

Resource Allocation for Reconfigurable Intelligent Surface assisted Wireless Communications

by

Mahmoud Saad Abouamer

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2025

© Mahmoud Saad Abouamer 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Dr. Ekram Hossain
Professor, Dept. of Electrical and Computer Engineering,
University of Manitoba, Canada

Supervisor(s): Dr. Patrick Mitran
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal Members: Dr. Ravi Mazumdar
Professor Emeritus, Dept. of Electrical and Computer Engineering,
University of Waterloo

Dr. Guang Gong
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal-External Member: Dr. Ricardo Fukasawa
Professor, Dept. of Combinatorics and Optimization,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Reconfigurable intelligent surfaces (RIS) offer a promising solution to meet the growing demand for connectivity while ensuring stringent quality-of-service (QoS) requirements. An RIS consists of a 2D array of nearly passive reflecting elements, which can be configured to dynamically shape incident electromagnetic waves, thereby engineering the channel between a transmitter and a receiver. An RIS, which typically lacks active components (e.g., RF chains), provides a cost-effective, nearly passive solution for smart radio environments. However, its inability to perform advanced digital signal processing presents challenges, requiring channel state information (CSI) acquisition and RIS configuration optimization to be handled at the transmitter or receiver side, with control signals fed back to the RIS. Thus, efficient channel estimation and optimization of RIS reflection coefficients are essential for practical integration into existing communications systems. Subsequently, this research focuses on addressing these challenges by investigating resource allocation problems in RIS-assisted systems and developing efficient RIS configuration schemes for various practical scenarios.

First, we focus on RIS design in multi-user frequency-division-duplexing (FDD) and time-division-duplexing (TDD) systems, proposing a joint uplink-downlink RIS design where the same configuration supports both transmissions. In FDD, this is essential as uplink (UL) and downlink (DL) occur simultaneously, requiring a joint configuration. In TDD, while not strictly necessary, a joint design reduces feedback overhead, power consumption, and configuration periods associated with updating the RIS. To compute the trade-off between uplink and downlink rates achieved by a joint design, a weighted-sum problem is formulated and optimized using a developed block-coordinate descent (BCD) algorithm. The resulting uplink-downlink trade-off regions are investigated by numerical simulation to gain insights into different scenarios. For many considered scenarios, the proposed joint design is shown to bring significant improvements over the fixed-uplink (fixed-downlink) heuristic of using the RIS configuration optimized for uplink(downlink) to assist downlink (uplink) transmissions. Moreover, the proposed joint design substantially bridges the gap to the individual design upper bound of allowing different RIS configurations in uplink and downlink.

In the second part, we develop a learning-based framework that directly exploits noisy pilots to optimize a multi-user RIS system while accommodating different service priorities and fairness via user weights. Towards this goal, an adaptive beamforming configuration problem is formulated to generate the RIS system's beamforming configurations that optimize the weighted sum-rate (WSR). Under mild regularity conditions, this problem is shown to attain a maximum. To learn approximate solutions, a novel hypernetwork-based

beamforming (HNB) framework is proposed. Particularly, a beamforming network (BFN) exploits available information, including noisy pilots, to generate optimized beamforming configurations. Rather than learning one BFN, a hypernetwork is trained to dynamically generate BFN learning parameters from an input conditioning vector. When the conditioning vector is chosen as the user weights, the trained HNB can tune the BFN to the user weights without the need for retraining. Numerical experiments demonstrate that tuning allows the proposed HNB to perform close to a block-coordinate descent with perfect CSI benchmark and significantly outperform static learning where a BFN is directly trained to optimize beamforming configurations. Additionally, employing the HNB to also tune the BFN to location information considerably reduces the pilots needed to generate optimized beamforming configurations.

In the third part, to enable RIS implementation with diverse technologies, we develop RIS configuration schemes under relaxed hardware constraints. Particularly, we consider the case where an RIS is configured with discrete phase shifts and the amplitude response associated with an RIS element is non-linearly dependent on the phase shift introduced by the element. This RIS configuration problem is formulated as a constrained virtual-channel selection problem. Subsequently, it is demonstrated that the configuration problem, when restricted to subsets of the virtual channels where the maximum phase-variation is $\theta_{\max} < \pi$, can be bounded between monotone upper bound (UB) and lower bound (LB) matroid-constrained problems. We illustrate discrete RIS configuration problems whose optimal solution satisfy this maximum phase-variation property. Moreover, under the maximum phase-variation property, performance guarantees are obtained using a low-complexity configuration scheme. Motivated by this, a practical low-complexity configuration framework is proposed to optimize general discrete RIS problems. Numerical experiments demonstrate the efficacy of the proposed RIS configuration scheme.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Patrick Mitran, for his unwavering support, guidance, and mentorship. His ability to provide the perfect balance between guidance and independence has been invaluable to my growth as a researcher. His mentorship has not only shaped my technical skills but also contributed to my personal growth. I am immensely grateful for all the knowledge he has shared with me, his patience throughout this journey, and his continuous encouragement.

I would also like to extend my sincere thanks to my committee members, Prof. Ravi Mazumdar, Prof. Guang Gong, Prof. Ricardo Fukasawa, and Prof. Ekram Hossain, for their insightful feedback and valuable contributions to my academic progress. Special thanks to Prof. Mazumdar, who taught me my first course at Waterloo and continues to offer invaluable guidance on both technical and broader professional matters.

A special acknowledgment goes to my undergraduate mentors, Prof. Ugur Abdulla, and Prof. Brian Lail whose encouragement and support motivated me to pursue graduate studies. Their belief in my potential played a crucial role in bringing me to this stage.

To my parents, Mohamed and Nadia, I owe everything. Their unconditional love, unwavering support, and relentless efforts—sometimes beyond what anyone could imagine—have shaped my path in life. They have never missed a chance to help me, push me forward, and ensure that I always had the best opportunities. No words can truly capture my gratitude for them.

To my sister, Esraa, my best friend and first teacher, thank you for being the person who has taught me nearly everything I know. Your kindness and support have been my guiding light. I am also grateful to Esraa and her husband, Abdalla Hussein, for always providing me with a home away from home, and for their endless love and care.

Finally, I want to thank my lifelong friends, Mohamed Fathy Hassan, Abdalla Hussein, and Ali Assem Mahmoud, for making this journey more enjoyable and bearable. Their friendship has been a source of joy and strength, and I am truly lucky to have them by my side.

Dedication

In the name of God, the Most Gracious, the Most Merciful. Grant me the power and ability that I may be grateful for Your favors which You have bestowed on me and on my parents, and that I may do righteous good deeds.

This is dedicated to my parents, Nadia and Mohamed and to my sister, Esraa.

Table of Contents

Examining Committee	ii
Author's Declaration	iii
Abstract	iv
Acknowledgements	vi
Dedication	vii
List of Figures	xiii
List of Tables	xvi
List of Abbreviations	xvii
1 Introduction	1
1.1 Sixth Generation Networks: Expectations, Challenges and Opportunities .	1
1.2 Reconfigurable Intelligent Surface (RIS)	2
1.2.1 What is an RIS?	2
1.2.2 RIS Use Cases	2
1.2.3 RIS vs Coverage Extension Predecessors	3
1.3 Motivation and Organization	4
1.4 Notation	7

2	Literature Review	8
2.1	RIS-Assisted Wireless Communication Systems	8
2.1.1	A SISO Example	8
2.1.2	RIS Design Objectives	10
2.1.3	RIS Prototypes and Field trials	10
2.2	Physical Layer Resource Allocation in RIS-Assisted Systems	11
2.2.1	Physical Layer Resource Allocation in Multi-antenna RIS systems .	11
2.3	CSI Acquisition	13
2.3.1	CSI Estimation For Single-User Systems	14
2.4	Data-Driven Approaches for CSI Acquisition and RIS Configuration	18
3	Joint Uplink-Downlink Resource Allocation for Multi-User RIS-Assisted Systems	20
3.1	Introduction	20
3.1.1	Contributions and Key Findings	22
3.2	System Model and Performance metrics	24
3.2.1	RIS-assisted system Model	24
3.2.2	Linear beamforming and performance metrics	25
3.3	Problem Formulation	26
3.3.1	Problem Formulation	26
3.3.2	UL/DL user weighting strategies	28
3.4	Optimizing the Joint Design, and Alternate Schemes	29
3.4.1	Updating the beamforming vectors and power control for fixed RIS configuration	30
3.4.2	Updating the RIS configuration for fixed beamforming vectors and uplink power control	32
3.4.3	Structure of beamforming vectors and channel reciprocity in RIS systems	34
3.4.4	Alternate Schemes	34

3.5	Simulation Results	35
3.5.1	System setup and parameters	35
3.5.2	Uplink-Downlink Trade-off Regions	37
3.5.3	Complexity Analysis	45
3.6	Conclusion	48
4	A Hypernetwork Framework for Learning Adaptive Beamforming Schemes in RIS Systems	50
4.1	Introduction	50
4.1.1	Contributions and Key Findings	52
4.2	System Model and Signal Transmissions	53
4.2.1	Downlink Data Transmission	53
4.2.2	Uplink Pilot-based CSI Acquisition	55
4.3	Problem Formulation	55
4.3.1	Adaptive beamforming configuration problem formulation	56
4.3.2	Analyzing the adaptive beamforming configuration problem	56
4.4	Hypernetwork-based Beamforming Framework	59
4.4.1	Hypernetwork-based beamforming	59
4.4.2	HNB Architecture	60
4.5	Proposed HNB Schemes and Benchmarks	61
4.5.1	Benchmarks	62
4.5.2	Proposed Schemes	63
4.6	Numerical Experiments	64
4.6.1	System and Channel Models	64
4.6.2	Channel Training and CSI Acquisition	66
4.6.3	Optimizing WSR under significant variation in user priorities	66
4.6.4	Optimizing WSR for Proportional-Fairness	73
4.6.5	Complexity Analysis and Comparisons	75
4.7	Conclusions	77

5	Practical Configuration of a Discrete RIS with Phase-dependent Amplitude Response	79
5.1	Introduction	79
5.1.1	Contributions and Key Findings:	80
5.2	Discrete RIS Configuration: Problem Formulation and Analysis	81
5.2.1	Virtual-antenna Selection Formulation	82
5.2.2	Analysis and a Performance Guarantee	83
5.3	Proposed RIS Configuration Scheme and Numerical Results	87
5.3.1	RIS Model and Proposed Scheme	87
5.3.2	Simulation Settings and Benchmarks	88
5.3.3	Numerical Experiments	89
5.3.4	Complexity Comparisons	93
5.4	Multi-antenna receivers	94
5.4.1	Virtual-antenna Selection Formulation and Analysis	94
5.4.2	Simulation Settings and Benchmarks	96
5.4.3	Numerical Results	97
5.5	Conclusions	100
6	Concluding Remarks and Future Work	102
6.1	Summary of Contributions and Key Findings	102
6.2	Future Research Directions	106
6.2.1	Extended RIS Architectures	106
6.2.2	RIS for Localization and Mapping Tasks	107
6.2.3	RIS deployment strategy	107
	References	108
	APPENDICES	123

