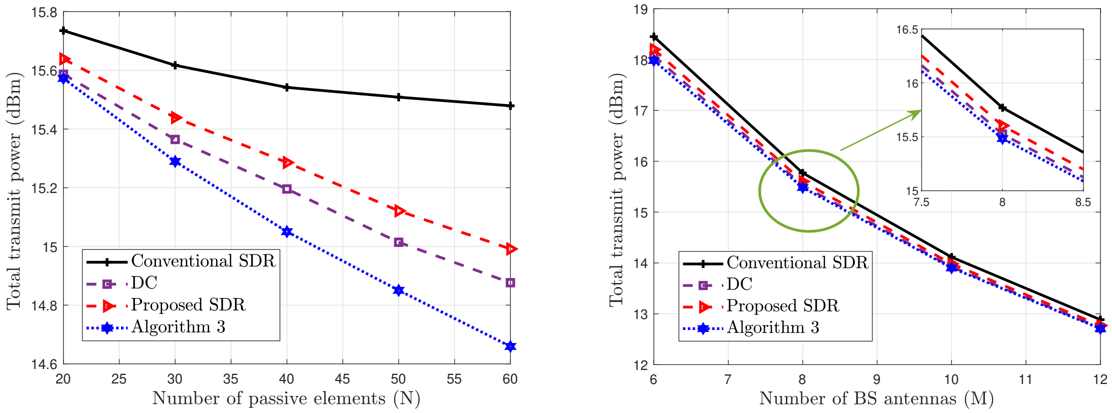
optimization directions/strategies with near-optimal performance and are scalable for large-sized RIS-assisted wireless networks.

4.5.5 Performance for Multiple Antenna with ZFBF

Fig. 4.7 depicts the performance of the proposed Algorithms in the MISO case. In the simulation setting, we adopt the same channel model and the simulation parameters used for the single antenna case. Fig. 4.7.a illustrates the performance of Algorithms while varying the number of passive elements at the RIS. As expected, the conventional SDR has higher power consumption than Algorithm 3, proposed SRR and the DC method. Also, the proposed Algorithm 3 has a slightly better performance than the DC method. In contrast, the proposed SDR has a slightly higher transmit power than Algorithm 3 and the DC method, however, it still perform better than the conventional SDR. Furthermore, the performance gap between Algorithm 3 and both the DC method and proposed SDR increases while the number of passive elements increases. On the other hand, Fig. 4.7.b shows the performance of the proposed Algorithms with respect to the number of antennas at the BS. It shows that by increasing the number of antennas at the BS, the performance gap between the conventional SDR and other schemes decreases. The reason behind this is that the diversity gain of using multiple antennas makes the directed channel link between the BS and users more dominant than the cascaded channel through RIS. Moreover, Fig. 4.7.b shows that the power consumed by the BS reduces by increasing the number of antennas at the BS. For instance, the transmit power decreases form 18 dBm to 14.8 for both Algorithm 3 and the DC method.

4.6 Summary

In this chapter, efficient methods were developed to reduce the complexity of optimizing the RIS phase shifts. We proposed two approaches to enhance the accuracy (optimality) and the complexity of the RIS phase shift design. First, based on the optimal RIS configurations of individual users, we proposed a novel RIS optimization framework wherein linear transformation is used to reduce the number of optimization variables. In this framework, the number of optimization variables is reduced from



(4.7.a) Total transmit power comparison with respect to the number of passive elements N. $U = 4$ and $M = 8$.

(4.7.b) Total transmit power comparison with respect to the number of BS antennas $M.U = 4$ and $N = 50$.

Fig. 4.7: Performance of proposed algorithms in terms of the power consumption dBm. a) We vary the number of passive elements N. b) We vary the number BS antennas M.

the number of passive elements to users, which is generally much smaller compared to the number of passive elements. Second, we proposed three iterative low complexity algorithms based on an approximated closed-form expression for the RIS phase shift design. Extensive simulations were performed to illustrate the performance of the proposed schemes. Simulation results showed that the proposed method has better performance and computational time than the conventional SDR baseline, making them more efficient in terms of optimality and scalability. Moreover, the proposed algorithms based on the derived approximated closed-form expression have a low computational time of milliseconds, making them suitable for practical, especially for the low-latency services (URLLC).

Chapter 5

Exploiting Sequence Similarity for Efficient Puncturing

5.1 Background, Related Works, and Contributions

As discussed in Chapter 1, 5G and beyond systems are anticipated to provide a variety of service classes with different requirements in terms of latency, reliability and connectivity [3, 5, 7]. This naturally raises concerns about their coexistence, especially after it has been shown that allocating a dedicated bandwidth for each service is not spectrally efficient [20]. In particular, providing a dedicated bandwidth for URLLC class of service has been shown to be poorly efficient where the effectively used bandwidth could be less than 5% of the total allocated resource. This is mainly due to URLLC traffic characteristics and requirements [34]. Meanwhile, given the sporadic characteristic of URLLC traffic and their short packet size, the allocated resources will only be used occasionally and for a short period [1, 5, 7, 23]. Therefore, on-demand resource allocation for URLLC transmissions is deemed a promising solution to make good use of spectral resources. Consequently, the 3GPP standard

The work done in this chapter leads to an IEEE published journal [101].

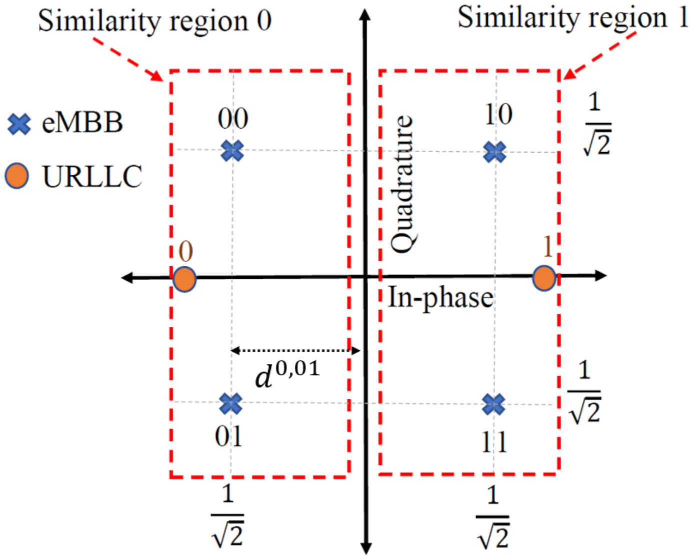


Fig. 5.1: Puncturing mechanism: URLLC and eMBB traffic are modulated using BPSK, and 4-QAM, respectively. URLLC symbols $\{0, 1\}$ are transmitted instead of eMBB symbols $\{00, 10, 01, 11\}$. $d^{0,01}$ is the minimum distance between the eMBB symbol 01 and the decision line of the URLLC symbol 0.

has suggested using superposition/puncturing for multiplexing URLLC and eMBB services in 5G networks [30, 102]. The main idea of the superposition/puncturing framework is to transmit URLLC packets in mini-slot basis, upon their arrival, over the resources occupied by ongoing service type transmissions. Specifically, If the BS allocates transmission power for both eMBB and URLLC traffic, then it is referred to as superposition. If the BS chooses zero transmission power for the eMBB traffic, then this is referred to as puncturing [30]. According to the puncturing mechanism, and assuming that the binary phase shift keying (BPSK) is used for the URLLC traffic and that the 4-Quadrature amplitude modulation (QAM) is used for the eMBB traffic, the URLLC symbols $\{0, 1\}$ are transmitted instead of the punctured eMBB symbols $\{00, 10, 01, 11\}$, as shown in Fig 5.1.

Superposition/puncturing is considered a promising option to allocate the URLLC traffic due to the tolerance of the latter in terms of latency and reliability. Hence, much work [3, 25, 26, 33, 35, 37, 38, 54, 103–105] focused on developing techniques based on coupling URLLC and conventional (eMBB) data transmission through superposition/puncturing. In [3, 25, 26, 33, 35, 37, 38, 54, 103–105], the authors investigated and developed novel superposition/puncturing approaches aiming to minimize the impact on eMBB in terms of the contaminated symbols. The advantages of using superposition for sharing resources in uplink communications between eMBB, mMTC, and URLLC devices were studied in [26]. To enable the non-orthogonal coexistence of URLLC and eMBB services, both uplink and downlink were analysed in [33]. The authors proposed processing URLLC traffic at the network edge to guarantee the latency requirements of the URLLC, while eMBB communication is handled at the cloud radio access network level to achieve high-spectral efficiency for eMBB traffic. For URLLC downlink MIMO-NOMA, network layer performance bounds and cross-layer power control were studied in [103]. A max-matching diversity (MMD) algorithm was proposed in [104] to allocate eMBB users, considering both heterogeneous orthogonal and non-orthogonal multiple access network slicing strategies. A machine learning approach for hybrid multiple access solution (HMA) was proposed in [105]. In fact, the classical methods of NOMA, such as power-domain, require perfect CSI at the BS such that the transmitted signal can be separated at the receiver with successive interference cancellation (SIC)[106]. A new class of NOMA, namely bits similarity NOMA, was proposed in [106]. It was shown that, without perfect CSI, bit-similarity NOMA can achieve better spectral efficiency and fairness among users compared to traditional NOMA techniques.

Another line of research that aims to deal with traffic superposition includes unequal error protection and/or hierarchical modulation, which have been developed mainly for simultaneous transmission of voice and data over fading channels [107–110]. The authors in [108] proposed unequal error protection codes to achieve the best use of channel redundancy. Particularly, the codeword specifies the multiplexing rule, then a codeword is selected from a fixed codebook to convey additional important information. Adaptive hierarchical modulation for simultaneous voice and multi-class data transmission was proposed in [109]. The authors derived closed-form expressions for the outage probability, achievable spectral efficiency, and average BER of both traffics over Nakagami-m fading. Pseudo-noise amplitude shift keying (PN-ASK) modulation, where covert symbols are mapped by shifting the amplitude of primary symbols, was proposed in [110]. One of the shortcomings of superposition techniques is that they may cause severe degradation to the URLLC reliability because the eMBB signal acts as an interference signal that increases the decoding errors of the URLLC traffic. Moreover, the lack of the URLLC CSI at the transmitter decreases the chances to superpose URLLC traffic on the eMBB [33]. Thus, puncturing is preferred as it conserves the URLLC reliability.

In order to study the impact of puncturing eMBB resources to accommodate URLLC transmission, the authors in [3] studied the problem of joint scheduling of eMBB and URLLC data transmission according to linear, convex and threshold models of the eMBB rate loss associated with the eMBB resources puncturing. A risk-sensitive approach was introduced in [35] to mitigate the risk of puncturing eMBB resources. A resource allocation scheduler was proposed in [37] where the formulated problem considered the overhead associated with the URLLC load segmentation while maximizing the rate utility. A null-space-based spatial puncturing scheduler for joint URLLC/eMBB traffic was proposed in [38]. The URLLC allocation problem in [54] was formulated as a dual objectives problem with the objective of maximizing eMBB utility while satisfying the URLLC constraints. The authors in [25] formulated a URLLC traffic allocation problem by adopting a superposition or puncturing scheme. Practically, when the URLLC service is initiated in the middle of the eMBB transport block, part of eMBB symbols are replaced by and/or superposed with the symbols of the URLLC packet. Accordingly, the reception quality of the eMBB services could be degraded severely.

Since eMBB tolerates delays, eMBB users can rely on long error-correction codes in combination with re-transmission techniques to compensate for the loss incurred by superposition/puncturing. *Retransmission-based puncturing* slows the eMBB traffic, and it requires more overhead including puncturing indicator (PI) to inform the eMBB user of the punctured resources, while the whole information block can be re-transmitted if decoding errors occur. Therefore, researchers have been thinking about using codes (*code-based puncturing*) to correct the erroneous symbols in the eMBB message and hence avoiding retransmissions and high overhead signal[20, 30]. Particularly, the gain achieved by retransmission based puncturing over *code-based puncturing* is moderate and less than 10% [20, 30]. Moreover, indicator-free scheme including a transmit precoding with blind detection is proposed for resource overhead reduction [111]. In general, the more punctured eMBB symbols there are, the higher the number of erroneous eMBB symbols and the lower the code rate (of the error correction code) we get, which subsequently results in low spectral efficiency.

In this chapter, we seek to develop a puncturing strategy such that the impact on the punctured eMBB symbols is minimized, which should essentially lead to better eMBB QoS and spectral utility. In other words, we aim at devising a puncturing strategy that can decrease the impact of simultaneous transmissions of URLLC and eMBB traffic on the eMBB traffic. Hence, there is no need to inform the eMBB users about punctured resources, i.e., avoid transmitting costly and unnecessary puncturing indicator signal. The contributions of the proposed downlink puncturing strategy are summarized as follows:

- Different from existing works, we exploit the possible similarity among the URLLC-eMBB symbols instead of random allocation. Indeed, upon the arrival of a URLLC packet, the BS scans the ongoing eMBB transmissions and selects the one that maximizes the number of similar symbols between the two services. In fact, increasing the similarity between the eMBB-URLLC symbols effectively reduces the impacted eMBB symbols and hence the possibility to enhance the used error correction code rate and then the spectral efficiency.

- While developing the proposed technique, we consider the case where an eMBB user could have different symbol constellations than that of URLLC users. Accordingly, we introduce the so-called similarity region to evaluate the similarity between the eMBB and URLLC with different constellations. We describe in detail the encoding and decoding processes for both eMBB and URLLC traffic.

- Taking into consideration the symbol errors occurring due to the channel impairment and the puncturing process, we derive a closed-form expression for the symbol error rate (SER) of the eMBB traffic. The expression shows that the SER of the eMBB traffic depends on the signal-to-noise ratio (SNR), the average URLLC load, and the average similarity. We also consider the SER to measure the reliability of the URLLC traffic, as conserving the SER of the URLLC preserves the minimum packet error rate. Moreover, other reliability improvement techniques, i.e., error control coding schemes [112], packet duplication [113], and HARQ [114], can be used to enhance the URLLC reliability. These enhancement techniques are outside the scope of this work.

- We demonstrate through several numerical examples the efficacy of the proposed scheme where we show that gains of up to 10 dB can be achieved in

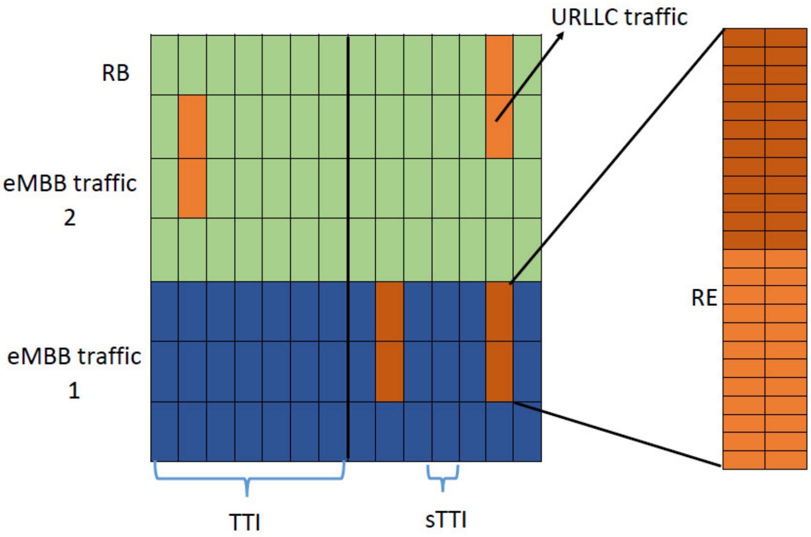


Fig. 5.2: Relation between frequency resources and puncturing mechanism.

comparison to the code-based puncturing technique. At high SNR, the eMBB SER is dominated by error occurring due to puncturing, i.e., the impact of channel diminishes. The opposite is true when the similarity increases, that is, the SER is greatly affected by the channel, not the puncturing. We also show that the proposed algorithm has low complexity computational time making it an efficient solution to be used in practice.