A	Proofs	124
A.1	Lemma 1	124
A.2	Lemma 2	125
A.3	Proof of Proposition 4.1	125
A.4	Proof of Proposition 5.1	127
A.5	Proof of Proposition 5.2	127
A.6	Proof of Proposition 5.3	128
A.7	Proof of Proposition 5.4	129
B	Static-GNN2 implementation	130

List of Figures

1.1	Programmable signal propagation using RIS	2
2.1	SISO RIS-assisted system	9
3.1	Three TDD cases showing possible allocation of time slots in cases 1, 2 and 3. Here, the update overhead and duration are reduced in cases 2 and 3 compared to case 1.	22
3.2	(a) System diagram and (b) top-view layout for MU-MISO RIS assisted system.	24
3.3	Uplink-Downlink trade-off region for $\alpha = 0, 0.1, \dots, 1$, where for a fixed α , $\beta = 0, 0.1, \dots, 1$, with $N = 200$ and under independent weights.	37
3.4	Uplink-Downlink Rate region for $N = 200$ and equal weights.	38
3.5	Uplink-Downlink Rate region for $N = 200$ and PF weights.	39
3.6	Uplink-Downlink region for $N = 200$ and $N = 400$, under independent user weights.	40
3.7	Uplink-Downlink Rate region achieved by joint design and slicing benchmarks for different N and independent weights.	41
3.8	Uplink-Downlink Rate region achieved by joint design and alternative schemes obtained using BCD algorithm 1 as well as individual design obtained using Couenne numerical solver. Here, $N = 200$ and independent weights are employed.	42
3.9	Uplink-Downlink Trade-off region for $N = 200$ under FDD/TDD for different M with independent weights.	43
3.10	Uplink-Downlink Rate region for $N = 200$ under ZF beamforming and independent weights.	44

4.1	Considered multi-user RIS system where $t \in \{DL, UL\}$, indicates the downlink or uplink channel.	54
4.2	Hypernetwork-based Beamforming architecture. A hypernetwork is trained to output the learning parameters $\mathbf{W}_D$ of the beamforming network.	60
4.3	Top-view for the considered multi-user RIS system.	65
4.4	Average downlink WSR on testing data after different number of training epochs, for independent user weights, $T = 20$, and Rician factor of 10 dB.	67
4.5	Average downlink WSR for different number of pilots per user T , under independent user weights and Rician factors of 10 dB, with user location (e.g., HNB-GNN2-loc) and without user location (e.g., HNB-GNN2).	69
4.6	Average downlink WSR on testing data for different Rician Factors and for different number of pilots with location information, under independent user weights.	71
4.7	Average downlink WSR on testing data for different UL noise powers, DL noise power and Rician factors. Here, $T = 10$ with location information and independent user weights are used.	72
4.8	Average downlink WSR on testing data after different number of training epochs with $T = 5$ and location information, obtained under a Rician factor of 10 dB, PF user weights and for $s = 300$	74
4.9	Average downlink WSR for different number of pilots per user T , under PF user weights and for $s = 300$, with user location (e.g., HNB-GNN2-loc) and without user location (e.g., HNB-GNN2).	76
5.1	This example illustrates a scenario involving 3 RIS elements and no direct TX – RX path, where in (a) each element’s channel is represented by a triangle, square, or circle. Each RIS element introduces one of four possible phase shifts, identified by different colors: black, green, blue, or red. In (b), the possible virtual channels associated with the first RIS element are shown. In (c), the 12 possible virtual channels associated with the 3 elements are illustrated. To coherently align the reflected channels, the selected effective channels should lie within a slice of the complex plane as shown in (c).	84
5.2	Figs. 5.2a and 5.2b show the empirical CDF of the spectral efficiency performance for ($k = 1$) binary RIS. Figs. 5.2c and 5.2d show average spectral efficiency performance vs number of bits k used to control the RIS under $\alpha_{min} = 0.2$ and $\alpha_{min} = 0.6$. Here, $N = 100$	91

5.3	Figs. 5.3a and 5.3b show the empirical CDF of the spectral efficiency performance for ($k = 1$) binary RIS. Figs. 5.3c and 5.3d show average spectral efficiency performance vs number of bits k used to control the RIS under $\alpha_{min} = 0.2$ and $\alpha_{min} = 0.6$. Here, $N = 200$	92
5.4	The average spectral efficiency performance vs number of bits used to control the RIS under $\alpha_{min} = 0.2$ and $\alpha_{min} = 0.6$. Here $M = 2$ and $N = 100$	98
5.5	The average spectral efficiency performance vs number of bits used to control the RIS for both $N = 100$ and $N = 200$. Here $M = 4$ and $\alpha_{min} = 0.2$	99
5.6	The average spectral efficiency performance vs number of antennas at the receiver for both $k = 1$ and $k = 2$ bits used to control the RIS. Here $N = 100$ and $\alpha_{min} = 0.2$	100

List of Tables

4.1	BFN Architectures	62
-----	-----------------------------	----

List of Abbreviations

AWGN additive white Gaussian noise

BCD block-coordinate descent

BFN beamforming network

CPP closest point projection

CSI channel state information

DL downlink

FDD frequency-division-duplexing

HNB hypernetwork-based beamforming

LB lower bound

MISO multiple-input single-output

MU multi-user

RCG Riemannian conjugate gradient

RIS Reconfigurable intelligent surfaces

RX receiver

TDD time-division-duplexing

TX transmitter

UB upper bound

UL uplink

WMMSE weighted minimum mean square error

WSR weighted sum-rate

Chapter 1

Introduction

1.1 Sixth Generation Networks: Expectations, Challenges and Opportunities

Sixth generation systems (6G) are envisioned to enable a plethora of wireless communication use cases including hyper reliable and low-latency communication, massive communication, and immersive communication [2, 108, 104]. Next generation wireless technologies must cater to the relentless demand for increased wireless connectivity while complying with stringent quality-of-service (QoS) requirements in terms of connectivity, latency, and coverage [53].

Current systems often enhance performance by introducing active infrastructure nodes and exploiting the bandwidth available in high-frequency bands. However, deploying active network resources poses numerous practical challenges—including increased deployment and operation costs, interference, hardware and signal processing complexity, power consumption, and overall network complexity [111]. Additionally, in high-frequency systems (e.g., millimeter wave (mmWave) networks), additional active antennas or nodes may be needed to overcome the increased path loss, chance of blockage, and limited coverage that characterize high-frequency bands [53]. These measures further amplify network complexity, costs, and power consumption.

Furthermore, the focus of communications technologies has traditionally been on optimizing the transmitter and receiver, leaving the configuration of the channel itself largely beyond the control of a designer. Therefore, innovative low-cost, low-interference, and energy-efficient technologies that support emerging use cases and address these issues are

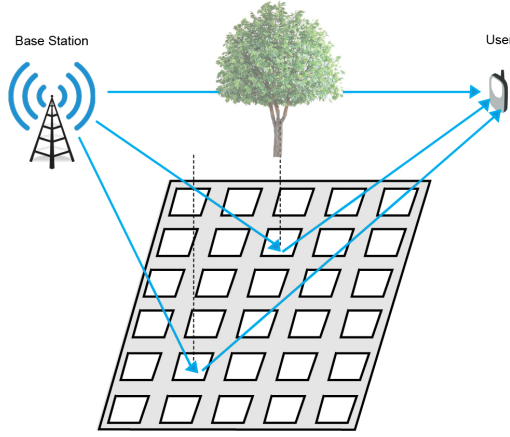


Figure 1.1: Programmable signal propagation using RIS

particularly attractive. To achieve these goals, a reconfigurable intelligent surface (RIS), also known as an *intelligent reflecting surface* (IRS), has emerged as a promising technology that offers a cost-effective and energy-efficient approach to engineering the channel between a transmitter and receiver, thereby enhancing communication system performance to support increasing QoS demands [14, 84, 28, 82].

1.2 Reconfigurable Intelligent Surface (RIS)

1.2.1 What is an RIS?

An RIS typically consists of a 2D array of a large number of tunable reflecting elements along with a low-power reconfiguration circuit. The properties of each element can be re-configured in order to dynamically reshape the impinging electromagnetic waves [28, 82]. An RIS can partially circumvent the random nature of wireless propagation by tuning a large number of reflecting elements that collectively modify the incident signal paths and engineer the channel in order to assist signal transmission, as shown in Fig. 1.1. Additionally, the nearly-passive RIS implementation inherits several favorable features such as operating in full-duplex with no self-interference, low-power consumption, and reduced hardware and signal-processing complexity. Subsequently, RIS technology promises significant system-level performance improvements with low power consumption, complexity and deployment/operation costs.

1.2.2 RIS Use Cases

An RIS can be deployed to support a large number of use cases. For instance, an RIS can create a virtual line-of-sight (LOS) signal path through the RIS cascaded channel

to overcome issues like blockage and poor coverage, as demonstrated in works such as [119, 127]. This would particularly bring significant improvements to the performance of high-frequency communications systems such as millimeter-wave systems which suffer from increased blockage and diminished coverage [80, 140]. Furthermore, multi-cellular RIS-assisted systems are considered in works like [83], where RIS deployment is shown to substantially extend the coverage of a BS, particularly benefiting cell-edge users. It has also been demonstrated that an RIS can deliver substantial performance improvements to all users in the network, whether or not they are affected by blockage [109].

Apart from virtual line-of-sight coverage extension, an RIS can provide additional degrees of freedom to optimize performance. Particularly, by tuning the wireless channel to constructively and destructively combine signal paths, the RIS can be designed to enhance the performance of the underlying wireless system. This can be viewed as a form of passive beamforming, which can be jointly optimized with the (active) beamforming at the transmitter and/or receiver [109, 14, 111]. Specifically, an RIS can be deployed to improve signal power at the receiver and diminish interference for unintended users [40, 135].

1.2.3 RIS vs Coverage Extension Predecessors

The paradigm of *cooperative communications* attempts to deploy additional network resources such as passive reflectors [56] and relays [58, 68] which provide additional signal-propagation paths to extend coverage and enhance the performance of the underlying wireless system. In this section, we outline how an RIS fits into the cooperative communications paradigm and highlight the key distinctions between RIS and previous technologies.

RIS vs Passive Reflectors

The wireless propagation environment can be enriched by deploying passive specular (mirror-like) reflectors that can establish additional scattering components. Depending on the locations of the users and environmental factors, the coverage of the transmitter can be extended and the overall performance of the system can be improved. However, due to their static nature, passive reflectors are incapable of adapting to the dynamic nature of wireless propagation channels [56]. Conversely, reconfigurable surfaces have tunable properties which can be programmed to create additional signals (or signal paths) and engineer the wireless channel to improve the performance of the underlying system [27].

RIS vs Active Relay

Active relays consume significant power in order to regenerate or amplify incident signals and may introduce additive noise components. Relay-assisted systems can be categorized as full-duplex or half-duplex. Half-duplex relays suffer from lower spectral efficiency as the transmitter and relay transmit on different time-frequency channel resources. Under full-duplex relaying, both the transmitter and relay transmit on the same time-frequency channel resources. However, a full-duplex relay suffers from self-interference which must be suppressed at the cost of additional hardware and signal processing [27]. This in turn increases the complexity, power consumption and costs associated with relayed systems.

Conversely, an RIS strikes a balance between passive reflectors and active relays as it is commonly realized using passive and controllable scattering elements. The properties of each element can be re-configured in order to dynamically reshape the impinging electromagnetic waves [28, 82]. A nearly-passive RIS implementation exhibits several favorable features such as operating in full-duplex with no self-interference, low-power consumption, and reduced hardware and signal-processing complexity.

Performance comparisons between RIS-assisted systems and relay-assisted systems are featured in several works, including [27, 48, 81]. In [16], the performance of an RIS-assisted single-input single-output (SISO) system is compared to that of a relayed system that employs a decode-and-forward (DF) protocol. It is shown that with a sufficient number of passive elements, an RIS outperforms a DF relay in terms of achievable rate. Moreover, the RIS-assisted system is shown to achieve a higher energy efficiency. Similarly, in [81], numerical experiments show that an RIS equipped with a sufficiently large number of elements is capable of outperforming a relay-supported system in terms of energy efficiency and achievable rate, even under high-insertion losses at the RIS elements.

1.3 Motivation and Organization

Due to the lack of active components (e.g., RF chains), an RIS provides a low-cost and nearly-passive solution to smart radio environments where the wireless propagation channel can be dynamically controlled. This, however, gives rise to many interesting technical challenges as the RIS is typically incapable of performing advanced digital signal processing. Hence, CSI acquisition and RIS configuration need to be performed at the BS/user side and then suitable control signals fed back to the RIS. Moreover, in a multi-antenna RIS-assisted system, the RIS configuration is jointly optimized along with the underlying active beamforming and/or combining. The coupling between the RIS configuration and

other optimization variables often leads to non-convex optimization problems. Therefore, efficiently obtaining high-quality channel estimates and optimizing the RIS reflection coefficients are critical for practical RIS integration into communication systems. This research focuses on addressing these challenges by investigating resource allocation problems in RIS-assisted systems and developing efficient RIS configuration schemes for various practical scenarios. The methods presented aim to support the integration of RIS into current communication systems.

The rest of this thesis is as follows:

In Chapter 2, a background on RIS operating principles and design objectives is provided. We highlight recent experimental works and RIS prototype development efforts which illustrate the significant gains that can be brought by deploying an RIS into a practical communication system. Next, we survey common optimization methods and outline CSI acquisition techniques associated with RIS-assisted systems. In addition, we briefly outline data-driven approaches that can be employed to configure the RIS and acquire the necessary CSI. Subsequently, we discuss the integration of an RIS into current systems and the associated practical deployment problems.

In Chapter 3, we propose the joint uplink-downlink configuration of a reconfigurable intelligent surface (RIS) for multi-user frequency-division-duplexing (FDD) and time-division-duplexing (TDD) systems.¹ This is motivated in FDD since uplink and downlink transmissions occur simultaneously and hence an RIS must be jointly configured for both transmissions. In TDD, while a joint design is not strictly necessary, it can significantly reduce feedback overhead as well as provide reductions to power consumption, and configuration periods associated with updating the RIS. To compute trade-offs between uplink and downlink rates achieved by a joint design, a weighted-sum problem is formulated and optimized using a developed block-coordinate descent algorithm. The resulting uplink-downlink trade-off regions are investigated by numerical simulation to gain insights into different scenarios.

In Chapter 4, we develop a learning-based framework that exploits limited pilot information to optimize RIS systems while catering to varying service priorities and fairness via user weights.² First, an adaptive beamforming configuration problem is formulated to

¹This chapter is based, in part on the paper [4]: M. S. Abouamer and P. Mitran, "Joint Uplink-Downlink Resource Allocation for Multiuser IRS-Assisted Systems," in *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10918-10933, Dec. 2022.

²This chapter is based, in part on the paper [5]: M. S. Abouamer and P. Mitran, "Flexible Resource Allocation in IRS-assisted Systems using Hypernetworks," *IEEE Wireless Communications and Networking Conference (WCNC)*, Glasgow, United Kingdom, 2023, pp. 1-6.

generate RIS beamforming configurations that optimize the weighted sum-rate. This problem is analyzed and shown, under mild regularity conditions, to attain a maximum. To learn approximate solutions, a novel hypernetwork-based beamforming (HNB) framework is proposed. Particularly, a beamforming network (BFN) exploits available information, including pilots, to generate optimized beamforming configurations. Rather than learning one BFN, a hypernetwork is trained to dynamically generate BFN learning parameters from an input conditioning vector. When the conditioning vector is chosen as the user weights, the trained HNB can tune the BFN to the user weights without the need for retraining. Additionally, employing the HNB to also tune the BFN to location information can be employed to reduce the pilots needed to generate optimized RIS configurations.

In Chapter 5, to relax hardware constraints and enable RIS implementation with diverse technologies, we consider configuring an RIS using discrete phase shifts that exhibit non-linear amplitude responses.³ This problem is formulated as a constrained virtual-channel selection problem. When considered over subsets with maximum phase-variation $\theta_{\max} < \pi$, the solution is shown to be bounded between two matroid-constrained problems. We illustrate configuration problems whose optimal solution lies within such subsets and show that a greedy framework guarantees a worst-case $|2 \log_2 \cos(\theta_{\max}/2)|$ optimality gap uniformly over all system parameters. Subsequently, a low-complexity framework is proposed for general discrete configuration problems. Numerical experiments demonstrate the efficacy of the proposed RIS configuration scheme.

In Chapter 6, concluding remarks and key findings of this research are presented. Additionally, potential directions for future research are discussed.

³This chapter is based, in part on the paper [6]: M. S. Abouamer and P. Mitran, "Practical Configuration of a Discrete RIS With Phase-Dependent Amplitude Response," in *IEEE Communications Letters*, vol. 28, no. 11, pp. 2678-2682, Nov. 2024.

1.4 Notation

The sets of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$. Moreover, $\Re(\cdot)$ and $\Im(\cdot)$ are used to represent the real and the imaginary parts of a complex variable, respectively. The transpose, conjugate and conjugate transpose operator are denoted using the superscripts $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^H$, respectively. Moreover, for $z = re^{j\theta} \neq 0$ where $-\pi < \theta \leq \pi$, $\arg(z) = \theta$. Vectors are denoted using lower case bold symbols (e.g $\mathbf{v}$) and matrices are denoted using upper case bold symbols (e.g $\mathbf{M}$). For a vector $\mathbf{v} \in \mathbb{C}^N$, $\text{diag}(\mathbf{v})$ is the $N \times N$ diagonal matrix whose diagonal entries are filled with the N elements of $\mathbf{v}$. Moreover, the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is indicated by $\mathbf{A} \otimes \mathbf{B}$. The expectation of a random variable (RV) is denoted by $\mathbb{E}\{\cdot\}$ and $\mathcal{CN}(0, \sigma^2)$ is used to denote a circularly symmetric complex Gaussian RV with a mean of 0 and variance σ^2 .

Chapter 2

Literature Review

In this chapter, we provide an overview of the operation and optimization of an RIS-assisted wireless communication system. We then discuss the challenges related to CSI acquisition and RIS configuration. Common optimization methods and CSI acquisition techniques developed to address these challenges are surveyed.

2.1 RIS-Assisted Wireless Communication Systems

2.1.1 A SISO Example

From an operation point of view, an RIS is a 2D array comprised of a large number of passive tunable elements whose properties can be reconfigured to dynamically shape the incident signals. For illustration, we consider a single-carrier single-input single-output (SISO) system assisted by an N -element RIS, as shown in Fig. 2.1. For an RIS of N elements, as commonly adopted in the literature [109, 138, 40, 121, 128, 52], the n^{th} element introduces a phase shift $\varphi_n \in [0, 2\pi)$ and the RIS reflection coefficients can be represented using $\boldsymbol{\theta} = [\theta_1, \dots, \theta_N]$, where $\theta_n = e^{-j\varphi_n}$ for $n \in \{1, \dots, N\}$. The RIS configuration is then expressed as $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta})$.

The signal reflected by the n^{th} RIS element is the product of the incident signal and the reflection coefficient θ_n . Hence, the effective reflected channel through the n^{th} RIS element is the product of the following three terms: transmitter-RIS channel coefficient, RIS configuration and the RIS-receiver channel coefficient. When the unit power information

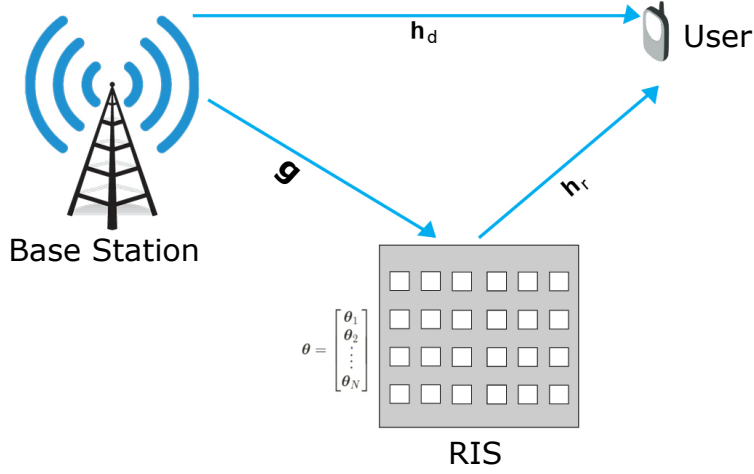


Figure 2.1: SISO RIS-assisted system

signal x is transmitted with power P_t , the received signal is given by

$$y = (h_d + \mathbf{g}\Theta\mathbf{h}_r)\sqrt{P_t}x + z, \quad (2.1)$$

where $h_d \in \mathbb{C}$, $\mathbf{h}_r \in \mathbb{C}^{N \times 1}$ and $\mathbf{g} \in \mathbb{C}^{1 \times N}$ are the baseband equivalent channels between the transmitter and the receiver, the transmitter and the RIS, and the RIS and the receiver, respectively. Moreover, $z \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at the receiver.

For this RIS-assisted system, the achievable rate is given by $r = \log_2(1 + \gamma)$ where γ is the received signal-to-noise ratio (SNR) which is given by

$$\gamma = \frac{P_t |h_d + \mathbf{g}\Theta\mathbf{h}_r|^2}{\sigma^2}. \quad (2.2)$$

Under this scenario, maximizing the achievable rate is equivalent to maximizing the received signal power and the associated optimization problem can be formulated as

$$\max_{\boldsymbol{\theta}} \left| h_d + \sum_{n=1}^N h_{r,n} g_n e^{-j\varphi_n} \right|^2 \quad (2.3a)$$

$$\text{subject to} \quad 0 \leq \varphi_n \leq 2\pi, \quad n = 1, \dots, N. \quad (2.3b)$$

In this simple SISO case, the solution of aligning the reflected channels through the RIS elements to the direct channel between the transmitter (TX) and receiver (RX) maximizes the objective in (2.3). Subsequently, the optimal phase shifts are $\varphi_n^* = -\angle h_d + (\angle g_n + \angle h_{r,n})$ for each RIS element $n \in \{1, \dots, N\}$. The received power here depends on the number of RIS elements N . Therefore, with sufficient number of elements, deploying an RIS can, in principle, bring substantial gain in the performance of a communication system [14, 111].