The rest of the chapter is organized as follows. Section 5.2 describes the adopted system model. The proposed puncturing strategy is described in 5.3. Section 5.4 provides performance analysis of the proposed strategy where closed-form expressions for the SER for both eMBB and URLLC users are derived. Numerical and simulations results are shown in 5.5. We conclude the chapter in Section 5.6.

5.2 System Model

We consider a downlink wireless system consisting of one BS that serves certain eMBB and URLLC traffics simultaneously. The system bandwidth is partitioned into L equally sized frequency resources, where each frequency resource is referred to as a resource element (RE). Each 12 REs constitute a resource block (RB) that is equivalent to 180 KHz. The time domain is divided into slots, also known as transmission time intervals (TTIs). The duration of each TTI is 1 ms. To support the low latency requirement of the URLLC traffic, each TTI is further divided into mini-slots, also known as small TTIs (sTTIs), where the duration of each sTTI is 0.143 ms [3]. The REs are assigned to the eMBB traffic at the beginning of each TTI, while the URLLC traffic arriving at each sTTI is directly transmitted in the next sTTI by puncturing the REs belonging to the eMBB load. Each URLLC packet is divided into blocks of ζ-symbols, with $\zeta \geq 1$, and it is allocated within one sTTI. Each eMBB receiver is assumed to decode its received data without knowledge of the punctured resources, i.e, each eMBB receiver is assumed to be unaware of the punctured resources at the transmission. Accordingly, the puncturing overhead is reduced. Based on this assumption, each eMBB receiver decodes its received data according to its decoder. Let $m \in \{2, 4, ..., M\}$ denote the order of the quadrature amplitude modulation (QAM) scheme adopted for the eMBB traffic with symbol error probability $P_m(\gamma_{\mathrm{e}})$, where γ_{e} denote the received eMBB signal-to-noise ratio (SNR). In addition, let $n \in \{2, 4, ..., N\}$ denote the order of the QAM scheme adopted for the URLLC traffic and let ϵ_{u} denote its target symbol error probability.[1] Practically, the URLLC modulation order n is low due to the following reasons. First, the high reliability constraint of the URLLC service in the absence of accurate channel estimation caused by the latency constraints of such traffic. Second, the size of each URLLC packet is assumed to be small, and hence, the achievable capacity follows the short-block regime [115–117].[2] Accordingly, we assume through this work there is no CSI at the BS for the URLLC users, hence BPSK will be used by default to encode the URLLC traffic to achieve the highest URLLC reliability. We list in Table 5.1 most of the variables used in the analysis throughout the chapter.

[1] By definition, both m and n are powers of 2.

[2] Although, the BPSK modulation is adopted for the URLLC traffic in this chapter, the proposed similarity-enhanced puncturing scheme can be extended to support higher modulation orders, such as quadrature-phase-shift-keying (QPSK) and 16-QAM.

Table 5.1: List of variables used in the analysis.

Symbol	Description
Ω	similarity region
n	URLLC modulation order
m	eMBB modulation order
L	BS downlink frequency resources
l	average URLLC traffic
γ_{e}	eMBB Signal to Noise Ratio
γ_{u}	URLLC Signal to Noise Ratio
L_m	eMBB frequency resources with modulation order m
$l_{n,m}$	punctured eMBB symbols of modulation order m by URLLC traffic of modulation order n
$\mathcal{L}_{n,m}$	effectively punctured eMBB symbols of modulation order m by URLLC traffic of modulation order n
$\overline{\mathcal{L}}_{n,m}$	non-effectively punctured eMBB symbols of modulation order m by URLLC traffic of modulation order n
p_m	the probability of encoding eMBB with modulation order m
$P(.)$	eMBB traffic symbol SER
$\mathcal{P}(.)$	URLLC traffic symbol SER
ζ	URLLC block size
$U_{n,m}(.)$	average similar symbols
K	similarity search space
s_u	URLLC symbol
s_e	eMBB symbol

5.3 Proposed Puncturing Scheme

5.3.1 Rationale

The main idea of the proposed puncturing strategy is to exploit the similarity between the symbols of the URLLC and the eMBB loads such that the punctured eMBB symbols are similar to the transmitted URLLC symbols. Instead of puncturing the eMBB traffic greedily by assuming the punctured parts are totally lost, we can search through the eMBB information block to exploit the eMBB sequence that has the highest similarity to the URLLC data block. Hence, some of the eMBB symbols will be received correctly due to their similarity with the transmitted URLLC symbols. Fig. 5.2 illustrates the mechanism of the proposed scheme in terms of frequency and time resources. At each mini-slot, one can transmit two OFDM symbols per RE [118]. In this case, if we consider a wireless network with 100 RBs, then a total of

$100 \times 12 \times 2 = 2400$ ODFM symbols can be transmitted in one sTTI. The BS counts the similarity between the URLLC sequence with the ongoing 2400 eMBB symbols. Then, it punctures the eMBB sequence that has the maximum similarity with the URLLC traffic. Let K denote the size of the search space, or equivalently, the number of possible candidate eMBB RBs to search over for similarity. In addition, let us consider a URLLC load with a length equal tof 2 RB, i.e., $2 \times 12 \times 2 = 48$ OFDM symbols. Assuming that the search window is one RB, the proposed algorithm evaluates the similarity between the URLLC sequence and all available eMBB blocks, i.e., $K = 99$. Afterwards, the BS selects and punctures the eMBB sequence that has the highest similarity to the URLLC sequence.

For more elaboration, let us assume that both eMBB and URLLC services employ binary phase shift keying (BPSK) modulation and that the transmitted URLLC symbol is **0**. Then the punctured eMBB symbol can be either **0** or **1**. If the punctured eMBB symbol is **0**, then the transmitted URLLC symbol and the punctured eMBB symbol are similar, and therefore, the error probability of the eMBB symbol is not affected by the puncturing scheme. However, if the punctured eMBB symbol is **1**, then the eMBB symbol will be received erroneously with a non zero probability. Therefore, it is recommended to puncture the eMBB traffic that has the maximum similarity to the URLLC traffic. Intuitively, increasing the similarity between the transmitted URLLC symbols and the punctured eMBB symbols will reduce the SER at the eMBB receiver, which reduces the retransmissions and the PI overhead.

5.3.2 Similarity Analysis

In practice, the modulation schemes used by the eMBB and the URLLC services can be different. In addition, the eMBB receiver, which is unaware of the punctured part of the transmission, decodes the received signal using a maximum likelihood receiver. Based on this, the probability of receiving the punctured eMBB symbols in error depends on the Euclidean distance between both the transmitted URLLC symbols and the punctured eMBB symbols. As an illustration, let us consider the case when the URLLC traffic employs BPSK modulation and the eMBB traffic employs 4-QAM modulation and let us suppose that the transmitted URLLC symbol is $\mathbf{\{0\}}$. As shown in Fig. 5.1, it is preferred to puncture the eMBB symbols **00** and **01**, since they have the lowest Euclidean distance to **0** as compared to the symbols **10** and **11**, and therefore, a lower resulting error probability. Specifically, as shown in Fig. 5.1, we can say that the symbols **{0, 00, 10}** belong to the same region which we refer to as the similarity region according to the following definition:

Definition 5.1. *Let us consider two QAM schemes with modulation orders* m *and* n*, respectively, and let us consider the diagram that has the superposition of their respective constellations. The similarity region of the two above modulation schemes is a region of the resulting constellation diagram that contains only one constellation point from the modulation that has the lowest order, i.e.* $\min(m, n)$*, and* $\frac{\max(m,n)}{\min(m,n)}$ *constellation points from the modulation that has the highest order, i.e.,* $\max(m, n)$*,* ***which have the minimum Euclidean distance with the constellation point of the modulation that has the lowest order****. Based on this, there exist exactly* $\min(m, n)$ *similarity regions.*

As an illustration for Definition 5.1, let us consider the case when the eMBB traffic has a modulation order of $m = 4$ and the URLLC traffic has a modulation order of $n = 2$. The superposition of the constellations diagrams of the eMBB and URLLC modulations is shown in Fig. 5.1. The resulting diagram can be divided into $\min(m, n) = 2$ similarity regions, namely, similarity region 0 and similarity region 1, where each similarity region contains only one constellation point from the URLLC's modulation and $\frac{\max(m,n)}{\min(m,n)} = 2$ constellation points from the eMBB modulation that have the lowest Euclidean distance with the included constellation point of the URLLC modulation. Moreover, according to Definition 5.1, the eMBB and URLLC symbols are divided into several sets and each set consists of several eMBB and URLLC symbols. The number of eMBB and URLLC symbols depends on the relation between the modulation orders of the eMBB and URLLC. In practice, the modulation schemes of the URLLC and eMBB services may have the same order (i.e, $m = n$) or different ones ($m \neq n$). Accordingly, we classify the relationship between the eMBB and URLLC modulation orders into the following classes:

- Similar-Modulation-Order: In this case, the URLLC and eMBB have the same modulation order, i.e., $m = n$. Hence, each similarity region consists of one eMBB symbol and one URLLC symbol. This symbol is named as the *Region-index-symbol.*

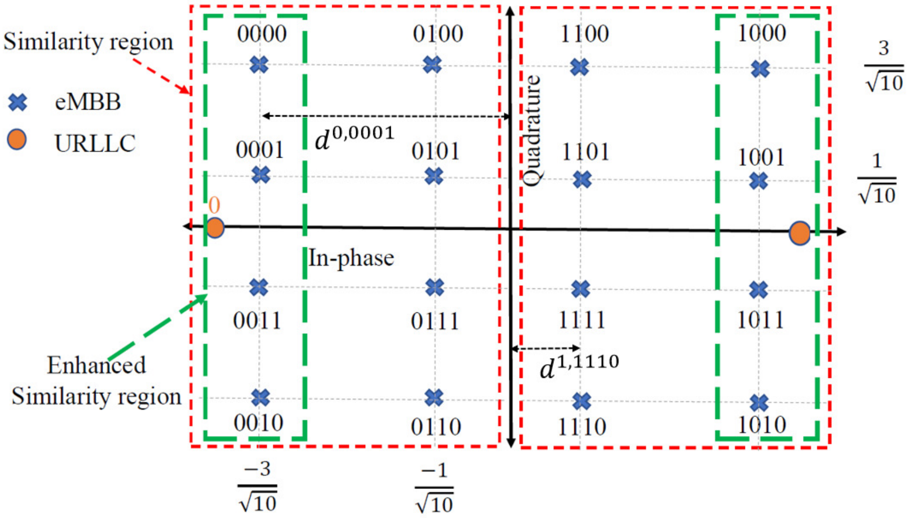


Fig. 5.3: Similarity region between eMBB traffic and URLLC load with 16-QAM (for eMBB) and BPSK (for URLLC). $d^{0,0001}$ is the minimum distance between the eMBB symbol 0001 and the decision line of the URLLC symbol 0.

- Lower-URLLC-Modulation-Order: In this case, the URLLC modulation order is lower than that of the eMBB, i.e., $m > n$. Accordingly, each similarity region consists of one URLLC symbol and $\frac{m}{n}$ eMBB symbols. Similarly, we can rename the URLLC symbol as the *Region-index-symbol* and the eMBB symbols as *mapping-symbols*.

- Higher-URLLC-Modulation-Order: In this case, the URLLC modulation order is higher than that of the eMBB, i.e., $m < n$. Accordingly, each similarity region consists of one eMBB symbol and $\frac{n}{m}$ URLLC symbols. We can rename the eMBB symbol as the *Region-index-symbol* (that denotes the similarity-region) and the URLLC symbols as *mapping-symbols*.

Practically, the Euclidean distance between the *mapping-symbols* and the *Region-index-symbol* varies according to their locations on the constellation. For clarity, as shown in Fig. 5.3, assume the URLLC and eMBB traffic are modulated by the BPSK and 16-QAM, respectively. According to Definition 5.1, the constellation is divided into two similarity regions. The first region has URLLC symbol 0 as the *Region-index-symbol* and the second region has the URLLC symbols 1 as the *Region-index-symbol*. Without loss of generality, let the transmitted URLLC symbol be 0. Then, the *mapping-symbols* belonging to the same similarity region, i.e. $\mathbf{\{0000, \ldots, 0111\}}$, can be treated as $\mathbf{0}$ based on the BPSK maximum-likelihood receiver. On the other hand, the eMBB receiver receives the transmitted symbol correctly; hence the eMBB SER is not degraded. On the other hand, the SER of the URLLC becomes worse, as the symbol energy varies based on the eMBB constellation, which is 16-QAM in this example. We note that symbols **{0000, 0010, 0011, 0001}** have the lowest Euclidean distance with URLLC symbol 0, hence they have lower SER at the URLLC receiver (See Fig. 5.3 for more elaboration.) In light of the above discussion, we define the symbol similarity as follows.

Definition 5.2. *The similarity relation between the Region-index-symbol s_x and the mapping-symbol s_y in the same similarity region can be:*

- *Absolute-similar: if $P(\hat{s} \neq s_x | s_x \text{ was sent}) - P(\hat{s} \neq s_x | s_y \text{ was sent}) \geq 0$.*
- *Strongly-similar: if $-\epsilon \leq P(\hat{s} \neq s_x | s_x \text{ was sent}) - P(\hat{s} \neq s_x | s_y \text{ was sent}) < 0$.*
- *Weakly-similar: if $P(\hat{s} \neq s_x | s_x \text{ was sent}) - P(\hat{s} \neq s_x | s_y \text{ was sent}) < -\epsilon$,*

where $\epsilon \approx 0$ depends on the target URLLC SER. Accordingly, we can call the set of symbols, which are absolute-similar and strongly-similar, as the enhanced similarity region.

Definition 5.3. *The enhanced similarity region is a subset of the similarity region which includes the Region-index-symbol and mapping-symbols that satisfy $P(\hat{s} \neq s_x | s_y \text{ was sent}) - P(\hat{s} \neq s_x | s_x \text{ was sent}) \leq \epsilon$.*

5.3.3 URLLC Encoding at the BS

According to the proposed puncturing strategy, it is preferred to puncture eMBB symbols such that the amount of symbol mismatch between the transmitted $\zeta-$symbols of the URLLC traffic and the punctured eMBB is minimized, i.e., smaller Hamming distance. Based on the URLLC-eMBB relationship, the encoding at the BS is illustrated as follows.

- Similar-Modulation-Order: The BS encodes the URLLC traffic according to the desired modulation order n while puncturing the eMBB symbol sequences that has maximum similarity (Absolute-similar), i.e., maximize the similar eMBB-URLLC OFDM symbols.

- Lower-URLLC-Modulation-Order: Similar to the Higher-URLLC-Modulation-Order case, the BS selects for puncturing the eMBB block that has a maximum number of absolute-similar, strongly-similar, and weakly-similar symbols. To accommodate the URLLC traffic, the BS can transmit either the encoded URLLC symbol or the ongoing eMBB symbol, as described below.

- Higher-URLLC-Modulation-Order: the BS encodes the URLLC traffic according to the desired modulation order n while puncturing the eMBB sequences that have a maximum similarity. In other words, the BS selects the eMBB symbol sequence that maximizes the number of absolute-similar, strong-similar, and weak-similar symbols. Compared to the Similar-Modulation-Order case, the impact of puncturing on the eMBB SER can not be eliminated.

When the URLLC modulation order is lower than that of the eMBB, the BS can transmit the URLLC symbol or keep the ongoing eMBB symbol, as follows.

- URLLC mapper: The BS transmits the encoded URLLC symbols. Hence, the impact of puncturing on the eMBB resources can not be eliminated. To elaborate, let us consider the following example. If we have the following URLLC sequence **{0, 1, 1, 0}**, and the punctured eMBB sequence is **{00, 11, 10, 01}** (See Fig. 5.1). According to the similarity region definition, these URLLC symbols are in the same similarity region of the punctured eMBB sequence. Assuming maximum-likelihood detection, we can roughly say 50% of the punctured eMBB symbols will be correctly received, which translates to a high SER at the eMBB receiver (Fig. 5.1.)