2.1.2 RIS Design Objectives

An RIS can be integrated with conventional multi-user communication systems and optimized to meet a variety of system design objectives. For instance, an RIS can be jointly optimized with active beamforming (at the transmitter and/or receiver) to maximize the weighted sum-rate (WSR) [40], minimize total power subject to SINR constraints [48, 42], enhance physical-layer security [21, 70, 75, 69], and improve capacity in both indoor and outdoor environments [86, 128]. Rate optimization of the MIMO broadcast (BC) channel was examined in [87] and more recently in [20].

Furthermore, an RIS can efficiently engineer the wireless channel to achieve favorable channel conditions. For example, an RIS can be used to improve the rank of MIMO channels, thereby enhancing spatial multiplexing capabilities [141, 135]. Additionally, by designing the RIS phase shifts to convert slow fading into fast fading, the RIS can help support vehicular communications [49].

The interaction between RIS designs and other emerging wireless paradigms, such as non-orthogonal multiple access (NOMA) [45], full duplex (FD) communication [132, 114], and simultaneous wireless information and power transfer (SWIPT) [66, 36, 33], is also explored in the literature. By introducing engineered signal paths, an RIS is expected to play a crucial role in enabling sensing and localization tasks [55]. Moreover, several recent works have discussed the advantages of deploying an RIS to support integrated sensing and communication (ISAC) paradigms [23, 72, 63, 30].

2.1.3 RIS Prototypes and Field trials

Both academic and industrial bodies have recently shown a growing interest in fabricating RIS prototypes. Most experimental studies focus on fabricating an RIS consisting of an array of passive reflecting elements, configurable using low-power circuits [112]. In [85], a sub-6 GHz RIS prototype was designed and deployed to provide virtual LoS, improving indoor user performance and delivering up to a 10 dB increase in received signal power compared to scenarios without an RIS. Similarly, the work in [116] demonstrated comparable performance improvements in millimeter wave (mmWave) communication systems. RIS prototypes were also integrated with 5G commercial networks in [62, 92] to enhance both cell-edge and overall system performance. For instance, [92] demonstrated a 4 dB gain in reference signal received power (RSRP) for cell-edge users and a 3.8 dB increase in average RSRP compared to non-RIS deployment. Consequently, with a focus on coverage extension, these field trials have validated the performance gains achieved through RIS deployment in both indoor and outdoor environments [85, 62, 92, 116].

2.2 Physical Layer Resource Allocation in RIS-Assisted Systems

The integration of an RIS into a wireless communication system necessitates the need to jointly configure the RIS reflection coefficients along with the configurations/resources associated with the underlying system. In this section, we review resource allocation problems in RIS-assisted systems. Moreover, we present an overview of the optimization methods employed to efficiently optimize different RIS design problems.

2.2.1 Physical Layer Resource Allocation in Multi-antenna RIS systems

Beyond the simple example presented in Section 2.1.1, optimizing the RIS configuration involves solving a non-trivial optimization problem. Particularly, for multi-antenna systems, the RIS is typically jointly optimized along with the underlying active beamforming and/or combining. Thus, the coupling between the RIS configuration and other optimization variables usually gives rise to non-convex optimization problems.

For illustration, consider a single-carrier downlink MISO system where the transmitter is equipped with M antennas and is assisted by an RIS composed of N passive elements. As commonly adopted in the literature, the goal is to configure the phase shift φ_n associated with each element $n \in \{1, \dots, N\}$ with reflection coefficient $\theta_n = e^{-j\varphi_n}$. This results in a unit-modulus constraint on each of the reflection coefficients, i.e., $|\theta_n| = 1$ and must be handled by the RIS optimization techniques.

Let $\mathbf{h}_d \in \mathbb{C}^{M \times 1}$, $\mathbf{G} \in \mathbb{C}^{M \times N}$, and $\mathbf{h}_r \in \mathbb{C}^{N \times 1}$ represent the channels between the transmitter and a single-antenna receiver, the transmitter and the N -element RIS, and the RIS and the receiver, respectively. The rate maximization (equivalently received power maximization) problem can then be formulated as

$$\max_{\boldsymbol{\Theta}, \mathbf{w}} \quad \left| (\mathbf{h}_d + \mathbf{G}\boldsymbol{\Theta}\mathbf{h}_r)^H \mathbf{w} \right|^2 \quad (2.4a)$$

$$\text{subject to} \quad \|\mathbf{w}\|^2 \leq P_t, \quad (2.4b)$$

$$|[\boldsymbol{\Theta}]_{n,n}| = 1, \quad n = 1, \dots, N. \quad (2.4c)$$

where $\mathbf{w} \in \mathbb{C}^{M \times 1}$ is the downlink beamforming vector and $\boldsymbol{\Theta}$ is diagonal. The unit modulus constraints and the coupling of the decision variables in the objective of (2.4a) result in a difficult non-convex optimization problem over $\boldsymbol{\Theta}$ and $\mathbf{w}$. Subsequently, a common

goal in the RIS literature is to devise heuristic designs and/or iterative algorithms that can efficiently find a sub-optimal solution of the corresponding RIS optimization problem [109].

Extending the RIS configuration problem to more general multi-user scenarios usually results in intractable coupled non-convex optimization problems. Consequently, block-coordinate descent (BCD) algorithms that decompose the coupled optimization variables into blocks of decoupled sub-problems are common in the RIS literature [40, 7, 18].

The decoupled beamforming and power control problems can be optimized using techniques such as fractional programming (e.g. [94, 40]) and the weighted MMSE algorithm as adopted in [126, 7]. Moreover, heuristic beamforming techniques such as zero-forcing (ZF), maximum ratio transmission (MRT) and maximum ratio combining (MRC) can be employed to optimize the beamforming/combining vectors in RIS-assisted systems. In particular, ZF, MRT and/or MRC can be leveraged to develop low-complexity algorithms that optimize the beamforming/combining vectors in RIS-assisted massive MIMO systems [107, 137].

RIS Configuration optimization

For the decoupled RIS optimization sub-problem, to address the unit-modulus constraints, the approaches adopted in the literature include constraint relaxation and manifold optimization. In the constraint relaxation approach, the unit modulus constraint is usually ignored, then the relaxed problem is optimized and finally a heuristic is employed to directly re-enforce the unit modulus constraint. A common heuristic is to normalize the solution of the relaxed problem (by dividing each of the n components of the solution to the relaxed problem by its norm) resulting in $|\theta_n| = 1$.

Another relaxation method is to re-formulate the joint RIS phase-shift/beamforming as a semi-definite problem (SDP) with a rank-1 constraint. The rank-1 constraint is then ignored, the relaxed SDP is optimized and then a heuristic (e.g. Gaussian randomization) is used to re-enforce the rank-one constraint [118, 71]. For the manifold optimization approach, the unit-modulus constraints on each phase shift constitutes a complex circle manifold and hence the phase shifts can be optimized using Riemannian conjugate gradient (RCG) methods. The use of manifold optimization methods to optimize the RIS phase shifts is featured in earlier works (e.g. [18, 113, 22]) as well as more recent works ([99, 138]).

RIS-assisted OFDM systems extend the benefits of an RIS into multi-carrier systems and are featured in works such as [65, 60, 118, 61, 136]. It is common to model an N -element RIS using N independent single-tap reflection coefficients, as described in Section 2.1.1.

Under this model, the frequency-selective cascaded TX-RIS-RX channel is the convolution of the TX-RIS multipath channel, RIS reflection coefficient and the RIS-RX multipath channel. This in turn means that a frequency non-selective RIS needs to be optimized with a single configuration that caters to all the OFDM subcarriers. In [118], successive convex approximation (SCA) and semi-definite relaxation (SDR) were used to develop an algorithm that converges to a stationary point of the problem of maximizing the achievable rate in a SISO RIS-assisted OFDM system. There, it was found that the RIS-assisted OFDM system achieves a performance improvement compared to an OFDM system with no RIS. In [65], a technique/algorithm referred to as *strongest tap maximization* (STM) was developed to optimize an OFDM RIS configuration problem by optimizing a number of narrowband configuration problems. The basic idea of STM is to select an RIS configuration that aligns the RIS phase shifts to the strongest time-domain channel tap. This heuristic was shown to work particularly well for a number of practical scenarios [65, 14].

Configuring the RIS with Discrete Phase shifts

Under the commonly adopted RIS reflection model, each element n can be configured with continuously controllable phase shift $\varphi_n \in [0, 2\pi)$. Indeed, RIS prototypes with continuously controllable phase shifts were demonstrated in works such as [32]. However, to relax hardware constraints and enable RIS implementation with diverse technologies, configuring the RIS with discrete phase shifts has been discussed in several works such as [110, 123, 89, 88]. For instance, [89, 88] considered directly configuring the RIS with discrete phase shifts. Moreover, works such as [110, 123] have demonstrated that the fine quantization of optimized continuous phase shifts (e.g., using a 4-bit quantization of phase shifts) allows a discrete RIS to achieve (nearly) the same performance as that associated with a continuous RIS. Thus, the commonly adopted continuous RIS reflection model can serve as a useful intermediate step in developing effective methods for discrete RIS implementations.

2.3 CSI Acquisition

For an RIS to provide substantial performance gains, a relatively large number of RIS elements is needed. Accordingly, RIS-related channel estimation problems are challenging as the number of channel coefficients scales with the number of RIS elements. A brief overview of CSI acquisition techniques in RIS systems is presented in this subsection.

2.3.1 CSI Estimation For Single-User Systems

Estimation Problem Setup

For illustration, we consider a single-carrier system where an M -antenna base station serves a single-antenna user and is assisted by an N -element RIS. For an uplink scenario and under the assumption of quasi-static fading channels, when a pilot symbol $x_i \in \mathbb{C}$ is transmitted in a particular time slot i , the signal received by the BS is

$$\mathbf{y}_i = (\mathbf{h}_d + \mathbf{G}\boldsymbol{\Theta}_i\mathbf{h}_r)\sqrt{P_t}x_i + \mathbf{z}_i, \quad (2.5)$$

where $\mathbf{h}_d \in \mathbb{C}^{M \times 1}$, $\mathbf{G} \in \mathbb{C}^{M \times N}$ and $\mathbf{h}_r \in \mathbb{C}^{N \times 1}$ are the channels between the BS and user, BS and RIS and RIS and user, respectively. Moreover, $\mathbf{z}_i \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ is the additive white Gaussian noise (AWGN) and $\boldsymbol{\Theta}_i = \text{diag}(\boldsymbol{\theta}_i)$ with $\boldsymbol{\theta}_i$ being the RIS configuration at time slot i . With the cascaded channel $\tilde{\mathbf{G}} = \mathbf{G} \text{diag}(\mathbf{h}_r)$, the received signal in (2.5) can be expressed as

$$\mathbf{y}_i = (\mathbf{h}_d + \tilde{\mathbf{G}}\boldsymbol{\theta}_i)\sqrt{P_i}x_i + \mathbf{z}_i. \quad (2.6)$$

Rather than estimating the individual components $\mathbf{G}$ and $\mathbf{h}_r$, the RIS channel estimation literature focuses on estimating the cascaded channel $\tilde{\mathbf{G}}$. This follows since the cascaded channels are sufficient for configuring the RIS and can be estimated using the pilots received at the BS/user [84, 134].

For each time slot $i \in \{1, \dots, I\}$ with I being the number of channel training time slots, a training symbol $x_i \in \mathbb{C}$ is sent and the RIS reflection coefficient is set to $\boldsymbol{\theta}_i \in \mathbb{C}^{N \times 1}$. For $\tilde{\mathbf{g}}_n$ being the n^{th} column of the cascaded channel matrix $\tilde{\mathbf{G}} \in \mathbb{C}^{M \times N}$ and $\mathbf{h}_c = [\mathbf{h}_d^T, \tilde{\mathbf{g}}_1^T, \dots, \tilde{\mathbf{g}}_N^T]^T$, the I received signals can then be compactly written as [100]

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_I \end{bmatrix} \quad (2.7a)$$

$$\mathbf{y}_i = \sqrt{P_t} \mathbf{X}_i \mathbf{h}_c + \mathbf{z}_i, \quad (2.7b)$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_I \end{bmatrix} \quad (2.8a)$$

$$\mathbf{X}_i = (x_i \tilde{\boldsymbol{\theta}}_i \otimes \mathbf{I}_M), \quad (2.8b)$$

$$\tilde{\boldsymbol{\theta}}_i = [1, \boldsymbol{\theta}_i^T]. \quad (2.8c)$$

The goal of the channel estimation problem is to design the pilot information matrix $\mathbf{X}$, described in (2.8), in order to find an estimate $\hat{\mathbf{h}}_c$ of the channel vector $\mathbf{h}_c$. Moreover, the quality/accuracy of these channel estimates is typically assessed using a particular criteria such as minimizing the mean-squared-error.

The CSI estimation techniques can then be classified into unstructured channel estimation methods and structured channel estimation methods. In unstructured channel models (that typically characterize sub-6 GHz channels), the channel coefficients are directly estimated. For millimeter-wave communication systems, sparse multipath scattering grants an efficient parameterization of the channel and hence the channel can be expressed in terms of a smaller number of parameters that can be estimated with lower complexity.

Unstructured channel estimation

For a given pilot information matrix $\mathbf{X}$, classical channel estimation can be extended to directly estimate $\mathbf{h}_c$. Three methods are common in the RIS literature: least-square (LS), linear minimum mean-squared-error (LMMSE) and data-driven approaches. Both LS and LMMSE are discussed in this section with the data-driven approaches discussed in Section 2.4.

Channel estimation in RIS-assisted systems is typically performed over a number of time slots I , which is referred to here as the *channel training period*. The channel training period affects the number time slots available for data transmission and hence it is an important factor in the performance of a practical RIS system.

When LS is employed to estimate the channel coefficient, then for a given $\mathbf{X}$, $\mathbf{h}_c$ is estimated as

$$\hat{\mathbf{h}}_c = \arg \min_{\mathbf{h}_c} \|\mathbf{y} - \sqrt{P_t} \mathbf{X} \mathbf{h}_c\|^2 \quad (2.9)$$

The estimation problem in (2.9) has a closed form solution

$$\hat{\mathbf{h}}_c = \frac{1}{\sqrt{P_t}}(\mathbf{X}^H \mathbf{X})^{-1} \mathbf{X}^H \mathbf{y}. \quad (2.10)$$

It then remains to optimize the pilot RIS configurations $\{\boldsymbol{\theta}_i\}_{i=1}^I$ which is featured in works such as [73, 118, 50, 136].

Under LS estimation methods, a simple scheme for choosing the pilot RIS configurations $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_I\}$ is to use an on/off scheme of $I = N + 1$ time slots as suggested in [73]. In the first time slot, all the RIS elements are switched off (assuming all $\{\alpha_n\}_{n=1}^N$ can be switched between 0 and 1) and for the next N time slots, only one RIS element is switched on. The received signals over the $N + 1$ time slots are stacked into the received vector $\mathbf{y}$ and then the closed-form of (2.10) can be used to find the LS estimate of $\mathbf{h}_c$.

The on/off schemes are simple in nature but are not optimized to reduce estimation errors [50]. Moreover, with only one element being turned on, the power reflected by an RIS and the received SNR are diminished. Thus, the accuracy of the channel estimates is reduced. In [50], the pilot RIS configurations are selected based on the columns of a DFT matrix. Compared to the aforementioned on/off schemes, the DFT matrix approach improves the received SNR as all the RIS elements are kept on during the channel training phase and hence estimation errors are shown to be smaller [50].

In order to exploit partial channel knowledge such as second order statistics, LMMSE can be employed for channel estimation. Under LMMSE estimation schemes, the goal is to find an estimator $\hat{\mathbf{h}}_c = \mathbf{W}\mathbf{y}$ such that

$$\mathbf{W} = \arg \min_{\tilde{\mathbf{W}}} \mathbb{E} \|\mathbf{W}\mathbf{y} - \mathbf{h}_c\|^2. \quad (2.11)$$

With the linear measurement model in (2.7), the LMMSE estimate of $\mathbf{h}_c$ can be found in closed form where the estimator depends on the observation vector $\mathbf{y}$ and the second order statistics of the channel and noise [54]. Similar to the LS case, the performance of the LMMSE channel estimates depends on the pilot RIS configurations $\{\boldsymbol{\theta}_i\}_{i=1}^I$. The pilot RIS configuration can be optimized to minimize the MSE, however, it amounts to solving a non-convex optimization problem [59]. Fortunately, the DFT-based channel training was shown to perform close to the optimal RIS pilots and hence it remains as a popular choice in the RIS literature [84].

Structured channel estimation

For millimeter wave communication systems, a geometric model is appropriate for modeling wireless channels due to sparse multipath scattering which is characterized by a few

dominant signal paths. Compared to direct estimation of the channel, a geometric model is specified by a smaller number of parameters and can potentially offer significant reduction in the channel estimation overhead [100].

The L -path channel between the BS and the RIS can be expressed as

$$\mathbf{G} = \sum_{l=1}^L \gamma_l^{BI} \mathbf{a}_I(\vartheta_l^I, \varphi_l^I) \mathbf{a}_B^H(\vartheta_l^B, \varphi_l^B), \quad (2.12)$$

where γ_l^{BI} is the complex gain associated with the l -th path between the BS and the RIS. The column vector $\mathbf{a}_I(\vartheta_l^I, \varphi_l^I)$ indicates the array response at the RIS and is dependent on the elevation angle-of-arrival (AoA) ϑ_l^I and the azimuth AoA φ_l^I . Similarly, the column vector $\mathbf{a}_B(\vartheta_l^B, \varphi_l^B)$ is the array response at the BS with elevation and azimuth angles of departure (AoD) given by ϑ_l^B and φ_l^B , respectively. Similarly, by parametrizing the direct BS-user channel $\mathbf{h}_d$ and the RIS-user channel $\mathbf{h}_r$, the channel estimation problem can be expressed in terms of two vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ that capture the unknown complex gains and AOA/AOD information, respectively.

In [84, 105], the signal received in time slot i in (2.5) is expressed as

$$\mathbf{y}_i = \mathbf{X}_i \mathbf{A}(\boldsymbol{\omega}) \boldsymbol{\gamma} + \mathbf{z}_i, \quad i \in 1, \dots, I \quad (2.13)$$

where $\mathbf{X}_i$ is a sensing matrix that contains the channel training pilots for time slot i as described in (2.7) and (2.8). Moreover, $\mathbf{A}(\boldsymbol{\omega})$ represents the array response and depends on the AOA/AOD information vector $\boldsymbol{\omega}$. The goal of channel estimation here is to design the sensing matrix $\mathbf{X}$ and then use the received signal $\{\mathbf{y}_1, \dots, \mathbf{y}_I\}$ to estimate the unknown gains and AOA/AoD information. This problem can then be treated as a sparse signal recovery problem, for which compressed sensing techniques can be applied [105].

From Single-User to Multi-User CSI Estimation

Until this point, the discussion was limited to estimating the CSI associated with a single-user RIS-assisted system. The trivial extension to a K -user system is to repeat the estimation process K times, [77], and then the channel estimation overhead scales with the number of users. However, the channel training overhead can be significantly reduced by exploiting the fact that the BS to RIS (i.e. $\mathbf{G}$) channel is shared by all the users [37, 106, 139].

To exploit the common BS to RIS channel, a two-step approach is suggested in [37], where the channel between the BS and RIS is first estimated and then the cascaded channels are estimated in a second stage. Alternatively, a three step approach is considered in

[106], where the direct channel between the BS and users is first estimated and then the cascaded channel associated with a particular user is estimated. Finally, the cascaded channel associated with the first user is considered as a reference and the channels associated with the rest of the users are then estimated as explained in [106]. Rather than requiring $KM(N+1)$ pilots, this scheme requires $K + M + \max\left(K - 1, \text{ceil}\left(\frac{N(K-1)}{M}\right)\right)$ pilots. Moreover, the common BS-RIS channel can also be exploited in IRS-assisted millimeter-wave systems to reduce the channel estimation overhead as discussed in [139].

2.4 Data-Driven Approaches for CSI Acquisition and RIS Configuration

Data driven approaches can be exploited to address several practical issues that pertain to the operation of RIS-assisted systems. For instance, deep learning was leveraged to improve RIS designs to aid intelligent IoT [117] and mobile edge computing [46]. Moreover, data-driven approaches were employed to assist localization with RIS [130] and to improve the security of RIS systems [75]. The focus here is on how data driven approaches can be exploited to address channel estimation and physical layer based resource allocation, including RIS phase shifts and active beamforming configurations.

In RIS-assisted problems, data-driven approaches can be exploited to estimate the CSI and then the RIS reflection coefficients can be optimized using the (mathematical) optimization methods outlined in Section 2.2.1. Several RIS works feature the use of data-driven models to explicitly estimate the CSI associated with RIS systems. Compared to parametric model approaches where the channel coefficients are estimated based on a pre-defined model, a data-driven approach enables a model to be trained and shaped based on samples of data (either measured or simulated). This allows the data-driven model to learn the structure/distribution of the data in a natural setting.

In [31], a convolutional neural network (CNN) is trained to estimate the direct and cascaded channel vectors. There, the trained CNN is shown to obtain high quality CSI estimates when fed with received pilots symbols. Furthermore, another CNN was used in [59] to address the MMSE channel estimation problem. In particular, the CSI estimation problem was reformulated as an image denoising problem and then the trained CNN was used to estimate the CSI. The idea of treating the CSI estimation problem as an image denoising problem was also featured in more recent works such as [96]. Further, compressed sensing and deep learning were leveraged in order to obtain the channel estimates in RIS-assisted millimeter-wave system as featured in works such as [101].

Meanwhile, a data-driven approach can be used to optimize the RIS configurations for a given CSI (including CSI estimated using another data-driven model). The joint configuration of RIS reflection coefficients and beamforming vectors usually results in a non-convex optimization problem. Compared to mathematical optimization, a data-driven approach shifts the computational complexity burden to the training phase. In this phase, measurements/ synthetic data are used for training the parameters of the data-driven model. In the deployment stage, the RIS configurations are obtained by performing low-complexity operations on the input CSI measurements and the parameters of the trained model.