- Similarity region mapper (SRM): To overcome the high SER of eMBB using the URLLC mapper, the BS transmits the eMBB symbol instead of the URLLC symbol, if they belong to the same similarity region, otherwise the URLLC symbol is transmitted. For example, assume that the modulation schemes of URLLC and eMBB are BPSK and 16-QAM, respectively, as shown in Fig. 5.3. Also assume that the transmitted URLLC symbol is **0**. Then, any eMBB symbol belonging to the same similarity region, i.e. **{0000, . . . , 0111}**, will be received as **0** at the URLLC receiver with error probability less than **1**. On the other hand, the eMBB user receives the transmitted symbol correctly as the symbol is not affected by the puncturing process; hence the eMBB SER

Algorithm 5.1: Proposed SRM/ESRM mapper

```
evaluate $\mathcal{P}_1$ ;
if $\mathcal{P}_1 > \epsilon_u$ then
    transmit the encoded URLLC symbols ;
else
    transmit the eMBB symbols that satisfy the similarity conditions in
      Definition 5.2;
end
```

will improve. On the other hand, the SER of the URLLC becomes worse, since the eMBB symbols have different minimum distances from the URLLC decision boundary, which is 16-QAM in this example.

- Enhanced Similarity region mapper (ESRM). To solve the high SER of the URLLC of the SRM, only the eMBB symbol belonging to the same enhanced similarity region, (that have better minimum distances from the URLLC decision boundary), i.e **{0000, 0010, 0011, 0001}**, are transmitted instead the URLLC symbol, otherwise the URLLC symbol is transmitted.(See Fig. 5.3 for more elaboration.)

Algorithm 1, refers to Algorithm 5.1, illustrates an example for the mechanism of the SRM/ESRM. Let ϵ_u and $\mathcal{P}_1$ be the targeted SER of the URLLC traffic and the expected SER if the SR/ESRM is used, respectively. The algorithm starts by checking the activation condition $\mathcal{P}_1 \leq \epsilon_u$, if the activation condition is satisfied, the eMBB symbol is transmitted if they are satisfying the similarity condition, otherwise the URLLC symbol will be sent. We emphasize here that the bit streams of both eMBB and URLLC traffics are the final bit streams to be transmitted by the BS, i.e., after the source/channel coding and modulation and just before transmission. Fig. 5.4 shows a real deployment of the proposed SRM/ESRM mapper. As illustrated in this figure, the proposed scheme exploits only the possible similarities between the URLLC segment and the eMBB symbols of the final bit streams.

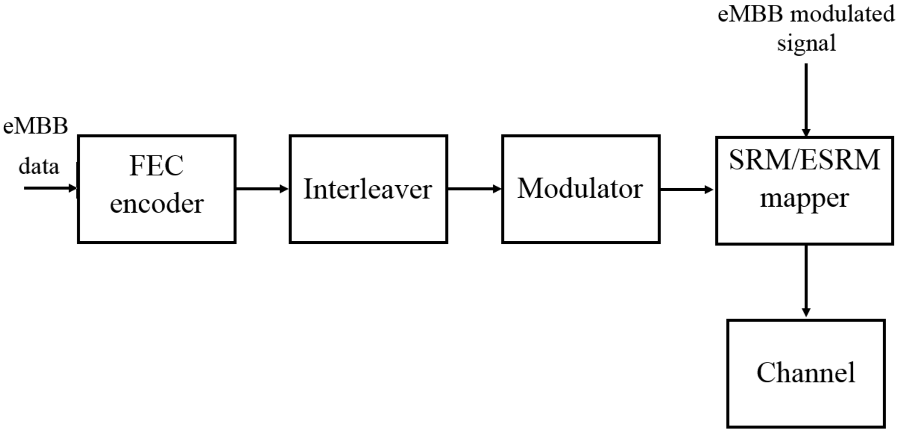


Fig. 5.4: Transmitter block diagram.

5.3.4 URLLC and eMBB Decoding

As mentioned above, the eMBB receiver does not have the knowledge of the punctured symbols and it decodes the received sequence as if no puncturing took place. On the other hand, the URLLC receiver will perform the decoding process normally based on their modulation scheme. For illustration, as shown in Fig. 5.1, we assume that BPSK and QPSK are used to encode both the URLLC and eMBB traffic, respectively. Let the transmitted URLLC symbol be **{0}**. The eMBB decoder will translate the received symbol with probability($\approx 50\%$) as **{00, 01}**. Similarly, the URLLC receiver will decode the received symbol **{00}** as **{0}** with probability $(1 - P_2(0.5\gamma_{\mathrm{e}}))$. For the case when the ESRM is used, i.e., keep transmitting the similar eMBB symbols, the URLLC receiver will decode the received signal assuming the desired decoder, BPSK in this example. Although the concept of the proposed scheme and hierarchical modulation are interference-free multiplexing schemes [107–110], they are based on completely disjoint concepts. In fact, in hierarchical modulation, also called layered modulation, two or more users' streams are modulated onto one stream with higher modulation order that is equal to the product of the all streams' constellation sizes. The proposed scheme, meanwhile, is based on the concept of similarity between the eMBB/URLLC transmitted symbols, where the final steam is modulation onto a single-layer constellation.

5.4 Performance Evaluation

5.4.1 eMBB SER Analysis

Although both SER and bit error rate (BER) can be used to represent the impact of puncturing on the eMBB, the SER can represent the puncturing in the RE (symbol) level instead of the RB level which gives more sense about the proposed strategy.[3] Hence, we use the SER, denoted by P, to measure the impact of the proposed puncturing strategy on the eMBB traffic. Without loss of generality, let $L_m = p_m L$ be the average number of eMBB symbols with modulation scheme m. Also, let the average number of eMBB symbols punctured due to the URLLC traffic with modulation order n be $l_{n,m}$. According to the total probability theorem, the SER of the eMBB traffic under the effect of both the wireless channel and the presence of URLLC load can be expressed as

$$\begin{aligned} P\left(\gamma_{\mathrm{e}}, l\right) = \sum_{m=2}^{M} p_m \times \Bigg[& P_m(\gamma_{\mathrm{e}}) \times \left(1 - \frac{\sum_{n=2}^{N} l_{n,m}}{L_m}\right) \\ & + \sum_{n=2}^{N} P_{n,m}\left(\gamma_{\mathrm{e}}, l_{n,m}\right) \times \frac{l_{n,m}}{L_m} \Bigg], \end{aligned} \tag{5.1}$$

where $P_m(\gamma_{\mathrm{e}})$ is the SER of the eMBB traffic with modulation order m due to the channel error only, and $P_{n,m}(\gamma_{\mathrm{e}}, l_{n,m})$ is the SER due to the channel and URLLC traffic with modulation order n. To quantify the actual effect on the eMBB traffic, we start with the following definition.

Definition 5.4. *Consider an eMBB and URLLC traffic with modulation orders m and n, respectively. The average effectively punctured symbols, $\mathcal{L}_{n,m}$, of eMBB traffic is a portion of the punctured eMBB symbols, $l_{n,m}$, in which the transmitted URLLC symbol, s_u, has a different similarity region from that of the punctured eMBB symbol, s_e, and its range is $0 \leq \mathcal{L}_{n,m} \leq l_{n,m}$.*

[3] We assume uncoded system is being used because this work is a proof of concept for the proposed puncturing scheme. However, the performance of the proposed scheme considering forward error correction will be addressed in future work.

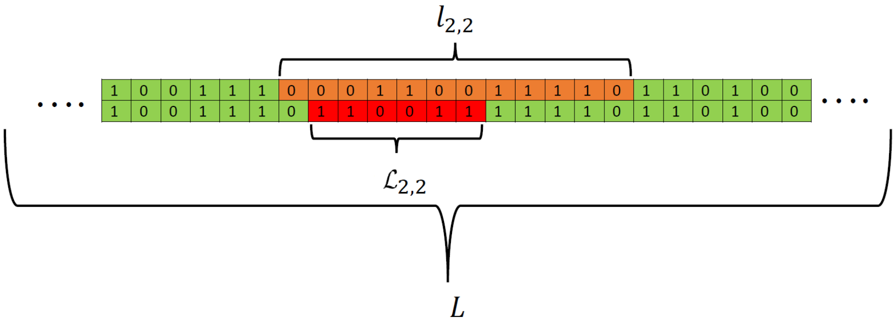


Fig. 5.5: Effectively punctured eMBB symbols.

Fig. 5.5 illustrates the relation between $\mathcal{L}_{n,m}$ and $l_{n,m}$ for the case of similar modulation schemes, BPSK in this example. It shows that, due to the similarity region between the punctured eMBB and URLLC symbols, some of the punctured eMBB symbols are not affected by the puncturing process. In general, for eMBB and URLLC traffic with modulation orders m and n, respectively, the expected number of the *effectively eMBB punctured* symbols is $\mathcal{L}_{n,m} <= l_{n,m}$. Accordingly, (5.1) can be written as follows.

$$\begin{aligned} P\left(\gamma_{\mathrm{e}}, l\right) = \sum_{m=2}^{M} p_m \times \Bigg[P_m\left(\gamma_{\mathrm{e}}\right) \times \left(1 - \frac{\sum_{n=2}^{N} l_{n,m}}{L_m}\right) + \\ \sum_{n=2}^{N} \left(P_{n,m}(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}) \times \frac{\overline{\mathcal{L}}_{n,m}}{L_m} + P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right) \times \frac{\mathcal{L}_{n,m}}{L_m} \right) \Bigg], \end{aligned} \tag{5.2}$$

where $P_{n,m}(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m})$ and $P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right)$ are the error probabilities (which will be derived in the sequel) of the eMBB traffic in $\overline{\mathcal{L}}_{n,m}$ and $\mathcal{L}_{n,m}$, respectively. Where $\overline{\mathcal{L}}_{n,m} = l_{n,m} - \mathcal{L}_{n,m}$.

Equation (5.2) shows the SER of the eMBB traffic under the impact of the wireless channel and the presence of URLLC load. The first term represents the average error probability for the fraction of eMBB sequence impacted by the channel errors only. The second term represents the average error probability of the punctured eMBB symbols that have the same similarity region to the URLLC symbols. The third term is the average error probability of the effectively punctured eMBB symbols.

5.4.2 Puncturing Parameters Evaluation

This section analyzes the parameters $\mathcal{L}_{n,m}$, $P_{n,m}(\gamma_{\mathrm{e}},\overline{\mathcal{L}}_{n,m})$ and $P_{n,m}\left(\gamma_{\mathrm{e}},\mathcal{L}_{n,m}\right)$ for the proposed puncturing strategy. The puncturing parameters depend on the modulation schemes of eMBB and URLLC, and how the URLLC traffic is distributed, i.e., $l_{n,m}$. Without loss of generality, the average punctured eMBB symbols (the average URLLC load) is assumed to be known as it depends on the arrival rate, λ, of the URLLC traffic. Also, the channel error, $P_m(\gamma_{\mathrm{e}})$, depends on the channel and the modulation scheme. For example, the SER of eMBB under additive white Gaussian noise (AWGN) and/or Rayleigh fading with SNR per symbol γ_e are [119, 120]:

$$P_m(\gamma_{\mathrm{e}}) \approx \begin{cases} 4a\,Q(\sqrt{\frac{3\gamma_e}{m-1}}), & \text{AWGN}, \\ 2a\,(1-b) - a^2(1-\frac{4b}{\pi}\tan^{-1}(\frac{1}{b})), & \text{Rayleigh}, \end{cases} \tag{5.3}$$

where $a = (1-\frac{1}{\sqrt{m}})$ and $b = \sqrt{\frac{3\gamma_{\mathrm{e}}}{2(m-1)+3\gamma_{\mathrm{e}}}}$.

5.4.2.1 Average Effectively Punctured Symbols

Without loss of generality, assume the distribution of the symbol similarity between a URLLC block and a punctured eMBB sequence with modulation orders n and m follows the Binomial distribution $B\ (\zeta, \eta_{n,m})$. Therefore, the average similarity between the two traffic blocks can be defined as $U_{n,m,\zeta} \triangleq \zeta \times \eta_{n,m}$, where $\eta_{n,m}$ (probability that any two symbols are similar) is

$$\eta_{n,m} = \sum_{j=0}^{n-1}\sum_{i=0}^{m-1} p_j\, p_i, \forall i,j \in \Omega, \tag{5.4}$$

where p_j and p_i are the probabilities of sending symbol j, i of the eMBB and URLLC traffic, respectively. Accordingly, we can represent the average effectively punctured eMBB symbols, in terms of the average URLLC load in the following definition.

Definition 5.5. *Consider a URLLC and eMBB traffic with modulation orders n and m, respectively. Also, let the eMBB block length and the average length of punctured* *eMBB symbols by the URLLC traffic be L_m and $l_{n,m}$, respectively. The average effectively punctured eMBB symbols (i.e., modified) is given by:*

$$\mathcal{L}_{n,m} = \left(1 - \frac{\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m)}{\zeta}\right) l_{n,m}\,(symbols), \tag{5.5}$$

where $\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m)$ is the expected number of eMBB symbols which have the same similarity region with the transmitted $\zeta-$symbols of URLLC block.

The term $\frac{\mathcal{U}_{n,m}(\zeta,l_{n,m},L_m)}{\zeta}$ represents the ratio (percentage) of similarity between both services. Then, $\left(1 - \frac{\mathcal{U}_{n,m}(\zeta,l_{n,m},L_m)}{\zeta}\right) l_{n,m}$ is the actual punctured eMBB symbols. The definition in (5.5) gives an expression for the average length of the effectively punctured (modified) eMBB symbols. However, it does not quantify the average similarity, $\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m)$, between the URLLC and eMBB sequences. In fact, $\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m)$ depends on the URLLC block size and the eMBB search space of size $N < L$. Lemma 5.1 gives an approximated value for $\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m)$.

Lemma 5.1. *Let L_m denote the average eMBB traffic, and let $l_{n,m}$ be the average number of punctured eMBB symbols. Assume that the URLLC traffic is divided into blocks with ζ-symbols each. An upper bound on the expected similarity between the URLLC and eMBB traffic is given by*

$$\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m) = \frac{1}{\lceil \frac{l_{n,m}}{\zeta} \rceil} \sum_{1}^{\lceil \frac{l_{n,m}}{\zeta} \rceil} \textstyle\sum_{k=0}^{\zeta-1} \left[1 - \{F(k)\}^{L_m - \zeta}\right], \tag{5.6}$$

where, $F(k) = \sum_{j=0}^{k} \begin{pmatrix} \zeta \\ j \end{pmatrix} (\eta_{n,m})_j\,(1 - \eta_{n,m})^{\zeta - j}$.

Proof. See Appendix C for the proof. □

For large ζ, $\mathcal{U}_{n,m}(\zeta, l_{n,m}, L_m)$ reduces to $\zeta\eta_{n,m}$. Hence, we can further reduce the block size, i.e., ζ, and consequently the average similarity will increase. Practically, decreasing the size of ζ will increase the signalling overhead. Therefore, a proper selection of ζ is important. Indeed, we examine the effect of different values of ζ on the performance of the proposed scheme in Section 5.5.

5.4.2.2 SER of the Effectively Punctured Symbols

The SER of the Effectively Punctured Symbols, $P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right)$, strictly depends on the relation between the modulation schemes of both URLLC and punctured eMBB traffic. $\mathcal{L}_{n,m}$ is the average number of eMBB symbols that are wrongly transmitted. In other words, the transmitted symbols belong to another region. $P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right)$ is then expressed as

$$P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right) = \sum_{s_j \in \Omega} \sum_{s_i \notin \Omega} p(s_j | s_i \text{ sent}) P_{err}(s_j | s_i \text{ sent}), \tag{5.7}$$

where s_j and s_i are the effectively punctured eMBB symbol and the transmitted URLLC symbol, respectively. We can upper bound $P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right)$ with the closed form expression

$$P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right) \leq 1 - P_n\left(\gamma_{\mathrm{e}}\right) \times \left(\frac{1}{m-1}\right) \approx 1, \tag{5.8}$$

where $P_n\left(\gamma_{\mathrm{e}}\right)$ is the probability of error under the URLLC modulation condition. For example, let BPSK be the modulation order used by both URLLC and eMBB traffic and $P_2(\gamma_{\mathrm{e}}) = 10^{-2}$, then $P_{2,2}(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}) = 1 - 10^{-2} \approx 1$.