In [35], a deep neural network is trained to maximize the achievable rate in an RIS-assisted downlink single-user system. There, a fully connected deep neural network is trained using actual CSI and is optimized to output the RIS phase shifts that maximize the achievable rate. It is shown that the RIS configuration of [35] achieves a performance close to that obtained by solving a more complex SDR of the optimization problem in (2.4a). In [34], a deep reinforcement learning based model is trained to optimize the RIS phase shifts that maximize the achievable rate. Furthermore, deep reinforcement learning was leveraged in [47] to maximize the sum-rate of a downlink multi-user system by jointly optimizing the beamforming vectors and RIS configuration.

In the previous discussion, the channel estimation problem and joint beamforming/RIS configuration problems are treated as separate problems. Conversely, another paradigm is to bypass explicit CSI estimation and directly optimize the RIS phase shifts using received pilots/measurements. This implicit channel estimation paradigm can be enabled by data-driven approaches where we allow the data to jointly shape the CSI estimation and RIS configuration processes. Implicit channel estimation methods, introduced in [51], bypass the explicit channel estimation step by training a deep learning model that directly learns how to maximize the sum-rate in a downlink RIS assisted system. In [51, 52], the deep learning model is fed with received pilots and is trained to optimize RIS phase shifts and beamforming vectors. This trained deep neural network was shown to outperform explicit channel estimation approach where the channel estimation problem and joint RIS configuration problem are treated as separate problems [51, 52].

Chapter 3

Joint Uplink-Downlink Resource Allocation for Multi-User RIS-Assisted Systems

3.1 Introduction

In this chapter, we consider the practical problem of jointly optimizing the uplink weighted sum-rate (WSR) and the downlink WSR for multi-user (MU) multiple-input single-output (MISO) FDD and TDD RIS-assisted systems. We adopt a joint uplink-downlink RIS design where the same RIS configuration is used to assist both uplink and downlink transmissions in FDD and TDD systems. Due to the non-convex and coupled nature of active and passive beamforming in RIS configuration problems, we develop a block-coordinate descent (BCD) algorithm to optimize the joint design. This BCD approach partitions the overall variable vector into manageable blocks and optimizes each block individually while keeping the others fixed—transforming the active/passive RIS beamforming problem into a series of simpler subproblems. By leveraging effective algorithms, including Riemannian conjugate gradient (RCG) for RIS passive beamforming and weighted minimum mean square error (WMMSE) for downlink active beamforming, achievable trade-off regions between uplink and downlink rates are computed.

Subsequently, the resulting uplink-downlink trade-off regions are investigated by numerical simulation to gain insights into different scenarios. To evaluate the performance of the proposed joint uplink-downlink design, we compare it to configuring the RIS for uplink (UL) and downlink (DL) transmissions individually and then time-sharing these two RIS

configurations, referred to as *Individual design*. Moreover, we propose two heuristic designs called *fixed-downlink* and *fixed-uplink*. In the fixed-uplink (fixed-downlink) heuristic, the RIS configuration for uplink (downlink) is used to assist downlink (uplink) transmissions.

For FDD systems, using the same RIS configuration is essential in order to support simultaneous uplink and downlink transmissions as they occur over the same time resources. On the other hand, while it is feasible for TDD systems to support different RIS configurations for uplink and downlink transmissions, a joint RIS design can trade separately optimized RIS phase shifts for reduced signalling overhead, configuration periods associated with updating the RIS, and optimization complexity. This trade-off may be particularly appealing when the loss in spectral efficiency due to a joint design (or fixed design) compared to the individual design is marginal, as will be demonstrated for many TDD scenarios considered in Section 3.5.

Depending on the technology used to implement the RIS, quiet guard periods while updating the RIS may be required to avoid the time-varying behavior of the channel due to a changing RIS configuration.¹ Subsequently, there is a time where an RIS may not be used for communication due to feedback and quiet guard periods. Moreover, updating the RIS imposes strict constraints on time synchronization between the BS and the RIS, which in turn increases overhead, hardware complexity and delays [78]. Thus, feedback overhead and delay associated with updating an RIS are among the major challenges that face RIS implementations in practical scenarios [76, 133, 120, 90, 122, 7, 15] as the feedback duration, depending on the RIS implementation technology, may significantly reduce the time available for data transmission and hence diminish the performance improvements provided by an RIS [122, 7]. In TDD, compared with separate individual designs for uplink and downlink, a joint design reaps the benefits of halving the feedback overhead as well as reducing the configuration periods. Furthermore, a joint RIS design can also reduce the power consumption of the RIS by reducing the need to switch the RIS configuration. This follows as most of the power consumed by an RIS is used in controlling and configuring the RIS [15].

Fig. 3.1 shows three possible update schedules for TDD operation: in case 1 the conventional approach of configuring the RIS before every uplink/downlink transmission; in case 2 the joint uplink/downlink approach where the RIS is only configured before each downlink transmission; and if the channel coherence time is long enough, then a joint design may require no update over several uplink/downlink intervals as shown in case 3 of Fig. 3.1, thus further reducing signalling overhead, configuration periods and power consumption.

¹For instance, micro-electro-mechanical systems (MEMS) is one technology envisioned for RIS implementation as they are cost-effective, have high quality factors and consume relatively low power [25, 26].

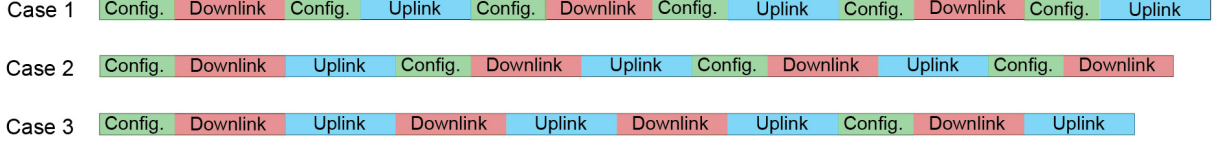


Figure 3.1: Three TDD cases showing possible allocation of time slots in cases 1, 2 and 3. Here, the update overhead and duration are reduced in cases 2 and 3 compared to case 1.

3.1.1 Contributions and Key Findings

In this chapter, a joint uplink-downlink design is proposed to support concurrent transmissions in practical multi-user RIS-assisted FDD systems as well as reduce overhead and configuration periods in practical multi-user RIS-assisted TDD systems. Moreover, two heuristic designs and two slicing benchmarks are devised in order to assess the performance of the proposed joint design under different scenarios. Subsequently, RCG, WMMSE and fractional programming are iteratively leveraged in a BCD algorithm to optimize the proposed joint design and investigate the achieved trade-off regions between achievable downlink and uplink rates in RIS systems. In particular, the achieved uplink-downlink trade-off regions are thus investigated by numerical simulation for different user priorities, beamforming strategies and FDD/TDD system parameters. Furthermore, we analyze the complexity associated with the proposed joint design and heuristic designs. This analysis is then used along with the computed regions to illustrate the trade-offs associated with different RIS designs in terms of spectral efficiency performance, overhead, configuration periods, and computational complexity. We summarize the contributions and key findings of this chapter as follows:

- We investigate joint uplink-downlink RIS design in multi-user multiple-input-single-output FDD and TDD systems. By leveraging RCG, WMMSE and FP, a joint RIS design trade-off region between UL and DL is obtained when the same RIS configuration is used.
- To evaluate the performance of the obtained joint UL-DL design, we compare it to configuring the RIS for UL and DL transmissions individually and then time-sharing these two RIS configurations, referred to as *Individual design*. Moreover, we propose two heuristic designs called *fixed-downlink* and *fixed-uplink*. In addition, two slicing benchmarks called *slicing-without-interference* and *slicing-with-interference* are also introduced.

However, one of the main drawbacks of MEMS is their slow configuration speed [26, 43].

- The impact of different user-weighting strategies (equal, proportional-fair, and independent weights) on the achieved uplink-downlink trade-off regions associated with the joint design and two fixed designs is analyzed by numerical simulation. It is found that the improvement due to joint design compared with the fixed-downlink/fixed-uplink designs becomes more significant as more flexibility is needed to accommodate user weights that vary substantially across the users.
- The impacts of both the ratio of RIS elements to active antennas at the BS and beamforming strategy (i.e., WMMSE, zero-forcing (ZF)) on the achieved trade-off regions are also investigated. As the ratio of RIS to BS elements is increased, the performance improvement due to the joint RIS design over fixed-uplink/fixed-downlink designs becomes more pronounced. Moreover, it is found that the improvement in performance due to joint design compared with the fixed uplink/downlink designs is similar under the two considered beamforming strategies.
- The uplink-downlink trade-off analysis demonstrates that the region achieved by the joint design strictly contains the regions due to slicing-without-interference and slicing-with-interference. This shows not only that a joint RIS design outperforms the slicing benchmarks but also demonstrates that interference alone is not responsible for the inferior performance achieved by a sliced RIS.
- The complexity associated with the optimized joint RIS design and fixed designs is analyzed. This analysis is then used along with the computed trade-off regions to gain insights into scenarios where a marginal loss in spectral efficiency can be traded off for reductions in overhead, configuration periods and complexity.
- In FDD, the proposed joint RIS design performs substantially better than the fixed-uplink and fixed-downlink heuristic designs and diminishes the gap to the individual design. In TDD, provided that the user weights are independent, the joint design substantially improves the spectral efficiency performance compared with the fixed-uplink and fixed-downlink heuristic designs. Otherwise, in TDD, the fixed-uplink and fixed-downlink designs perform almost as well as the individual design upper bound, offering reduced complexity, overhead, and configuration periods with almost no loss in spectral efficiency.

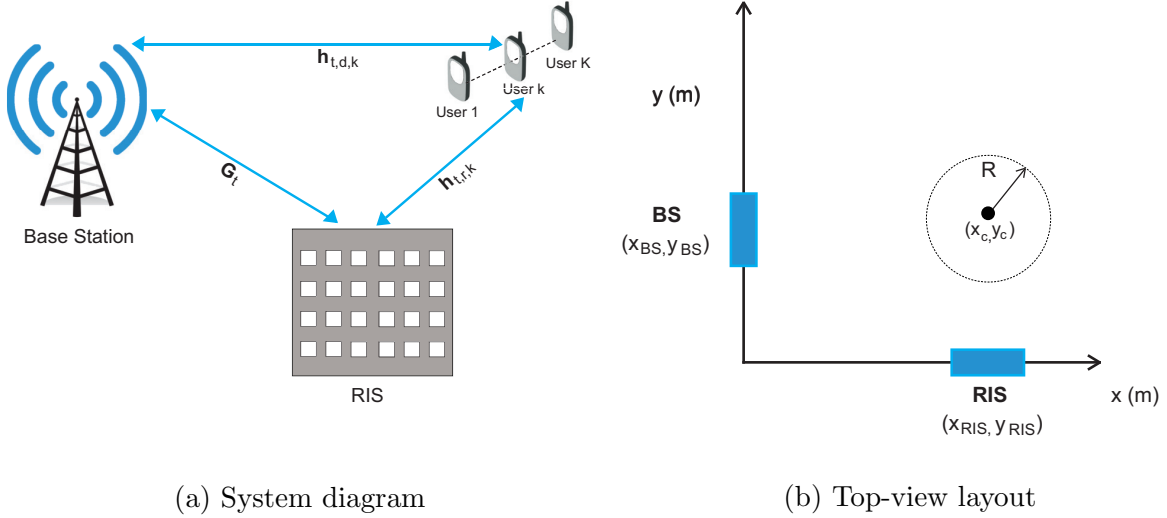


Figure 3.2: (a) System diagram and (b) top-view layout for MU-MISO RIS assisted system.