5.4.2.3 SER of the non-Effectively Punctured Symbols

The SER of the non-Effectively punctured symbols, $P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right)$ is summarized as follows.

- Similar-Modulation-Order: The modulation schemes of URLLC and eMBB are similar. Accordingly, non-effectively punctured symbols are similar to the punctured symbol, hence $P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right)$ is expressed as:

$$P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right) = P_m\left(\gamma_{\mathrm{e}}\right). \tag{5.9}$$

- Lower-URLLC-Modulation-Order: similar to Similar-Modulation-Order, in this case, the non-effectively punctured eMBB symbols are only affected by the channel conditions. Hence, $P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right)$ is expressed as follows:

$$P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right) = P_m\left(\gamma_{\mathrm{e}}\right). \tag{5.10}$$

- Higher-URLLC-Modulation-Order: As the energy of the transmitted URLLC symbols varies, an exact expression for the URLLC SER is not easy to obtain. Hence, we obtain an upper bound for $P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right)$ based on the minimum distance, $d^{i,j}$, the transmitted URLLC symbol and the decision boundary of the eMBB symbol. Accordingly, $P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right)$ is expressed as:

$$P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right) = \sum_{s_j\in\Omega}\sum_{s_i\in\Omega} p(s_j|s_i \text{ sent}) P_{err}(s_j|s_i \text{ sent}), \tag{5.11}$$

where $P_{err}(s_j|s_i \text{ sent})$ is expressed as

$$P_{err}(s_j|s_i \text{ sent}) = P_j(\gamma_{\mathrm{e}} {d^{i,j}}^2). \tag{5.12}$$

5.4.3 URLLC SER Analysis

The SER of the URLLC traffic is only affected when the URLLC traffic is modulated using the SRM/ESRM. This is because the energy of the transmitted symbols are different than the actual URLLC symbols. In other words, the energy of the non-effectively punctured eMBB symbols is varies. Accordingly, an exact expression for the URLLC SER is not easy to obtain. Hence, we drive an upper bound expression for the URLLC SER based on the minimum distance of the transmitted symbol and decision boundary, as

$$\begin{aligned}\mathcal{P}_{n,m}\left(\gamma_{\mathrm{u}}\right) = &\left(1 - \sum_{s_j}\sum_{s_i} p\left(s_i|s_j \text{ sent}\right)\right)\mathcal{P}_n\left(\gamma_{\mathrm{u}}\right)\\ &+ \sum_{s_j}\sum_{s_i} p\left(s_i|s_j \text{ sent}\right)\mathcal{P}_n\left(\gamma_{\mathrm{u}} {d^{i,j}}^2\right).\end{aligned} \tag{5.13}$$

Equation (5.13) shows the SER of the URLLC traffic when the SRM/ESRM is used. The first term represents the average error probability for the fraction of

URLLC sequence impacted by the channel errors only. The second term represents the average error probability of the URLLC symbols that have the same similarity region to the eMBB symbols (encoded by the SRM). In fact, the SER loss is equivalent to a power loss, W_{dB} which has the following expression:

$$W_{dB} \approx 10 \sum_{s_j} \sum_{s_i} p\left(s_i | s_j \text{ sent}\right) \log_{10}\left({d^{i,j}}^2 / {d^{i}}^2\right). \tag{5.14}$$

The expression in (5.14) evaluates the average URLLC loss in dB. The term $\log_{10}\left({d^{i,j}}^2 / {d^{i}}^2\right)$ is the power loss for each URLLC symbol in terms of the ratio between the distance of the URLLC and the transmitted eMBB symbols form the decision boundary.

5.4.4 SER Scaling

In light of the above discussion, we can observe that the eMBB SER is a function the SNR and the average similarity of the punctured eMBB symbols. When the SNR increases, the SER improves and it is asymptotically equal to the puncturing errors. Then, the eMBB SER based on (5.2) is approximated as

$$P\left(l\right) \approx \sum_{m=2}^{M} p^m \frac{\sum_{n=2}^{N} \mathcal{L}_{n,m}}{L_m}. \tag{5.15}$$

On the other hand, as the similarity increases (L increases or ζ decreases), the eMBB SER reduces to only channel errors, which can be approximated as

$$\begin{aligned} P\left(\gamma_{\mathrm{e}}, l\right) \approx \sum_{m=2}^{M} p_m \times \Bigg[& P_m\left(\gamma_{\mathrm{e}}\right) \times \left(1 - \frac{\sum_{n}^{N} \overline{\mathcal{L}}_{n,m}}{L_m}\right) \\ & + \sum_{n}^{N} P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right) \frac{\overline{\mathcal{L}}_{n,m}}{L_m} \Bigg]. \end{aligned} \tag{5.16}$$

We observe that the eMBB performance strictly depends on the SNR and the average similarity. As the SNR increases, the eMBB SER becomes dominated by the puncturing errors, while increasing the similarity will reduce the SER to errors due to the channel condition.

5.4.5 eMBB Loss Function

The function that represents the eMBB loss associated to the puncturing schemes can be either a linear, convex or a threshold function [3]. The linear loss function, i.e. $h(x) = \alpha\, x$, has been largely used to study the impact of the superposition/puncturing scheme [3, 35]. The expected loss of an eMBB traffic, based on the linear function, is the ratio between the punctured eMBB symbols to total eMBB symbols:

$$E[h\left(l_m\right)] = \frac{\sum_n^N l_{n,m}}{L_m}, \tag{5.17}$$

where $l_m = \sum_n l_{n,m}$. The loss function in (5.17) is widely coupled with the eMBB rate [3, 26, 35, 37, 103–105]. Taking into consideration the effectively punctured symbols and similar symbols, and using results in (2) and (5), the expected loss in (5.17) can be generalized as follows:

$$\begin{aligned} E[h\left(l_m\right)] = \frac{1}{L_m} \sum_n^N \Big(& P_{n,m}\left(\gamma_{\mathrm{e}}, \overline{\mathcal{L}}_{n,m}\right) \overline{\mathcal{L}}_{n,m} \\ & + P_{n,m}\left(\gamma_{\mathrm{e}}, \mathcal{L}_{n,m}\right) \mathcal{L}_{n,m} \Big). \end{aligned} \tag{5.18}$$

For clarity, assume a retransmission-based puncturing is adopted. Then, all punctured eMBB symbols are lost, i.e., $E[h\left(l_m\right)] = \frac{l_m}{L_m}$.

5.4.6 Proposed Search Algorithms

The latency constraint is a critical factor to maintain QoS of the URLLC service. Therefore, we present a fast search algorithm, of time complexity $O(K)$, that exploits the similarity between the URLLC block over multiple eMBB traffic sequences with different modulation schemes. Initially, the algorithm associates a counter c_k to each eMBB in K. The subset, K, of the possible eMBB blocks for puncturing depends on the latency requirement of the URLLC traffic. Specifically, K should be small for the URLLC traffic with strict latency constraint, and a relatively larger K otherwise. Moreover, the URLLC traffic with weak latency constraint implies that several mini-slots (URLLC-slot) is allowed for allocating the URLLC block.

As shown in **Algorithm 2**, refers to Algorithm 5.2, the proposed algorithm has

Algorithm 5.2: Proposed search Algorithm

```
$c_k \leftarrow 0\, \forall k \in [1, K]$;
Step 1: Similarity weight calculation;
for  $k = 1 \rightarrow K$ do
    for $t = 1 \rightarrow \zeta$ do
        count symbols in the same similarity region;
        if $s_e^t \,\&\, s_u^t \in \Omega$ then
            $\mathbf{c}_k \leftarrow \mathbf{c}_k + 1$;
        end
    end
end
Step 2: eMBB block selection;
$k^* \leftarrow \arg\max_{k \in [1,K]} c_k$;
```

two steps: the first step counts the similar symbols between the eMBB blocks and the URLLC block; the second step selects the suitable eMBB block for puncturing. The first step contains two loops, inner and outer loop, which describe the number of eMBB blocks for possibly punctured and the number of URLLC symbols per block. As illustrated in Algorithm 2, the similarity region of the eMBB symbol, s_e^t, is compared with the similarity region of the URLLC symbol, s_u^t. Accordingly, the counter c_k is incremented by one when both symbols are in the same similarity region, otherwise it remains not incremented. In the second part, the BS selects the eMBB block that has maximum similarity with the URLLC block. In other words, the punctured eMBB block k^* should have a maximum count of similar symbols with the URLLC block. **Algorithm 3**, refers to Algorithm 5.3, performs cross correlation between the eMBB symbols and the URLLC symbols to find the eMBB block that has the highest similarity with the URLLC block. [4]

To mitigate adverse impact on eMBB traffic, the URLLC load is segmented into smaller blocks, i.e., 1 RB [37]. Hence, Algorithm 3 guarantees that all URLLC segments are delivered in the correct order, i.e., the URLLC symbols sequence is correct. The algorithm initially divides the search space, K, into ordered and equal subsets,

[4] The implemented Matlab function xcorr() can be used to evaluate the correlation coefficient of Algorithm 3.

Algorithm 5.3: Correlation based search Algorithm

```
$c_k \leftarrow 0\,\forall k \in [1, K]$;
**Step 1:** Similarity weight calculation;
**for** $k = 1 \rightarrow K$ **do**
    evaluate the cross correlation between the eMBB and the URLLC
      signals,$x_e^k$ and $x_u$, respectively ;
    $\mathbf{c}_k \leftarrow E\{x_u, x_e\}$ ;
**end**
**Step 2:** eMBB block selection;
$k^* \leftarrow \arg\max_{k \in [1,K]} c_k$;
```

$K_1 < K_2 < K_3 ... < K_Z$, where Z is the number of URLLC segments. Then, Algorithm 2 is applied to each segment on the corresponding subset. To enhance the algorithm, the remaining eMBB blocks of subset, K_k, which satisfy the inequality $k_z^* < k_z < K_k$, are merged with the next subset, K_{k+1}, using the $merge()$ function.

Search algorithm time complexity: It can be easily shown that the proposed algorithm has a time complexity of $O(K)$ which makes it an efficient and practical solution. For clarity, the search algorithm (Algorithm 2) consists of two steps, namely, **Step 1** and **Step 2**. **Step 1** consists of one outer loop and one inner loop of K and ζ iterations, respectively. Hence, for a URLLC block with fixed number of symbols ζ, **Step 1** has a time complexity of $O(K)$. On the other hand, **Step 2** aims to select the maximum counter over K elements. In general, **Step 2** performs K comparisons which means it has a time complexity of $O(K)$. As a result, the proposed algorithm (Algorithm 2) has a low time complexity of $O(K)$.

Considering the overhead resulting from the similarity search operation, it is negligible in practice. In fact, the total overhead ϱ, measured in bits, of the proposed algorithm can be evaluated in practice as follows.

$$\varrho = Z \times \log_2(L/12) + Z \times \lceil \log_2(\frac{K}{L/12}) \rceil + Z \times \lceil \log_2(\frac{\zeta}{24}). \rceil \tag{5.19}$$

Accordingly, for the case when $L = 1200$, $\zeta = 256$, $K = 300$, and no segmentation, the overhead is equal to $\varrho 11 + 2 + 4 = 17$ *bits*, which is an extremely low overhead.

Algorithm 5.4: Search segmentation Algorithm.

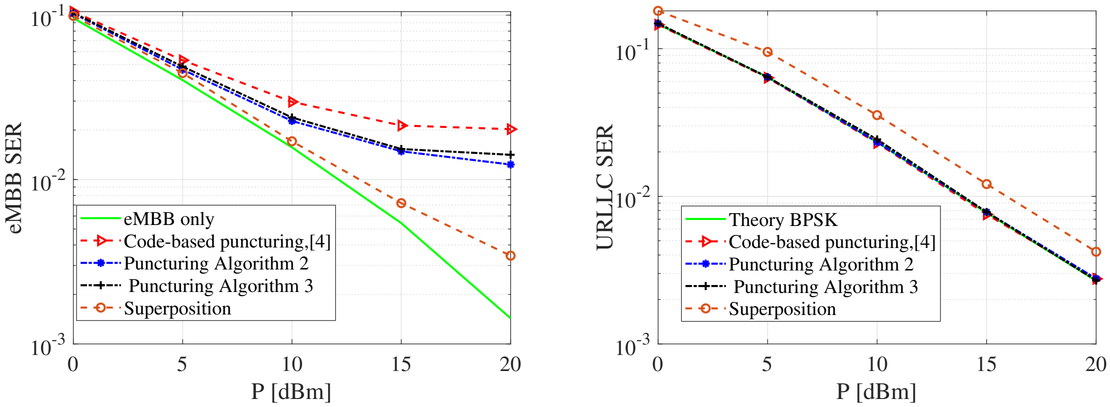
```
$k_z^* \leftarrow 0\,\forall z \in [1, Z]$;
for  $z = 1 \rightarrow Z$ do
    $k_z^* \leftarrow Algorithm\,1(K_z)$;
    $N_{k+1} \leftarrow merge(k_z^* + 1 \rightarrow K_z, K_{z+1})$;
end
```

5.5 Simulation Results

5.5.1 Simulation Settings

In this section, we carry out various simulations to evaluate the performance of the proposed puncturing strategy. We consider a wireless network which consists of one BS and eMBB and URLLC traffic. We assume that the eMBB traffic belongs to 10 eMBB users and the URLLC packet arrival follows the Poisson distribution with arrival rate, λ and each packet size is $\zeta = 96$ and $\zeta = 256$ bits. We also assume that the BS has $L = 1200$ downlink frequency resources (RE). We assume the channel between the BS and the eMBB and URLLC receivers is Rayleigh fading, wherein the eMBB channel gain remains constant for two time slots (14-sTTI) and one eMBB block is transmitted within this period. Moreover, we assume the served users are located in different distance from the BS, hence the path loss is considered with a path loss exponent equals 3. The noise at the receiver is assumed to be complex AWGN $\mathcal{CN}(N_0, 0)$, where $N_0 = 10^{-3}$ is the noise power. The BS can use BPSK, 4-QAM, 16-QAM or 64-QAM to modulate the eMBB traffic. Particularly, the BS adopts the modulation order $m \in \{4, 16, 64\}$ such that the channel SER is less than or equal 0.01, otherwise BPSK is adopted. Moreover, we assume that the CSI of the URLLC traffic is not available at the BS. Hence, the URLLC traffic is modulated using only BPSK to achieve the maximum reliability.

The performance of the similarity puncturing strategy is evaluated for different transmitting power and arrival rate, i.e., $\lambda = 7$ and $\lambda = 3.5$ packets per millisecond (p/msec). we assume that the eMBB users are unaware of the punctured resources, so we consider the code-based puncturing proposed in [20, 30] as a baseline algorithm. Generally, when the eMBB receiver is unaware of the punctured resources of the transmission, the received signal is decoded as useful signal. Resource proportional (RP) placement is used to allocate the URLLC traffic as it gives the optimal solution



(5.6.a) eMBB SER vs transmission power in dBm.

(5.6.b) URLLC SER vs transmission power in dBm.

Fig. 5.6: Puncturing against superposition in terms of SER of eMBB and URLLC.

for the linear loss model [3].

In the following, we start by illustrating the performance of the proposed puncturing scheme compared to the superposition solution. Second, we show the advantage of the proposed strategy on the performance of the eMBB traffic in terms of the spectral efficiency, SER and reliability. Finally, we investigate the performance of the proposed strategy on the URLLC traffic by considering the URLLC SER and reliability.

5.5.2 Puncturing Against Superposition

We compare in Fig. 5.6 the performance of the puncturing and superposition schemes with respect to the transmitted power in dBm. As shown in the figure, the superposition scheme achieves better eMBB SER compared to that of the puncturing scheme. This is attributed to the fact that, unlike puncturing, the superimposed eMBB signal can be recovered using the upgraded hierarchical receiver such as SIC. Moreover, the superposition scheme requires a control signal to help recover the superimposed eMBB signal. In this case, the control signal is mandatory to inform the eMBB receivers of the superimposed resources. Furthermore, the superposition scheme severely impacts the URLLC SER as illustrated in Fig. 5.6.b. This suggests that puncturing is preferred as it conserves the URLLC reliability.

Fig. 5.6 shows that Algorithm 2 and Algorithm 3 have similar performance in terms of SER of both eMBB and URLLC traffics. However, Algorithm 2 achieves the

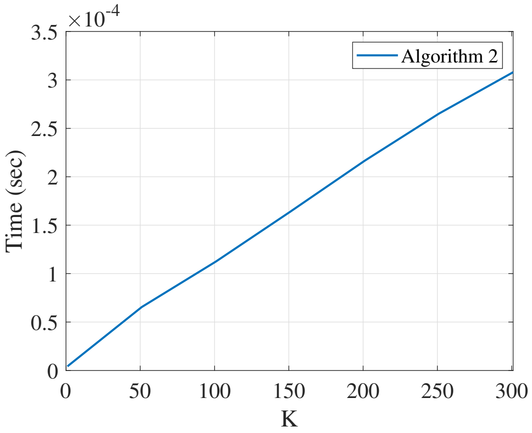


Fig. 5.7: Time complexity of the proposed algorithm

optimal solution since it exhaustively calculates similar symbols through all eMBB blocks. Consequently, Fig. 5.7 shows the time complexity of the proposed algorithm. The algorithm was implemented in Matlab using a machine with the following characteristics: System Type: x64-based PC Processor: Intel(R) i7-8700H CPU @3.20GHz. The results show that the processing time of the Algorithm 2 is less than 1 ms. As the cloud radio access network has a powerful computational resources, the running time of the proposed algorithm will be further reduced; Hence, the latency of the URLLC will be surely met.

5.5.3 eMBB Traffic Performance

5.5.3.1 Spectral Efficiency

In (5.18), we express the loss function of the eMBB traffic in terms of the SER of the punctured symbols. To measure the efficiency of the proposed strategy and select the optimal K, we evaluate the average eMBB loss in terms of the contaminated eMBB symbols for both the URLLC mapper and the ESRM while varying the size of the search space K (see Fig. 5.8). The results show that the percentages of the contaminated, or lost, eMBB symbols for the ESRM are 18% and 44% as compared to 59% and 93% for the URLLC mapper for BPSK-4QAM and BPSK-16QAM, respectively. This enhancement of the ESRM results from transmitting the punctured eMBB symbols that fall in the same similarity region of the transmitted URLLC symbols, i.e., only the eMBB symbols that have different similarity regions are effectively lost due to puncturing. Moreover, the results show that the ESRM for the

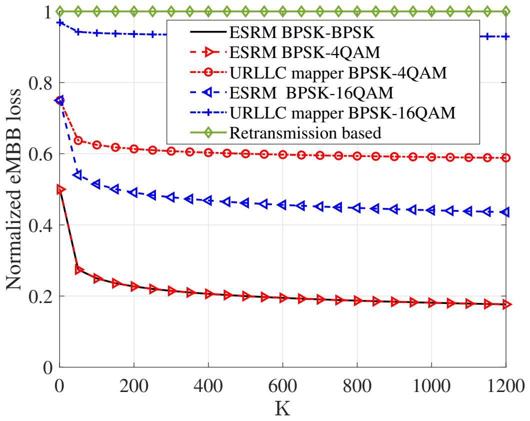


Fig. 5.8: Average eMBB loss relative to the punctured symbols. Adopted $\zeta = 24$ (1 RB) and $\gamma_e = 40$ dB

BPSK-4QAM case achieves the same loss as that of the BPSK-BPSK case. This is attributed to the fact that the probability of similarity in the similarity region is the same, which is equal to 0.5. However, the eMBB SER is enhanced for BPSK-4QAM using ESRM, and this comes at the expense of a deterioration in the URLLC SER by about 2.5 dB (according to (5.14)). The results also demonstrate that the proposed scheme performs better than the code-based puncturing scheme, i.e., $K = 1$ for the URLLC mapper. When K is small, it indicates that a small number of eMBB blocks can be punctured, as per the eMBB QoS requirement. For example, when $K = 300$, the search algorithm scans only 300 eMBB blocks (possibilities) out of $K = 1200$ to allocate the URLLC traffic. In other words, the QoS requirement limits K (number of eMBB blocks the URLLC is compared with). This implies that if the eMBB traffic has stricter QoS requirements, the search space K becomes smaller. Moreover, for the case of the transmission based puncturing, all punctured resources are lost through PI signal. Hence the loss of the retransmission based puncturing is always 100% of the punctured resources.

5.5.3.2 eMBB SER

Fig. 5.9 and 5.10 illustrate the performance of the proposed puncturing scheme in terms of the SER of eMBB. We make the following observations from the results.

- The proposed algorithm achieves better SER for the eMBB traffic compared to the code-based baseline. In other words, the proposed scheme can achieve the

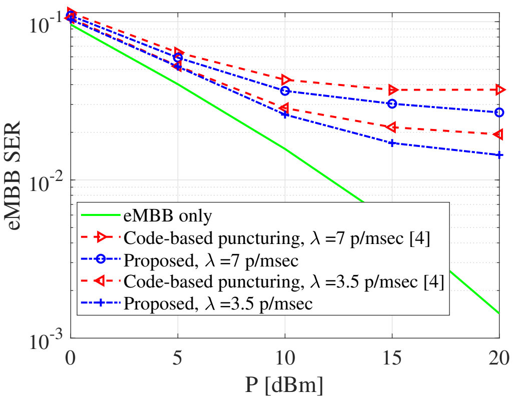


Fig. 5.9: eMBB SER vs transmission power in dBm for different URLLC arrival rate.

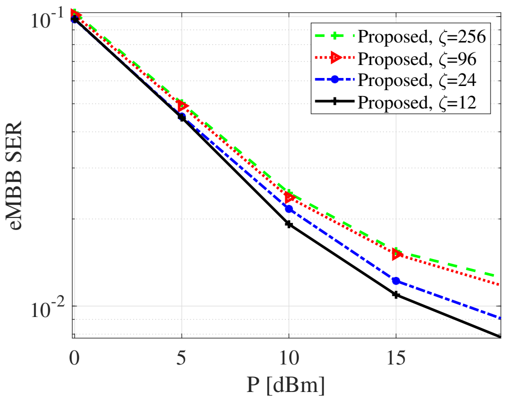
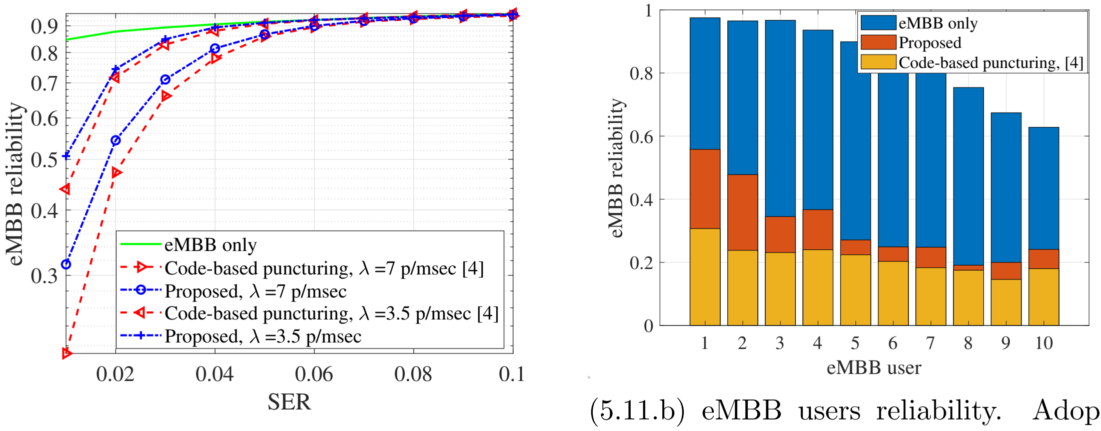


Fig. 5.10: eMBB SER vs transmission power in dBm for different URLLC block (segment) size.

target eMBB SER with lower transmission power. Particularly, the achieved gain increases with the SNR (transmission power) (it reaches 10 dB at high SNR), as shown in Fig. 5.9. Also, the figure shows that the gain of our proposed method is negligible at low transmission power (low SNR) since the channel errors are the dominant here, however the gain improves as the SNR increases since the error at high SNR is dominated by puncturing. Note also that the gain saturates at high SNR according to (5.15), with no further improvement as the BS allocates more transmit power for the eMBB symbols (puncturing errors dominates).

- It is intuitive that the SER of the eMBB traffic deteriorates as the URLLC load increases, as illustrated in Fig. 5.9. For instance, the eMBB SER saturates at 0.02 and 0.04 at both $\lambda = 3.5$ and $\lambda = 7$, respectively. This increase in SER is due to that as the URLLC load increases, the more eMBB resources are punctured, which leads to more SER.

- Fig. 5.10 illustrates the eMBB SER for different URLLC block segment, ζ. For instance, the eMBB SER is enhanced when the URLLC block size ζ becomes smaller, as illustrated in Fig 9b. Reducing ζ increases the probability of similarity, which decreases the effectively punctured symbols. In this case, the trade-off between reducing ζ and the overhead should be optimized to achieve better spectral efficiency. In other words, as ζ decreases, the overhead increases.



(5.11.a) eMBB reliability vs the SER.

(5.11.b) eMBB users reliability. Adopted SER=.01 and $\lambda = 7p/msec$

Fig. 5.11: eMBB traffic reliability of the proposed algorithm. $P = 10\ dBm$

5.5.3.3 Reliability

In this section, we evaluate the reliability of the eMBB traffic as a function of the achieved SER. In general, we define the reliability of both URLLC and eMBB traffic

$$\text{reliability} = \frac{\text{Number of blocks satisfying the targeted SER}}{\text{Total transmitted blocks}}. \tag{5.20}$$

Fig. 5.11.a presents the eMBB reliability while varying the targeted BER, for different URLLC arrival rates. The figure shows that the proposed puncturing strategy achieves better reliability compared the puncturing baseline. For instance, at $P = 0.01$, the proposed puncturing strategy achieves reliability of 31% compared to 20% for the puncturing baseline. This means more eMBB blocks, about 50% enhancement, are received correctly, hence less re-transmissions and better spectral efficiency. The gain of the proposed algorithm decreases while increasing the targeted eMBB SER, this is because the SER becomes dominated by the channel errors at high SER. Moreover, Fig. 5.11.b presents the reliability for each eMBB user. Compared to the baseline, the figure also shows that the proposed algorithm considerably enhances the reliability of the eMBB users, which means less retransmission for the eMBB traffic.

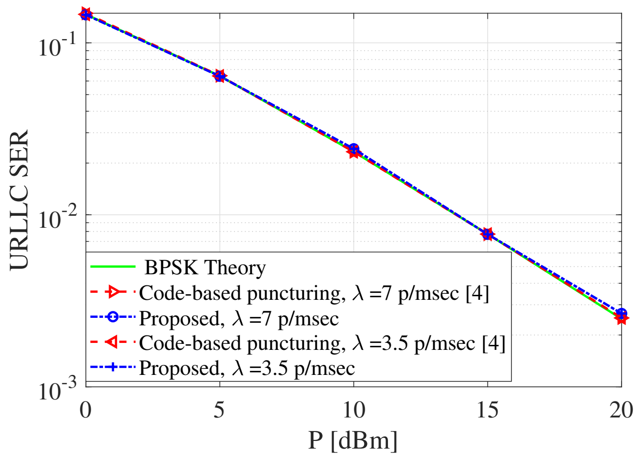


Fig. 5.12: URLLC SER vs transmission power in dBm.

5.5.4 URLLC Performance

In this section, the performance of the URLLC traffic is investigated in terms of the SER and reliability.

5.5.4.1 URLLC SER

Fig. 5.12 illustrates the URLLC SER while varying the transmitted power. As shown in Fig. 5.12, the proposed puncturing scheme preserves the SER of the URLLC traffic while enhancing the SER of the eMBB traffic. The SER loss of the proposed strategy is negligible while taking into account the coding gain. We emphasize here that according to the target SER or BER for both the URLLC and eMBB, the scheduler can select either the URLLC mapper or the ESRM.

5.5.4.2 URLLC Reliability

Fig. 5.13 illustrates the URLLC reliability (success rate) while varying the transmitted power. The figure shows the URLLC reliability for different ϵ_u. The figure shows that the proposed puncturing strategy preserves the URLLC reliability, which makes the proposed strategy a practical method for efficient multiplexing between eMBB and URLLC traffics.

In summary, the proposed puncturing scheme represents a low complexity solution that optimizes between the SER and the spectral efficiency of the eMBB traffic while preserving the reliability of the URLLC services. The proposed algorithm gain, for

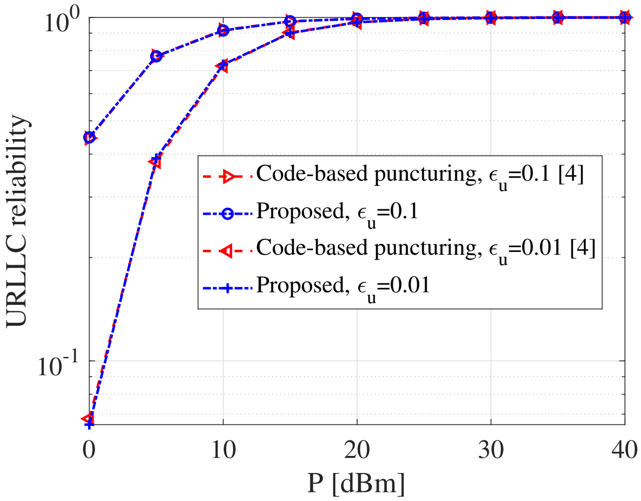


Fig. 5.13: URLLC reliability versus the transmitting power. $\lambda = 7\ p/msec$

the eMBB traffic, starts at 0 dB at low SNR and reach up to 10 dB at high SNR, and it enhances the eMBB reliability up to 50%.

5.6 Summary

In this chapter, we proposed a downlink puncturing strategy in an effort to reduce the impact of transmitting URLLC traffic simultaneously with eMBB traffic. The proposed strategy mitigates the impact on the eMBB traffic by exploiting the region similarity between the eMBB and URLLC symbols to reduce the effectively punctured eMBB symbols. The introduced strategy covers all relations between the eMBB and URLLC modulation schemes. Throughout the analysis, it was shown that the eMBB SER depends on the channel gain, the URLLC load, and the average similarity between the URLLC and eMBB traffic. At high SNR, the eMBB SER asymptotically saturates to the errors due to puncturing, and it is proportional to the ratio between the effectively punctured eMBB symbols to the total eMBB load. Also, when the URLLC block is small or the search space increases, the eMBB SER reduces to the errors due to the channel. Numerical and simulation results demonstrated that the proposed puncturing strategy enhances the system information rate by doubling the URLLC load for the same SER compared to the baseline. While preserving the URLLC quality of service requirements, the proposed puncturing scheme can achieve gains of up to 10 dB as compared to the baseline scheme.

Chapter 6

Conclusions and Future Research Directions

6.1 Conclusions

Through this dissertation, we first provided a brief overview of the envisioned services in the next generation of wireless networks, their enabling technologies and challenges. Chapter 1 discussed the difficulties of scheduling URLLC traffic and corresponding key enabling technologies. Then, the shortcoming in accommodating heterogeneous services as well as research contributions are summarized.

The first contribution of this thesis, which is the content of Chapter 2, focused on accommodating the coexisting eMBB and URLLC traffic through superposition and puncturing. Considering the URLLC reliability and latency requirements alongside the eMBB rate constraint, we formulated the URLLC allocation problem as MINLP, which is generally very hard to solve in polynomial time. Subsequently, the developed optimization was simplified as a one-to-one pairing problem in which the feasibility region and the optimal solutions for the power and spectral resource allocation were derived. Then, the proposed low complexity solution is generalized to support the many-to-many pairing. Using the proposed solution, extensive simulations showed that 30% more URLLC users could be accommodated without degrading their QoS while having minimal impact on the eMBB rate.