3.2 System Model and Performance metrics

3.2.1 RIS-assisted system Model

In this chapter, we consider a scenario where an M -antenna BS is assisted by an N -element RIS in order to serve K single-antenna users. The users are indexed by $\mathcal{K} = \{1, \dots, K\}$. The system diagram and layout are shown in Fig. 3.2.

In the DL, the precoded signal transmitted by the BS is expressed as $\mathbf{x}^{DL} = \sum_{k=1}^K \mathbf{w}_k^* s_k^{DL}$ where $\mathbf{w}_k \in \mathbb{C}^{M \times 1}$ and $s_k^{DL} \in \mathbb{C}$ are the downlink transmit beamforming vector and the information symbol in the downlink associated with user $k \in \mathcal{K}$, respectively. The information symbols are zero-mean and satisfy $\mathbb{E}\{s_k^{DL} (s_i^{DL})^*\} = \delta_{i,k}$, i.e., are i.i.d. with unit power.

In the UL, the signal transmitted by user k can be expressed as $x^{UL} = \sqrt{p_k} s_k^{UL}$ where $p_k > 0$ and $s_k^{UL} \in \mathbb{C}$ are the uplink power transmitted by user k and the information symbol in the uplink associated with user k , respectively. The UL information symbols satisfy $\mathbb{E}\{s_k^{UL} (s_i^{UL})^*\} = \delta_{i,k}$. Additionally, $\mathbf{v}_k \in \mathbb{C}^{M \times 1}$ is the unit-norm receive combining vector at the BS for user k .

The elements of the RIS are indexed by $\mathcal{N} = \{1, \dots, N\}$ with the vector containing the RIS phase shifts expressed as $\boldsymbol{\theta} = [\theta_1, \dots, \theta_N]$. For each RIS element $n \in \mathcal{N}$, the reflection

coefficient is represented as $\theta_n = e^{-j\varphi_n}$ where $\varphi_n \in [0, 2\pi)$ is the phase shift associated with element n . The phase-shift matrix associated with the RIS can then be expressed as $\Theta = \text{diag}(\boldsymbol{\theta})$.

All channels are quasi-static flat and it is assumed that channel state information (CSI) is perfectly known. Let $\mathbf{h}_{t,d,k} \in \mathbb{C}^{M \times 1}$ denote the baseband equivalent direct channel between the BS and user k , $\mathbf{h}_{t,r,k} \in \mathbb{C}^{N \times 1}$ denote the channel between the RIS and user k , and $\mathbf{G}_t \in \mathbb{C}^{M \times N}$ denote the direct channel between the BS and RIS, where $t \in \{UL, DL\}$ indicates whether the uplink channel or the downlink channel is considered. In TDD systems, the downlink and uplink transmissions occur over the same frequency bands but at different time slots. Moreover, channel reciprocity in TDD, as experimentally established for RIS-assisted systems, implies that the uplink and downlink channels are the same [85, 102]. In FDD, the downlink and uplink transmissions occur concurrently over different frequency bands, thus the channels are generally different for uplink and downlink transmissions.

The effective channels between the BS and user k in DL and UL are expressed as

$$\mathbf{h}_{DL,k} = \mathbf{h}_{DL,d,k} + \mathbf{G}_{DL} \Theta_{DL} \mathbf{h}_{DL,r,k} \quad (3.1a)$$

$$\mathbf{h}_{UL,k} = \mathbf{h}_{UL,d,k} + \mathbf{G}_{UL} \Theta_{UL} \mathbf{h}_{UL,r,k}. \quad (3.1b)$$

It is important to note that in general TDD systems, the RIS configuration Θ can be different for UL and DL transmissions, although at the cost of increased overhead, power consumption and configuration periods, whereas for FDD the same RIS configuration is used to support simultaneous UL-DL transmissions. In this chapter, a joint RIS configuration is used for UL and DL transmissions. However, the case where the RIS configurations can be different for UL and DL transmissions is also considered to evaluate the performance of the joint design.

3.2.2 Linear beamforming and performance metrics

The signal received by user k in the DL is

$$y_k^{DL} = \underbrace{(\mathbf{h}_{DL,k})^T \mathbf{w}_k^* s_k^{DL}}_{\text{Intended signal}} + \underbrace{\sum_{i=1, i \neq k}^K (\mathbf{h}_{DL,k})^T \mathbf{w}_i^* s_i^{DL}}_{\text{MU-interference}} + \underbrace{n_k^{DL}}_{\text{AWGN noise}}, \quad (3.2)$$

where $n_k^{DL} \sim \mathcal{CN}(0, \sigma_{DL}^2)$ is circularly symmetric complex additive white Gaussian noise (AWGN) at user k . Similarly, the signal received by the BS in UL is expressed as

$$\mathbf{y}^{UL} = \sum_{i=1}^K \sqrt{p_i} \mathbf{h}_{UL,i} s_i^{UL} + \mathbf{n}_k^{UL} \quad (3.3)$$

If $\mathbf{v}_k$ is the combining vector at the BS for user k , the combined signal is given by

$$\mathbf{v}_k^H \mathbf{y}^{UL} = \underbrace{\sqrt{p_k} \mathbf{v}_k^H \mathbf{h}_{UL,k} s_k^{UL}}_{\text{Intended signal}} + \underbrace{\mathbf{v}_k^H \left(\sum_{i=1, i \neq k}^K \sqrt{p_i} \mathbf{h}_{UL,i} s_i^{UL} \right)}_{\text{MU-interference}} + \underbrace{\mathbf{v}_k^H (\mathbf{n}_k^{UL})}_{\text{AWGN noise}}, \quad (3.4)$$

where $\mathbf{n}_k^{UL} \sim \mathcal{CN}(0, \sigma_{UL}^2 \mathbb{I}_M)$ is the AWGN at the BS.

The SINR corresponding to user k in DL and UL are given by

$$\gamma_k^{DL} = \frac{\left| (\mathbf{h}_{DL,k})^H \mathbf{w}_k \right|^2}{\sigma_{DL}^2 + \sum_{i=1, i \neq k}^K \left| (\mathbf{h}_{DL,k})^H \mathbf{w}_i \right|^2} \quad (3.5a)$$

$$\gamma_k^{UL} = \frac{p_k \left| (\mathbf{h}_{UL,k})^H \mathbf{v}_k \right|^2}{\mathbf{v}_k^H \left(\sigma_{UL}^2 \mathbf{I}_M + \sum_{i=1, i \neq k}^K p_i \mathbf{h}_{UL,i} (\mathbf{h}_{UL,i})^H \right) \mathbf{v}_k}. \quad (3.5b)$$

3.3 Problem Formulation

In this section, we formulate the RIS-assisted UL and DL weighted sum-rate problems for FDD and TDD systems. Then, in order to obtain a trade-off region between uplink and downlink rates achieved by a joint RIS design, a weighted sum problem is formulated.

3.3.1 Problem Formulation

In DL and UL, the achievable data rates in bit/sec/Hz associated with user k are given by

$$\tilde{R}_k^{DL} = \log_2(1 + \gamma_k^{DL}) \quad [\text{bit/sec/Hz}] \quad (3.6a)$$

$$\tilde{R}_k^{UL} = \log_2(1 + \gamma_k^{UL}) \quad [\text{bit/sec/Hz}]. \quad (3.6b)$$

Asymmetrical demands for UL vs DL transmission arise in practical communications systems. In TDD systems, flexible resource allocation becomes feasible by adjusting the proportion of time resources dedicated for UL vs DL. Rather than assigning equal resources to UL and DL, a normalized weight, $\alpha \in [0, 1]$, is introduced in order to capture the relative proportion of time resources are dedicated to DL vs UL in TDD systems.

In a TDD system with bandwidth W Hz, the rate in UL/DL in bit/sec is then given by

$$\bar{R}_k^{DL} = \alpha W \log_2(1 + \gamma_k^{DL}) \quad [\text{bit/sec}] \quad (3.7a)$$

$$\bar{R}_k^{UL} = (1 - \alpha) W \log_2(1 + \gamma_k^{UL}) \quad [\text{bit/sec}] \quad (3.7b)$$

where α is the fraction of time used for DL. Normalizing (3.7a) and (3.7b) by W , we have

$$R_k^{DL} = \alpha \log_2(1 + \gamma_k^{DL}) \quad [\text{bit/sec/Hz}] \quad (3.8a)$$

$$R_k^{UL} = (1 - \alpha) \log_2(1 + \gamma_k^{UL}) \quad [\text{bit/sec/Hz}]. \quad (3.8b)$$

Likewise, in an FDD system with a bandwidth of αW dedicated to downlink and $(1 - \alpha)W$ dedicated to uplink, the spectral efficiency is again given by (3.8a) and (3.8b) respectively.

The first objective is to maximize the weighted sum-rate in DL by jointly optimizing the phase-shifts at the RIS and the transmit beamforming vectors, $\mathbf{W} := \{\mathbf{w}_1, \dots, \mathbf{w}_K\}$, at the BS which are subject to a sum-power constraint. The aforementioned problem is given by

$$\max_{\boldsymbol{\Theta}_{DL}, \mathbf{W}} \quad J_{DL} = \sum_{k=1}^K \varepsilon_k^{DL} R_k^{DL} \quad (3.9a)$$

$$\text{subject to} \quad \sum_{k=1}^K \|\mathbf{w}_k\|^2 \leq P_{max}^{DL}, \quad (3.9b)$$

$$|[\boldsymbol{\Theta}_{DL}]_{n,n}| = 1, \quad n = 1, \dots, N. \quad (3.9c)$$

where $\varepsilon_k^{DL} > 0$ is the weight associated with user k in DL, and captures the relative priority of user k in the DL. Moreover, P_{max}^{DL} denotes the maximum sum-power at the BS.

In TDD, $P_{max}^{DL} = W S_{max}^{DL}$ where S_{max}^{DL} is the maximum downlink power spectral density in Watts/Hz. Similarly, the noise power in downlink is expressed as $\sigma_{DL}^2 = W N_0$ where N_0 is the noise power spectral density. In FDD, $P_{max}^{DL} = \alpha W S_{max}^{DL}$ and $\sigma_{DL}^2 = \alpha W N_0$.