The second part of this dissertation, in chapters 3 and 4, focused on integrating

RIS to support URLLC traffic in next-generation wireless networks. Specifically, the joint scheduling of eMBB and URLLC problem aided by RIS was investigated in Chapter 3. This contribution was the first work that explored integrating RIS to achieve the URLLC QoS while preserving the eMBB rate constraint. To pursue this purpose, a time slot eMBB and mini-slot URLLC allocation problems were developed with the objectives of maximizing the eMBB rate and the URLLC admitted packets, respectively. Then, a low complexity framework was proposed to avoid violating the URLLC latency and reliability requirements by proactively optimizing the RIS phase shift. The proactively designed RIS configuration would then be used in the presents of URLLC packets. Simulation results showed that the proposed scheme achieves around 99.99% URLLC packets admission rate, for only 60 passive elements, compared to 95.6% when there is no RIS, while also achieving up to 70% enhancement on the eMBB sum rate.

In Chapter 4, two novel schemes were proposed to reduce the complexity of optimizing the RIS phase shift matrix: a transformation and an element-wise KKT approaches. The former method applied linear transformation on the RIS phase matrix using the optimal configurations of individual users. Accordingly, optimization variables are reduced from the number of passive elements to the number of connected users. On the other hand, the latter scheme is based on closed-form expression derived using the KKT conditions with certain approximations. Extensive simulations were performed where they showed the superior performance of the proposed solutions in terms of optimality and complexity for large-scale RIS, making them practical for URLLC applications.

Finally, Chapter 5 introduced the symbol similarity puncturing scheme as a spectrally efficient mechanism to multiplex the coexisting URLLC and eMBB traffic. Then, we generalized the proposed method for different modulation schemes by introducing the symbol region similarity concept. We depicted the performance of the proposed puncturing mechanism analytically and through simulations. By employing the proposed scheme, simulation results showed that the eMBB spectral efficiency is improved by puncturing fewer symbols leading to better SER and eMBB reliability. Moreover, the URLLC data is accommodated while maintaining its reliability.

6.2 Future Works

Although this thesis addresses several research questions concerning coexisting services, specifically eMBB and URLLC, in the next-generation wireless networks, other challenges remain that we plan to undertake in future works. In the sequel, we list some of the potential directions.

6.2.1 Practical Considerations on the Puncturing/Superposition Framework

In Chapter 5, the main motivation of the proposed puncturing scheme is to reduce the contaminated eMBB symbols leading to protect the effected eMBB user. This chapter showed that, under uncoded traffic, the proposed scheme achieves better eMBB spectral efficiency and SER. However, in practice, error correction codes are used for controlling channel and puncturing errors. It would therefore be worth studying the performance of the proposed puncturing scheme under a coded system in which the eMBB and URLLC streams are coded. The rate and targeted block error rates should be incorporated while allocating the URLLC load. Moreover, next-generation networks are expected to enable massive antennas, where spatial multiplexing and spatial diversity provide significant throughput gains. It would be also necessary to study the performance under the multi-cell scenario where multiple cells cooperate to server the cell edge user (eMBB or URLLC).

In fact, these possible extensions that we plan to pursue are also a possible directions for the superposition and the RIS framework RIS-aided proposed in chapters 2 and 3, respectively. For instance, we plan to generalize the proposed methodologies to the case of multi-cells scenario. As a result, eMBB and URLLC users are divided into cell-center and cell-edge users, with the association problem being handled first. The cell-center users (eMBB and URLLC) will are served by a single BS. In contrast, the cell-edge users (eMBB and URLLC) should be accommodated by multiple adjacent and cooperating cells via the coordinated multi-point (CoMP) technique [52, 53]. To guarantee the reliability of URLLC traffic and the eMBB QoS requirements, the proposed methodologies can be applied for both user clusters while applying an inter-cell interference mitigation mechanism.

6.2.2 Next-Generation Multiple Access

NGMA has to support hundreds of thousands of connected users with limitless wireless capacity, high data rates, high reliability, and ultra-low latency [6, 7, 9, 10]. In this regard, RSMA is suggested as a potential NGMA technique for upcoming wireless networks [13]. In RSMA, as shown in Fig. 6.1, the transmitted messages are split into common and private messages [4, 13]. The common message may contain data from multiple users, so multiple users decode it. On the other hand, private messages are decoded by their corresponding users. Particularly, users rely on SIC to first decode the common messages before obtaining the private messages by only considering the private messages of other users as noise. It has been proved that RSMA outperforms NOMA in terms of spectrum and energy efficiency, even under imperfect channel state information [4, 121]. Moreover, RSMA is expected to post the performance of the URLLC and eMBB service classes. [122]. Consequently, and motivated by the aforementioned benefits of RSMA, it would be worth investigating the feasibility of adopting RSMA in the coexistence problem for downlink and uplink wireless networks. In this research direction, we have considered RSMA in our published paper in [123].

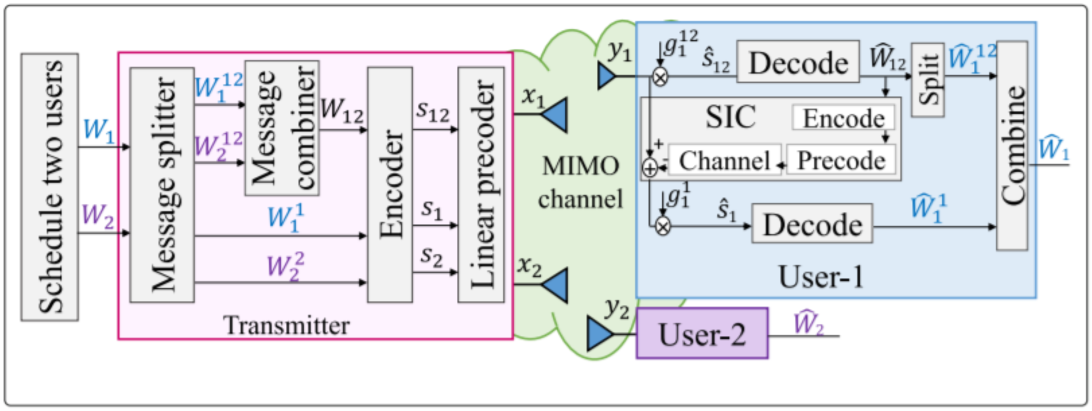


Fig. 6.1: One-layer RSMA for two users [4].

6.2.3 Leveraging Machine Learning Tools

Although two low complexity frameworks were proposed to accommodate the URLLC traffic in RIS-aided wireless networks, integrating RIS in wireless networks comes with its delay sources that may impact the URLLC latency. These latency sources include the delays for the CSI estimation, the phase shift optimization and configuration, which may affect the URLLC latency and reliability and hence the rate requirement of coexisting eMBB traffic. As an attractive solution, machine learning (ML) is capable of solving very complex problems, like the case of RIS-aided wireless networks. For example, deep reinforcement learning (DRL) and distributed machine learning (DML) may be used to effectively deploy the RIS, improve and minimize the complexity of the CSI estimate, and optimize the passive and active elements at the RIS and the BS, respectively. In this direction, we plan to study the chance of building a unified ML model for the BSs-RIS-users association, CSI estimation, passive and active beamforming, and URLLC allocation. Accordingly, instead of performing the CSI-estimation, uses-association and RIS configuration in a cascade way, the BS will intelligent perform the CSI-estimate and RIS configuration for only the associated (subset) users based on the environment measurements (states), i.e., mobility, locations, service class etc...

Bibliography

[1] M. Bennis, M. Debbah, and H. V. Poor. Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale. *Proceedings of the IEEE*, 106(10):1834–1853, Oct. 2018.

[2] Gordon J. Sutton *et al.* Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives. *IEEE Commun. Surveys Tuts.*, 21(3):2488–2524, 3rd quarter 2019.

[3] A. Anand, G. De Veciana, and S. Shakkottai. Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks. In *Proc. IEEE INFOCOM*, Honolulu, HI, USA, Oct. 2018.

[4] Yijie Mao, Bruno Clerckx, and Victor OK Li. Rate-Splitting Multiple Access for Downlink Communication Systems: Bridging, Generalizing, and Outperforming SDMA and NOMA. *EURASIP J. Wireless Commun. Netw.*, 2018(1):1–54, May 2018.

[5] W. Saad, M. Bennis, and M. Chen. A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. *IEEE Network*, 34(3):134–142, May 2020.

[6] F. Tariq *et al.* A Speculative Study on 6G. *IEEE Wireless Commun.*, 27(4):118–125, Aug. 2020.

[7] K. B. Letaief *et al.* The Roadmap to 6G: AI Empowered Wireless Networks. *IEEE Commun. Mag.*, 57(8):84–90, Aug. 2019.

[8] Wei Jiang *et al.* The Road Towards 6G: A Comprehensive Survey. *IEEE Open J. Commun. Soc.*, 2:334–366, Feb. 2021.

[9] Mohamed Amine Arfaoui *et al.* Physical Layer Security for Visible Light Communication Systems: A Survey. *IEEE Commun. Surveys Tuts.*, 22(3):1887–1908, 3rd quarter 2020.

[10] Xinyue Pei *et al.* Next-Generation Multiple Access Based on NOMA With Power Level Modulation. *IEEE J. Sel. Areas Commun.*, 40(4):1072–1083, Apr. 2022.

[11] Zhiguo Ding *et al.* A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE J. Sel. Areas Commun.*, 35(10):2181–2195, Oct. 2017.

[12] SM Riazul Islam, Nurilla Avazov, Octavia A Dobre, and Kyung-Sup Kwak. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Commun. Surveys Tuts.*, 19(2):721–742, 2016.

[13] Onur Dizdar *et al.* Rate-Splitting Multiple Access: A New Frontier for the PHY Layer of 6G. In *6G," in Proc. IEEE 92nd Veh. Technol (VTC2020-Fall)*, Victoria, BC, Canada, Nov. 2020.

[14] P. Popovski *et al.* Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC). *IEEE Trans. Commun.*, 67(8):5783–5801, Aug. 2019.

[15] Yong Niu *et al.* A Survey of Millimeter Wave Communications (mmWave) for 5G: Opportunities and Challenges. *Wireless networks*, 21(8):2657–2676, Nov. 2015.

[16] Nan Chi *et al.* Visible Light Communication in 6G: Advances, Challenges, and Prospects. *IEEE Vehicular Technology Magazine*, 15(4):93–102, Dec. 2020.

[17] Marco Di Renzo *et al.* Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How It Works, State of Research, and The Road Ahead. *IEEE J. Sel. Areas Commun.*, 38(11):2450–2525, Nov. 2020.

[18] Yuanwei Liu *et al.* Reconfigurable Intelligent Surfaces: Principles and Opportunities. *IEEE Communications Surveys Tutorials*, 23(3):1546–1577, May 2021.

[19] E. Basar *et al.* Wireless Communications Through Reconfigurable Intelligent Surfaces. *IEEE Access*, 7:116753–116773, Aug. 2019.

[20] H. Ji *et al.* Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects. *IEEE Wireless Commun.*, 25(3):124–130, June 2018.

[21] He Chen *et al.* Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches. *IEEE Communications Magazine*, 56(12):119–125, Dec. 2018.

[22] Y Rao, J Jing, et al. New Services & Applications with 5G Ultra-Reliable Low Latency Communication. *5G Americas, Bellevue, WA, USA, Tech. Rep*, 2018.

[23] Jihong Park *et al.* Extreme URLLC: Vision, Challenges, and Key Enablers. *arXiv preprint arXiv:2001.09683*, Jan. 2020.

[24] Ali A Zaidi *et al.* Waveform and numerology to support 5G services and requirements. *IEEE Commun. Mag.*, 54(11):90–98, Nov. 2016.

[25] Aunas Manzoor *et al.* Contract-Based Scheduling of URLLC Packets in Incumbent EMBB Traffic. *IEEE Access*, 8:167516–167526, Sep. 2020.

[26] P. Popovski *et al.* 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. *IEEE Access*, 6:55765–55779, Sep. 2018.

[27] Zhengquan Zhang *et al.* 6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies. *IEEE Veh. Technol. Mag.*, 14(3):28–41, Sep. 2019.

[28] D. C. Melgarejo *et al.* Reconfigurable Intelligent Surface-Aided Grant-Free Access for Uplink URLLC. In *Proc. IEEE 6G SUMMIT*, Levi, Finland, Mar. 2020.

[29] Mohammed Almekhlafi *et al.* Enabling URLLC Applications Through Reconfigurable Intelligent Surfaces: Challenges and Potential. *IEEE Internet Things Mag.*, 5(1):130–135, Mar. 2022.

[30] Klaus Pedersen *et al.* Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband. In *Proc. IEEE VTC*, Toronto, ON, Canada, Sep. 2017.

[31] Mohammed Almekhlafi *et al.* Superposition-Based URLLC Traffic Scheduling in 5G and Beyond Wireless Networks. *IEEE Trans. Commun.*, Sept., 2022.

[32] Mohammed Almekhlafi *et al.* Joint Resource and Power Allocation for URLLC-eMBB Traffics Multiplexing in 6G Wireless Networks. In *Proc. IEEE ICC*, June 2021.

[33] R. Kassab *et al.* Non-Orthogonal Multiplexing of Ultra-Reliable and Broadband Services in Fog-Radio Architectures. *IEEE Access*, 7:13035–13049, Jan. 2019.

[34] Z. Li *et al.* 5G URLLC: Design Challenges and System Concepts. In *Proc. IEEE ISWCS*, Oct. 2018.

[35] M. Alsenwi *et al.* eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach. *IEEE Commun. Lett.*, 23(4):740–743, Apr. 2019.

[36] Madyan Alsenwi *et al.* Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach. *IEEE Trans. Wireless Commun.*, 20(7):4585–4600, Feb. 2021.

[37] A. Karimi *et al.* Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G. In *Proc. IEEE VTC*, Kuala Lumpur, Malaysia, Apr./May 2019.

[38] N. B. Khalifa *et al.* Low-Complexity Channel Allocation Scheme for URLLC Traffic. *IEEE Trans. Commun.*, 69(1):194 – 206, Jan. 2021.

[39] Mohammed Y. Abdelsadek, Yasser Gadallah, and Mohamed H. Ahmed. A Critical MTC Resource Allocation Approach for LTE Networks With Finite Blocklength Codes. *IEEE Trans. Veh. Technol.*, 69(5):5598–5609, May 2020.

[40] Anupam Kumar Bairagi *et al.* A matching based coexistence mechanism between eMBB and uRLLC in 5G wireless networks. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 2377–2384, Apr. 2019.

[41] Daniel Maaz, Ana Galindo-Serrano, and Salah Eddine Elayoubi. URLLC User Plane Latency Performance in New Radio. In *Proc. IEEE ICT*, Saint-Malo, France, June 2018.

[42] Y. Saito *et al.* Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access. In *2013 IEEE VTC Conf.*, pages 1–5, June 2013.

[43] Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[44] Y. Long, Y. Gao, and T. Yang. Research on Ultra-Reliable and Low-Latency Wireless Communications in Smart Factory with Finite Block-Length. In *In Proc. IEEE ICCC Workshops*, pages 158–162, Aug. 2018.

[45] Yan Huang *et al.* A Deep-Reinforcement-Learning-Based Approach to Dynamic eMBB/URLLC Multiplexing in 5G NR. *IEEE Internet Things J.*, 7(7):6439–6456, July 2020.

[46] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019.

[47] Jing Cheng, Chao Shen, and Shuqiang Xia. Robust URLLC Packet Scheduling of OFDM Systems. In *Proc. IEEE WCNC*, May 2020.

[48] Mira Morcos *et al.* Optimal resource preemption for aperiodic URLLC traffic in 5G Networks. In *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1–6, Sep. 2020.

[49] Mohammed Almekhlafi *et al.* Joint Resource Allocation and Phase Shift Optimization for RIS-Aided eMBB/URLLC Traffic Multiplexing. *IEEE Trans. Commun.*, 70(2):1304–1319, Feb. 2022.

[50] S. Ebbesen, P. Kiwitz, and L. Guzzella. A generic particle swarm optimization Matlab function. In *2012 American Control Conf. (ACC)*, pages 1519–1524, June 2012.

[51] P. Dinh *et al.* A Low-Complexity Framework for Joint User Pairing and Power Control for Cooperative NOMA in 5G and Beyond Cellular Networks. *IEEE Trans. Commun.*, pages 1–1, Nov. 2020.

[52] Mohamed Elhattab, Mohamed Amine Arfaoui, and Chadi Assi. A joint comp c-noma for enhanced cellular system performance. *IEEE Commun. Lett.*, 24(9):1919–1923, Sep. 2020.

[53] Ralf Irmer *et al.* Coordinated multipoint: Concepts, performance, and field trial results. *IEEE Commun. Mag.*, 49(2):102–111, Feb. 2011.

[54] Hao Yin, Lyutianyang Zhang, and Sumit Roy. Multiplexing URLLC Traffic within eMBB Services in 5G NR: Fair Scheduling. *IEEE Trans. Commun.*, 69(2):1080 – 1093, Feb. 2021.

[55] Mohammed Almekhlafi *et al.* Joint Scheduling of eMBB and URLLC Services in RIS-Aided Downlink Cellular Networks. In *Proc. IEEE ICCCN*, pages 1–9, July 2021.

[56] S. Gong *et al.* Toward Smart Wireless Communications via Intelligent Reflecting Surfaces: A Contemporary Survey. *IEEE Commun. Surv. Tuts.*, 22(4):2283–2314, June 2020.

[57] Moataz Samir *et al.* Optimizing Age of Information Through Aerial Reconfigurable Intelligent Surfaces: A Deep Reinforcement Learning Approach. *IEEE Trans. Veh. Technol.*, 70(4):3978–3983, Mar. 2021.

[58] Q. Wu and R. Zhang. Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming. *IEEE Trans. Wireless Commun.*, 18(11):5394–5409, Aug. 2019.

[59] X. Yu, D. Xu, and R. Schober. MISO Wireless Communication Systems via Intelligent Reflecting Surfaces : (Invited Paper). In *Proc. IEEE ICCC*, Changchun, China, Oct. 2019.

[60] G. Yang, X. Xu, and Y. Liang. Intelligent Reflecting Surface Assisted Non-Orthogonal Multiple Access. In *Proc. IEEE WCNC*, Seoul, Korea (South), May 2020.

[61] Xuelin Cao *et al.* AI-Assisted MAC for Reconfigurable Intelligent-Surface-Aided Wireless Networks: Challenges and Opportunities. *IEEE Commun. Mag.*, 59(6):21–27, July 2021.

[62] Walid R. Ghanem, Vahid Jamali, and Robert Schober. Joint Beamforming and Phase Shift Optimization for Multicell IRS-aided OFDMA-URLLC Systems. In *Proc. IEEE WCNC*, Mar./Apr. 2021.

[63] A. Ranjha and G. Kaddoum. URLLC Facilitated by Mobile UAV Relay and RIS: A Joint Design of Passive Beamforming, Blocklength and UAV Positioning. *IEEE Internet Things J.*, 8(6):4618 – 4627, Mar. 2020.

[64] Xuelin Cao *et al.* Reconfigurable Intelligent Surface-Assisted Aerial-Terrestrial Communications via Multi-Task Learning. *IEEE J. Sel. Areas Commun.*, 39(10):3035–3050, Oct. 2021.

[65] Praveenkumar Korrai *et al.* A RAN Resource Slicing Mechanism for Multiplexing of eMBB and URLLC Services in OFDMA Based 5G Wireless Networks. *IEEE Access*, 8:45674–45688, Mar. 2020.

[66] Yiyu Guo *et al.* Intelligent Reflecting Surface Aided Multiple Access Over Fading Channels. *IEEE Trans. Commun.*, 69(3):2015–2027, Mar. 2021.

[67] Qianqian Zhang, Walid Saad, and Mehdi Bennis. Millimeter Wave Communications with an Intelligent Reflector: Performance Optimization and Distributional Reinforcement Learning. *IEEE Trans. Wireless Commun.*, pages 1–1, Sep. 2021.

[68] Senglee Foo. Liquid-Crystal Reconfigurable Metasurface Reflectors. In *Proc. IEEE ISAP*, San Diego, CA, USA, July 2017. IEEE.

[69] Lei Zhang *et al.* Space-Time-Coding Digital Metasurfaces. *Nature communications*, 9(1):1–11, Oct. 2018.

[70] Wankai Tang *et al.* Programmable Metasurface-Based RF Chain-Free 8PSK Wireless Transmitter. *Electron. Lett.*, 55(7):417–420, Apr. 2019.

[71] Mohamed Elhattab *et al.* Reconfigurable Intelligent Surface Enabled Full-Duplex/Half-Duplex Cooperative Non-Orthogonal Multiple Access. *IEEE Trans. Wireless Commun.*, 21(5):3349–3364, May. 2022.

[72] Jiakuo Zuo, Yuanwei Liu, and Naofal Al-Dhahir. Reconfigurable Intelligent Surface Assisted Cooperative Non-orthogonal Multiple Access Systems. *IEEE Trans. Commun.*, 69(10):6750–6764, July 2021.

[73] X. Liu *et al.* RIS Enhanced Massive Non-Orthogonal Multiple Access Networks: Deployment and Passive Beamforming Design. *IEEE J. Sel. Areas Commun.*, 39(4):1057–1071, Apr. 2021.

[74] Zijian Zhang and Linglong Dai. A Joint Precoding Framework for Wideband Reconfigurable Intelligent Surface-Aided Cell-Free Network. *IEEE Trans. Signal Process.*, 69:4085–4101, June 2021.

[75] Emil Bjornson, Ozgecan Ozdogan, and Erik G. Larsson. Intelligent Reflecting Surface Versus Decode-and-Forward: How Large Surfaces are Needed to Beat Relaying? *IEEE Commun. Lett.*, 9(2):244–248, Oct. 2020.

[76] Mohamed Almekhlafi *et al.* A Low Complexity Passive Beamforming Design for Reconfigurable Intelligent Surface (RIS) in 6G Networks. *IEEE Trans. Veh. Technol.*, 2022.

[77] Sarah Basharat *et al.* Reconfigurable Intelligent Surfaces: Potentials, Applications, and Challenges for 6G Wireless Networks. *IEEE Wireless Commun.*, 28(6):184–191, 2021.

[78] Mohamed ElMossallamy *et al.* Reconfigurable Intelligent Surfaces for Wireless Communications: Principles, Challenges, and Opportunities. *IEEE Trans. Cogn. Commun. and Netw.*, 6(3):990–1002, Sep. 2020.

[79] Sixian Li *et al.* Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming. *IEEE Wireless Commun. Lett.*, 9(5):716–720, May 2020.

[80] Sixian Li *et al.* Robust secure UAV communications with the aid of reconfigurable intelligent surfaces. *IEEE Trans. Wireless Commun.*, Oct. 2021.

[81] Tianwei Hou *et al.* Reconfigurable intelligent surface aided NOMA networks. *IEEE J. Sel. Areas Commun.*, 38(11):2575–2588, Nov. 2020.

[82] Chao Zhang *et al.* Downlink analysis for reconfigurable intelligent surfaces aided NOMA networks. In *Proc. IEEE GLOBECOM*, Taipei, Taiwan, Dec. 2020. IEEE.

[83] Vetrivel Chelian Thirumavalavan and Thiruvengadam S Jayaraman. BER analysis of reconfigurable intelligent surface assisted downlink power domain NOMA system. In *Proc. Int. Conf. Commun. Syst. Netw. (COMSNETS)*. IEEE, Jan. 2020.

[84] Zhaohui Yang *et al.* Energy Efficient Rate Splitting Multiple Access (RSMA) with Reconfigurable Intelligent Surface. In *Proc. IEEE ICC*, Dublin, Ireland, June 2020.

[85] Tianyu Fang, Yijie Mao, Shanpu Shen, Zhencai Zhu, and Bruno Clerckx. Fully Connected Reconfigurable Intelligent Surface Aided Rate-Splitting Multiple Access for Multi-User Multi-Antenna Transmission. In *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, Seoul, Korea, May 2022.

[86] Ramin Hashemi *et al.* Deep Reinforcement Learning for Practical Phase Shift Optimization in RIS-aided MISO URLLC Systems. *IEEE Internet Things J.*, 2022 (Early Access).

[87] Shivani Dhok *et al.* Non-Linear Energy Harvesting in RIS-Assisted URLLC Networks for Industry Automation. *IEEE Trans. Commun.*, 69(11):7761–7774, Nov. 2021.

[88] Yiqing Li *et al.* Joint beamforming design in multi-cluster miso noma reconfigurable intelligent surface-aided downlink communication networks. *IEEE Trans. Commun.*, 69(1):664–674, Jan. 2021.

[89] Haseeb Ur Rehman *et al.* Joint Active and Passive Beamforming Design for IRS-Assisted Multi-User MIMO Systems: A VAMP-Based Approach. *IEEE Trans Commun.*, pages 1–1, July 2021.

[90] Huayan Guo *et al.* Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks. *IEEE Trans. Wireless Commun.*, 19(5):3064–3076, Feb. 2020.

[91] Min Fu, Yong Zhou, and Yuanming Shi. Intelligent Reflecting Surface for Downlink Non-Orthogonal Multiple Access Networks. In *Proc. IEEE Globecom*, Waikoloa, HI, USA, Dec. 2019.

[92] Keming Feng *et al.* Deep Reinforcement Learning Based Intelligent Reflecting Surface Optimization for MISO Communication Systems. *IEEE Wireless Communications Letters*, 9(5):745–749, May 2020.

[93] Gilsoo Lee *et al.* Deep Reinforcement Learning for Energy-Efficient Networking with Reconfigurable Intelligent Surfaces. In *Proc. IEEE ICC*, June 2020.

[94] Chao Feng *et al.* Wireless Communication with Extremely Large-Scale Intelligent Reflecting Surface. In *Proc. IEEE/CIC ICCC*, Xiamen, China, July 2021.

[95] Emil Björnson and Luca Sanguinetti. Demystifying the Power Scaling Law of Intelligent Reflecting Surfaces and Metasurfaces. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 549–553, 2019.

[96] Razan Abdulhammed *et al.* Features dimensionality reduction approaches for machine learning based network intrusion detection. *Electronics*, 8(3):322, Feb. 2019.

[97] Adilet Otemissov. *Dimensionality reduction techniques for global optimization.* PhD thesis, University of Oxford, 2020.

[98] Coralia Cartis and Adilet Otemissov. A dimensionality reduction technique for unconstrained global optimization of functions with low effective dimensionality. *Information and Inference: A Journal of the IMA*, 11(1):167–201, Mar. 2022.

[99] Ami Wiesel, Yonina C. Eldar, and Shlomo Shamai. Zero-Forcing Precoding and Generalized Inverses. *IEEE Trans Signal Process.*, 56(9):4409–4418, Sep. 2008.

[100] Mohamed Elhattab *et al.* RIS-Assisted Joint Transmission in a Two-Cell Downlink NOMA Cellular System. *IEEE J. Sel. Areas Commun.*, 40(4):1270–1286, Apr. 2022.

[101] Mohammed Almekhlafi *et al.* A Downlink Puncturing Scheme for Simultaneous Transmission of URLLC and eMBB Traffic by Exploiting Data Similarity. *IEEE Trans. Veh. Technol.*, 70(12):1–1, Dec. 2021.

[102] 3GPP. TSG RAN WG1 Meeting 87. Technical report, 3rd Generation Partnership Project (3GPP), Nov. 2016.

[103] C. Xiao *et al.* Downlink MIMO-NOMA for Ultra-Reliable Low-Latency Communications. *IEEE J. Sel. Areas Commun.*, 37(4):780–794, Apr. 2019.

[104] E. J. dos Santos *et al.* Network Slicing for URLLC and eMBB with Max-Matching Diversity Channel Allocation. *IEEE Commun. Lett.*, 24(3):658 – 661, Dec. 2019.

[105] A. Azari, M. Ozger, and C. Cavdar. Risk-Aware Resource Allocation for URLLC: Challenges and Strategies with Machine Learning. *IEEE Commun Mag.*, 57(3):42–48, March 2019.

[106] M. Chraiti, A. Ghrayeb, and C. Assi. A NOMA Scheme Exploiting Partial Similarity Among Users Bit Sequences. *IEEE Trans. Commun.*, 66(10):4923–4935, Oct. 2018.

[107] Hua and Sun *et al.* Five decades of hierarchical modulation and its benefits in relay-aided networking. *IEEE Access*, 3:2891–2921, Dec. 2015.

[108] A. R. Calderbank and N. Seshadri. Multilevel codes for unequal error protection. *IEEE Trans. Inf. Theory*, 39(4):1234–1248, July 1993.

[109] Md J Hossain, Pavan K Vitthaladevuni, M-S Alouini, Vijay K Bhargava, and Andrea J Goldsmith. Adaptive hierarchical modulation for simultaneous voice and multiclass data transmission over fading channels. *IEEE Trans. Veh. Technol.*, 55(4):1181–1194, July 2006.

[110] Salvatore D'Oro, Francesco Restuccia, and Tommaso Melodia. Hiding Data in Plain Sight: Undetectable Wireless Communications Through Pseudo-Noise Asymmetric Shift Keying. In *Proc. IEEE INFOCOM*, Paris, France, Apr./May 2019.

[111] W. Wu, P. Lin, and Y. Lee. An Indicator-Free eMBB and URLLC Multiplexed Scheme for 5G Downlink System. In *Proc. IEEE VTC*, Honolulu, HI, USA, Sep. 2019.

[112] Aarti Sharma and Mohammad Salim. Polar Code: The channel Code Contender for 5G Scenarios. In *Proc. IEEE Comptelix*, Jaipur, India, July 2017.

[113] Jaya Rao and Sophie Vrzic. Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance analysis. *IEEE Network*, 32(2):32–40, Apr. 2018.

[114] Nils Strodthoff *et al.* Enhanced Machine Learning Techniques for Early HARQ Feedback Prediction in 5G. *IEEE J. Sel. Areas Commun.*, 37(11):2573–2587, Aug. 2019.

[115] Chih-Hsiu Zeng and Kwang-Cheng Chen. Downlink Multiuser Detection in the Virtual Cell-Based Ultra-Low Latency Vehicular Networks. *IEEE Trans. Veh. Technol.*, 68(5):4651–4666, Feb. 2019.

[116] Seda Doğan, Armed Tusha, and Hüseyin Arslan. NOMA With Index Modulation for Uplink URLLC Through Grant-Free Access. *IEEE J. Sel. Topics Signal Process.*, 13(6):1249–1257, May. 2019.

[117] Xiaoning Wu *et al.* Low-Rate PBRL-LDPC Codes for URLLC in 5G. *IEEE Wireless Commun. Lett.*, 7(5):800–803, Apr. 2018.

[118] Amitabha Ghosh. 5G New Radio (NR): Physical Layer Overview and Performance. In *Proc. IEEE CTW*, Lisbon, Portugal, Oct. 2018.

[119] H. Soury, F. Yilmaz, and M. Alouini. Exact Symbol Error Probability of Square M-QAM Signaling Over Generalized Fading Channels Subject to Additive Generalized Gaussian Noise. In *Proc. IEEE ISIT*, Istanbul, Turkey, July 2013.

[120] Andrea Goldsmith. *Wireless Communications.* Cambridge university press, 2005.

[121] Bruno Clerckx *et al.* Rate-Splitting Unifying SDMA, OMA, NOMA, and Multicasting in MISO Broadcast Channel: A Simple Two-User Rate Analysis. *IEEE Wireless Commun. Lett.*, 9(3):349–353, Mar. 2020.

[122] Onur Dizdar *et al.* Rate-Splitting Multiple Access for Enhanced URLLC and eMBB in 6G: Invited Paper. In *Proc. IEEE ISWCS*, 2021.

[123] Shreya Khisa *et al.* Full Duplex Cooperative Rate Splitting Multiple Access for a MISO Broadcast Channel with two Users. *IEEE Commun. Lett.*, 26:1913–1917, Aug, 2022.

[124] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[125] Barry C Arnold, Narayanaswamy Balakrishnan, and Haikady Navada Nagaraja. *A First Course in Order Statistics*, volume 54. Siam, 1992.

Appendix A

Proofs and Derivations of Chapter 3

A.1 Solution Approach of Problem 3.19

Let us start by defining $\boldsymbol{\vartheta}^H = [\vartheta_1, \vartheta_2, \ldots, \vartheta_N]^H$, where $\vartheta_n = e^{j\phi_n}$. Therefore, we obtain $|h_{\mathrm{BS},e} + \mathbf{h}^H_{\mathrm{RIS},e} \boldsymbol{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}|^2 = |h_{\mathrm{BS},e} + \boldsymbol{\vartheta}^H \boldsymbol{\Theta}|^2 = \boldsymbol{\vartheta}^H \boldsymbol{\Theta}\, \boldsymbol{\Theta}^H \boldsymbol{\vartheta} + h_{\mathrm{BS},e}\, \boldsymbol{\Theta}^H \boldsymbol{\vartheta} + \boldsymbol{\vartheta}^H \boldsymbol{\Theta}\, h^{\dagger}_{\mathrm{BS},e} + |h_{\mathrm{BS},e}|^2$, where $\boldsymbol{\Theta} = \mathrm{diag}(\mathbf{h}^H_{\mathrm{RIS},e})\, \mathbf{f}_{\mathrm{BS,RIS}}$, and $\boldsymbol{\Theta} \in \mathbb{C}^{N\times 1}$. By introducing an auxiliary variable ρ, problem $\mathcal{P}_{1,2}$ can be equivalently transformed to

$$\mathcal{P}_{1,3} : \max_{\bar{\boldsymbol{\vartheta}}} \; \sum_{e=1}^{E} \log_2 \left(1 + \frac{p_e\, (\bar{\boldsymbol{\vartheta}}^H\, \boldsymbol{Q}_e\, \bar{\boldsymbol{\vartheta}} + |h_{\mathrm{BS},e}|^2)}{\Gamma_{\mathrm{eMBB}}\, \sigma^2} \right) \tag{A.1a}$$

$$\text{s.t. } (1-\delta) \log_2 \left(1 + \frac{p_e\, (\bar{\boldsymbol{\vartheta}}^H\, \boldsymbol{Q}_e\, \bar{\boldsymbol{\vartheta}} + |h_{\mathrm{BS},e}|^2)}{\Gamma_{\mathrm{eMBB}}\, \sigma^2} \right) \geq \frac{r_{\mathrm{th}}}{W\, b}, \quad \forall\, e \in \mathcal{E}_f, \tag{A.1b}$$

$$|\bar{\boldsymbol{\vartheta}}_n| = 1, \qquad \forall n = 1, \ldots N+1, \tag{A.1c}$$

where

$$\boldsymbol{Q}_e = \begin{bmatrix} \boldsymbol{\Theta}\, \boldsymbol{\Theta}^H & \boldsymbol{\Theta}\, h^{\dagger}_{\mathrm{BS},e} \\ h_{\mathrm{BS},e}\, \boldsymbol{\Theta}^H & 0 \end{bmatrix}, \quad \text{and} \quad \bar{\boldsymbol{\vartheta}} = \begin{bmatrix} \boldsymbol{\vartheta} \\ \rho \end{bmatrix}, \tag{A.2}$$

such that $\bar{\boldsymbol{\vartheta}}^H \boldsymbol{Q}_e \bar{\boldsymbol{\vartheta}} = \mathrm{tr}(\boldsymbol{Q}_e \bar{\boldsymbol{\vartheta}} \bar{\boldsymbol{\vartheta}}^H)$. In addition, we define $\boldsymbol{S} = \bar{\boldsymbol{\vartheta}}\bar{\boldsymbol{\vartheta}}^H$, which needs to satisfy $\mathrm{rank}(\bar{\boldsymbol{\vartheta}}) = 1$. This rank one constraint is a non-convex constraint [71]. By dropping this constraint, we reach

$$\mathcal{P}_{1,4} : \max_{\boldsymbol{S}} \sum_{e=1}^{E_f} \log_2 \left(1 + \frac{p_e \left(\mathrm{tr}(\boldsymbol{Q}_e \boldsymbol{S}) + |h_{\mathrm{BS},e}|^2\right)}{\Gamma_{\mathrm{eMBB}}\, \sigma^2}\right) \tag{A.3a}$$

$$\text{s.t. } (1-\delta)\, b \log_2 \left(1 + \frac{p_e \left(\mathrm{tr}(\boldsymbol{Q}_e \boldsymbol{S}) + |h_{\mathrm{BS},e}|^2\right)}{\Gamma_{\mathrm{eMBB}}\, \sigma^2}\right) \geq r_{\mathrm{th}}, \quad \forall\, e \in \mathcal{E}_f, \tag{A.3b}$$

$$\boldsymbol{S}_{n,n} = 1, \qquad \forall n \in \{1, 2, \ldots, N+1\}, \tag{A.3c}$$

$$\boldsymbol{S} \succeq 0. \tag{A.3d}$$

It can be easily seen that problem $\mathcal{P}_{1,4}$ is a semi-definite programming (SDP) problem, which can be optimally solved using one of the convex optimization solvers such as CVX [124]. In general, the optimal $\bar{\boldsymbol{\vartheta}}$ obtained by solving problem $\mathcal{P}_{1,4}$ does not satisfy the rank-one constraint [71]. Consequently, the Gaussian randomization technique is applied to get a rank-one solution [71].

A.2 Solution Approach of Problems 3.20 and 3.21

We provide here the solution for problems in (3.20) and (3.21). Problems in (3.20) and (3.21) (which aim at maximizing the URLLC channels and URLLC-eMBB channels respectively) have similar formulation. We start by solving (3.20). By adding a auxiliary variable ζ, (3.20) is re-written as

$$\mathcal{P}_{3,1} : \quad \max_{\zeta, \boldsymbol{\Phi}_u} \zeta \tag{A.4a}$$

$$\text{s.t.} \quad |g_{\mathrm{BS},u} + \boldsymbol{g}_{\mathrm{RIS},u}^H \boldsymbol{\Phi}_u \, \mathbf{f}_{\mathrm{BS,RIS}}|^2 \geq \zeta \ \forall u \in \mathcal{U}, \tag{A.4b}$$

$$0 \leq \phi_n \leq 2\pi, \quad \forall n = 1, \ldots, N+1 \tag{A.4c}$$

$$\zeta \geq 0. \tag{A.4d}$$

Then, using the same results in (A.3), problem $\mathcal{P}_{3,1}$ is re-written as

$$\begin{aligned}
\mathcal{P}_{3,2}: \quad & \max_{\zeta} \zeta && \text{(A.5a)}\\
\text{s.t.} \quad & \mathrm{tr}(V_u S) + |g_{\mathrm{BS},u}|^2 \geq \zeta, \ \forall u \in \mathcal{U}, && \text{(A.5b)}\\
& S_{n,n} = 1, \ \forall n = 1, \ldots, N+1, && \text{(A.5c)}\\
& S \succeq 0, && \text{(A.5d)}\\
& \zeta \geq 0. && \text{(A.5e)}
\end{aligned}$$

Finally, problem (3.21) can be solved by following the same steps as in problem (3.20).

Appendix B

Proofs and Derivations of Chapter 4

B.1 Proof of Theorem 4.1

In this part, we provide the proof of **Theorem 4.1**. We defined the Lagrangian of problem in (4.14) as

$$\begin{aligned} L(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{u=1}^{U} \frac{\sigma^2 \gamma_u^{\mathrm{th}}}{|h_{\mathrm{BS},1} + \mathbf{h}_{\mathrm{RIS},1}^{H} \boldsymbol{\Phi}\, \mathbf{f}_{\mathrm{BS,RIS}}|^2} \\ - \sum_{n=1}^{N} \alpha_n \times \phi^n + \sum_{n=1}^{N} \beta_n \times (\phi^n - 2\pi). \end{aligned} \tag{B.1}$$

For simplicity, we can rewrite as

$$\begin{aligned} L(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{u=1}^{U} \frac{\sigma^2 \gamma_u^{\mathrm{th}}}{|h_{\mathrm{BS},u} + \sum_{n=1}^{N} h_{c,u}^{n}\, e^{j\phi^n}|^2} \\ - \sum_{n=1}^{N} \alpha_n \times \phi^n + \sum_{n=1}^{N} \beta_n \times (\phi^n - 2\pi). \end{aligned} \tag{B.2}$$

Consequently, the corresponding KKT conditions of (B.2) are:

$$
\begin{aligned}
0 = \sum_{u=1}^{U} & \frac{\sigma^2 \gamma_u^{\text{th}}}{|h_{\text{BS},u} + \sum_{n=1}^{N} h_{\text{c},u}^{n}\, e^{j\phi^n}|^4} \times, \\
& \qquad \frac{\partial(|h_{\text{BS},1} + \sum_{n=1}^{N} h_{\text{c},u}^{n}\, e^{j\phi^n}|^2)}{\partial \phi^n}, \forall n \in \mathcal{N}, && \text{(B.3a)} \\
& \alpha_n \times \phi^n = 0, \forall n \in \mathcal{N}, && \text{(B.3b)} \\
& \beta_n \times (\phi^n - 2\pi) = 0, \forall n \in \mathcal{N}, && \text{(B.3c)} \\
& \phi^n \geq 0, \forall n \in \mathcal{N}, && \text{(B.3d)} \\
& \phi^n \leq 2\pi, \forall n \in \mathcal{N}. && \text{(B.3e)}
\end{aligned}
$$

From (B.3), one can easily obtain that $\alpha_n = 0$ and $\beta_n = 0$. By applying the gradient operation, problem (B.3) can be written as:

$$
0 = \sum_{u=1}^{U} \frac{2\,\sigma^2\,\gamma_u^{\text{th}}\,|h_{d,u}^{n}|\,|h_{\text{c,u}}^{n}|\,\sin(\theta_{d,u}^{n} - \theta_{\text{c},u}^{n} - \phi^n)}{|h_{\text{BS},u} + \sum_{k=1}^{N} h_{\text{c},u}^{k}\, e^{j\phi^k}|^4}, \tag{B.4}
$$

where

$$
\begin{aligned}
|h_{d,u}^{n} + h_{\text{c},u}^{n}\, e^{j\phi^n}|^2 = |h_{d,u}^{n}|^2 + |h_{\text{c},u}^{n}|^2 + \\
2\,|h_{d,u}^{n}|\,|h_{\text{c},u}^{n}|\,\cos(\theta_{d,u}^{n} - \theta_{\text{c},u}^{n} - \phi^n) \forall n \in \mathcal{N},
\end{aligned} \tag{B.5}
$$

and

$$
h_{d,u}^{n} = h_{\text{BS},u} + \sum_{k \neq n}^{N} h_{\text{c},u}^{k}\, e^{j\phi^k}. \tag{B.6}
$$

To remove the coupling between phase shifts of passive elements, the set of equations in (B.4) can be solved iteratively by assuming that the $h_{\text{BS},u} + \sum_{k \in \mathcal{N}, k \neq n} h_{\text{c},u}^{k}\, e^{j\phi^k}$ is constant for all $n \in \mathcal{N}$. Moreover, at N is large or E is large, the term $|h_{\text{BS},u} + \sum_{n=1}^{N} h_{\text{c},u}^{n}\, e^{j\phi^n}|^4$ can be approximated as $|h_{d,u}^{n} + h_{\text{c},u}^{n}\, e^{j\phi^n}|^4 = |h_{d,u}^{n}|^4$, i.e., $|h_{\text{c},u}^{n}| \ll |h_{d,u}^{n}|$. Then, equation (B.4) can be approximated as:

$$
0 \approx \sum_{u=1}^{U} \frac{2\,\sigma^2\,\gamma_u^{\text{th}}\,|h_{\text{c,u}}^{n}|\,\sin(\theta_{d,u}^{n} - \theta_{\text{c},u}^{n} - \phi^n)}{|h_{d,u}^{n}|^3}. \tag{B.7}
$$

By applying the subtraction theorem $\sin(a-b) = \sin(a)\cos b - \cos a \sin(b)$, equation (B.7) can be expressed as

$$0 \approx \sum_{u=1}^{U} C_u^n \cos(\phi^n) - \sum_{u=1}^{U} D_u^n \sin(\phi^n), \tag{B.8}$$

where

$$C_u^n = \frac{2\,\sigma^2\,\gamma_u^{\text{th}}\,|h_{\text{c,u}}^n|\,\sin(\theta_{d,u}^n - \theta_{\text{c},u}^n)}{|h_{d,u}^n|^3}, \tag{B.9a}$$

$$D_u^n = \frac{2\,\sigma^2\,\gamma_u^{\text{th}}\,|h_{\text{c,u}}^n|\,\cos(\theta_{d,u}^n - \theta_{\text{c},u}^n)}{|h_{d,u}^n|^3}. \tag{B.9b}$$

Accordingly, equation (B.8) can be written as $0 \approx \sin(\theta_{\text{effective}}^n - \phi^n)$, where $\theta_{\text{effective}}^n$ is effective direction (angle) of the combined direct and cascaded channels of all users, which can be approximately obtained by

$$\theta_{\text{effective}}^n \approx \arctan\left(\frac{\sum_{u=1}^{U} C_u^n}{\sum_{u=1}^{U} D_u^n}\right), \; \theta_{\text{effective}}^n \in \{0, 2\pi\}. \tag{B.10}$$

Similar to the single-user case, the optimal phase shift configuration, ϕ^{n*}, of passive element n, satisfies $\phi^{n*} = \theta_{\text{effective}}^n$. This completes the proof.

Appendix C

Proofs and Derivations of Chapter 5

C.1 Proof of Lemma 5.1

Consider the Binomial distribution with $B\left(\zeta, \eta_{n,m}\right)$ to exploit the similar $\zeta-$symbols blocks between the URLLC load and the eMBB sequence. Then, the CDF $F(k)$ is expressed as

$$F(k) = \sum_{j=0}^{k} \left(\begin{array}{c} \zeta \\ j \end{array} \right) \left(\eta_{n,m}\right)^{j} \left(1 - \eta_{n,m}\right)^{\zeta - j}. \tag{C.1}$$

The expected number of similar symbols between both the URLLC and eMBB blocks is $\mu = \eta_{n,m}\,\zeta$. Under the assumption that the eMBB blocks and the URLLC packet are *i.i.d*, and by searching over the search space ($K_m = L_m - \zeta + 1$), the order statistic after arranging the random samples in an increasing order is $Y_1 \leq Y_2 \leq \cdots \leq Y_{K_m}$. Based on the results of [125], the pmf of Y_z becomes for all $k = 0, 1, \ldots, \zeta$

$$\begin{aligned} f_z(k) = \sum\nolimits_{r=z}^{K^m} \left(\begin{array}{c} K^m \\ r \end{array} \right) \big[& \{F(k)\}^r \{1 - F(k)\}^{K_m - r} \\ & - \{F(k-1)\}^r \{1 - F(k-1)\}^{K_m - r} \big]. \end{aligned} \tag{C.2}$$

Considering the case when the largest ordered sample has at least k similar symbols. Also, considering that $L_m >> l_{n,m}$, the expected number of similar symbols can be approximated as [125]

$$U_{n,m,\zeta} = \sum\nolimits_{k=0}^{\zeta-1} \left[1 - \{F(k)\}^{L_m - \zeta}\right]. \tag{C.3}$$

Averaging over the number of ζ-blocks in $l_{n,m}$, we arrive at

$$U_{n,m,\zeta}(l_{n,m}) = \frac{1}{\lceil \frac{l_{n,m}}{\zeta} \rceil} \sum_{1}^{\lceil \frac{l_{n,m}}{\zeta} \rceil} \sum\nolimits_{k=0}^{\zeta-1} \left[1 - \{F(k)\}^{L_m - \zeta}\right]. \tag{C.4}$$

This completes the proof.