The second objective is to maximize the weighted sum-rate in UL by jointly optimizing the phase-shifts at the RIS, the unit-norm receive beamforming vectors at the BS defined using $\mathbf{V} := \{\mathbf{v}_1, \dots, \mathbf{v}_K\}$ and the power control at the users defined using $\mathbf{P} := [p_1, \dots, p_K]$ where each p_k is subject to a max-power constraint P_{max}^{UL} . Maximizing the WSR in UL is

expressed as

$$\max_{\Theta_{UL}, \mathbf{V}, \mathbf{P}} \quad J_{UL} = \sum_{k=1}^K \varepsilon_k^{UL} R_k^{UL} \quad (3.10a)$$

$$\text{subject to} \quad 0 \leq p_k \leq P_{max}^{UL}, \quad k = 1, \dots, K. \quad (3.10b)$$

$$|[\Theta_{UL}]_{n,n}| = 1, \quad n = 1, \dots, N. \quad (3.10c)$$

where $\varepsilon_k^{UL} > 0$ is the weight associated with user k in UL that captures the relative priority of user k in UL. Moreover, P_{max}^{UL} denotes the maximum power that can be transmitted by a user.

In TDD, $P_{max}^{UL} = W S_{max}^{UL}$ where S_{max}^{UL} is the maximum uplink power spectral density in Watts/Hz. Additionally, the noise power in uplink is $\sigma_{UL}^2 = W N_0$. In FDD, $P_{max}^{UL} = (1 - \alpha) W S_{max}^{UL}$ and $\sigma_{UL}^2 = (1 - \alpha) W N_0$.

For a joint RIS design, problems (3.9) and (3.10) are coupled through the RIS phase shifts $\Theta_{DL} = \Theta_{UL} := \Theta$. The multi-objective optimization problem is scalarized by formulating a weighted sum problem (WSP) by introducing a weight $\beta \in [0, 1]$ that captures the relative priority given to DL vs UL when optimizing the RIS configuration. The WSP problem is then formulated as

$$\max_{\Theta, \mathbf{V}, \mathbf{W}, \mathbf{P}} \quad J_{WSP} = \alpha \beta \sum_{k=1}^K \varepsilon_k^{DL} \log_2(1 + \gamma_k^{DL}) + (1 - \alpha)(1 - \beta) \sum_{k=1}^K \varepsilon_k^{UL} \log_2(1 + \gamma_k^{UL}) \quad (3.11a)$$

$$\text{subject to} \quad (3.9b), (3.10b) \quad (3.11b)$$

$$|[\Theta]_{n,n}| = 1, \quad n = 1, \dots, N. \quad (3.11c)$$

3.3.2 UL/DL user weighting strategies

The weighting strategy of the UL/DL user is an important design parameter that can be chosen to prioritize users. Three weighting strategies are considered: *equal*, *proportional-fair* and *independent* weights, and the weights are always normalized such that $\sum_{k=1}^K \varepsilon_k^{UL} = 1$ and $\sum_{k=1}^K \varepsilon_k^{DL} = 1$.

Under the equal weight strategy, $\varepsilon_k^{UL} = \varepsilon_k^{DL} = \frac{1}{|\mathcal{K}|}$ which corresponds to providing uniform priorities to the users. In order to provide flexibility in assigning resources while maintaining fairness, the weights in (3.11) can also be chosen based on a proportional-fair

(PF) strategy. In particular, memory is incorporated into the system in order to capture the history of the data transmissions provided for each user in the past. Let $r_{k,t}^s$ be the data rate associated with user k at time slot s in the DL or the UL indicated by $t \in \{UL, DL\}$. The objective associated with PF at time-slot s is [93, 57, 19]

$$\max_{\{r_{k,t}^s\}_{k=1}^K} \prod_{k=1}^K \left(r_{k,t}^s + \sum_{l < s} r_{k,t}^l \right) \iff \max_{\{r_{k,t}^s\}_{k=1}^K} \sum_{k=1}^K \log \left(1 + \frac{r_{k,t}^s}{\sum_{l < s} r_{k,t}^l} \right) + \sum_{k=1}^K \log \left(\sum_{l < s} r_{k,t}^l \right).$$

Then for $\sum_{l < s} r_{k,t}^l \gg r_{k,t}^s$, approximating $\log(1+x) \approx x$ and by ignoring $\sum_{k=1}^K \log(\sum_{l < s} r_{k,t}^l)$ (which is a constant for each slot), the weights in (3.11) are given by $\varepsilon_k^t = \frac{\Gamma_t}{\sum_{l < s} r_{k,t}^l}$ where Γ_t is a normalization constant chosen such that $\sum_{k=1}^K \varepsilon_k^t = 1$. Finally, we also consider a scenario where the weights are uncoupled from user channel quality and position. Specifically, we consider *independent* user weights that are randomly generated, uniformly and independently on $[0, 1]$, and then normalized. Here, the weights are randomly chosen independently of all user/network parameters and correspond to a scenario where the demand for downlink and uplink transmissions vary substantially across the users and the variance of the weights reflect the flexibility needed to accommodate the requests for transmissions.

3.4 Optimizing the Joint Design, and Alternate Schemes

To tackle the non-convex weighted sum optimization problem in (3.11), a block-coordinate descent (BCD) algorithm is developed. We iteratively leverage RCG, WMMSE and fractional programming (FP) in a BCD algorithm in order to optimize the WSP formulated in (3.11). The coupled optimization variables of the WSP in (3.11) are decomposed into blocks of decoupled sub-problems that are iteratively optimized.

Specifically, for a fixed RIS configuration Θ , the WSP in (3.11) decouples into separate downlink and uplink sub-problems. The WMMSE algorithm is used to optimize the transmit beamforming vectors $\{\mathbf{w}_k\}_{k=1}^K$ of the downlink sub-problem. For fixed Θ and uplink power $\{p_k\}_{k=1}^K$, an MMSE filter is used to optimize the receive beamforming vectors $\{\mathbf{v}_k\}_{k=1}^K$ and fractional programming (FP) is used to optimize the uplink power $\{p_k\}_{k=1}^K$ for fixed Θ and $\{\mathbf{v}_k\}_{k=1}^K$. Moreover, for fixed $\{\mathbf{w}_k\}_{k=1}^K$, $\{\mathbf{v}_k\}_{k=1}^K$ and $\{p_k\}_{k=1}^K$, RCG is used to update Θ . The algorithm proceeds by iterating over the decoupled blocks and updating the decision variables of a particular block while fixing the variables associated with the other blocks.

3.4.1 Updating the beamforming vectors and power control for fixed RIS configuration

For fixed Θ , the WSP in (3.11) decouples into a separate downlink sub-problem in (3.9) and uplink sub-problem in (3.10).

Updating transmit beamforming vectors $\{\mathbf{w}_k\}_{k=1}^K$

When Θ is fixed, optimizing the transmit beamforming vectors in (3.9) corresponds to a DL-WSR maximization problem. As in [98], by introducing the auxiliary variables $\{u_k\}_{k=1}^K$ and $\{t_k\}_{k=1}^K$ and with $\mathbf{u} = [u_1, \dots, u_K]$, $\mathbf{t} = [t_1, \dots, t_K]$, the DL-WSR problem in (3.9) is equivalent to the WMMSE problem written as

$$\begin{aligned} \min_{\mathbf{u}, \mathbf{t}, \mathbf{W}} \quad & \sum_{k=1}^K \varepsilon_k^{DL} (u_k e_k - \log_2 u_k) \\ \text{subject to} \quad & \sum_{k=1}^K \|\mathbf{w}_k\|^2 \leq P_{max}^{DL}, \end{aligned} \quad (3.12)$$

where the MSE associated with user k is $e_k = \mathbb{E}_{\mathbf{s}, n_k^{DL}} \{(s_k^{DL} - \hat{s}_k^{DL})(s_k^{DL} - \hat{s}_k^{DL})^H\}$ where $\mathbf{s} = [s_1^{DL}, \dots, s_K^{DL}]$ and the estimated DL symbol at user k is given by $\hat{s}_k^{DL} = t_k y_k^{DL}$.

The equivalent WMMSE problem lends itself to a tractable iterative algorithm as the problem in (3.12) is convex in the optimization variables when the other variables are fixed to the values of their previous iteration [98]. Proceeding similar to [98], the problem in (3.12) is optimized by iterating over the following update rules for each user $k \in \mathcal{K}$:

$$u_k = \frac{\sum_{i=1}^K |\mathbf{h}_{DL,k}^H \mathbf{w}_i|^2 + \sigma_{DL}^2}{\sum_{i=1, i \neq k}^K |\mathbf{h}_{DL,k}^H \mathbf{w}_i|^2 + \sigma_{DL}^2} \quad (3.13a)$$

$$t_k = \frac{\mathbf{h}_{DL,k}^H \mathbf{w}_k}{\sum_{i=1}^K |\mathbf{h}_{DL,k}^H \mathbf{w}_i|^2 + \sigma_{DL}^2} \quad (3.13b)$$

$$\mathbf{w}_k = \varepsilon_k^{DL} u_k t_k \left(\nu \mathbf{I}_M + \sum_{i=1}^K \varepsilon_i^{DL} u_i |t_i|^2 \mathbf{h}_{DL,i} \mathbf{h}_{DL,i}^H \right)^{-1} \mathbf{h}_{DL,k}, \quad (3.13c)$$

and $\nu \geq 0$ is the optimal Lagrange multiplier chosen such that complementary slackness associated with the sum-power constraint is achieved. By iterating over (3.13a), (3.13b) and (3.13c), Theorem 3 of [98] guarantees convergence to a stationary point of (3.12).

Updating receive beamforming vectors $\{\mathbf{v}_k\}_{k=1}^K$

When Θ and $\{p_k\}$ are fixed, the objective in the uplink sub-problem of (3.10), is maximized by the MMSE receiver. For each user $k \in \mathcal{K}$, the receive beamforming vector is given as

$$\mathbf{v}_k = \frac{\left(\sigma_{UL}^2 \mathbf{I}_M + \sum_{i=1}^K p_i \mathbf{h}_{UL,i} \mathbf{h}_{UL,i}^H\right)^{-1} \mathbf{h}_{UL,k}}{\left\| \left(\sigma_{UL}^2 \mathbf{I}_M + \sum_{i=1}^K p_i \mathbf{h}_{UL,i} \mathbf{h}_{UL,i}^H\right)^{-1} \mathbf{h}_{UL,k} \right\|}. \quad (3.14)$$

Updating uplink power control $\{p_k\}_{k=1}^K$

When Θ and $\{\mathbf{v}_k\}_{k=1}^K$ are fixed, fractional programming (see [94], [95]) is used to optimize $\{p_k\}_{k=1}^K$ in (3.10). The Lagrangian dual reformulation is applied by introducing the auxiliary variable $\gamma = [\gamma_1, \dots, \gamma_K]$. The power control problem of (3.10) is then equivalent to:

$$\begin{aligned} \max_{\gamma, \mathbf{P}} \quad & \sum_{k=1}^K \varepsilon_k^{UL} \ln(1 + \gamma_k) - \sum_{k=1}^K \varepsilon_k^{UL} \gamma_k + \sum_{k=1}^K \frac{\varepsilon_k^{UL} (1 + \gamma_k) |\mathbf{v}_k^H \mathbf{h}_{UL,k}|^2 p_k}{\sum_{i=1}^K |\mathbf{v}_k^H \mathbf{h}_{UL,i}|^2 p_i + \sigma_{UL}^2}, \\ \text{subject to} \quad & 0 \leq p_k \leq P_{max}^{UL}, \quad k = 1, \dots, K. \end{aligned} \quad (3.15)$$

For fixed $\{p_k\}_{k=1}^K$, setting the derivative of the objective in (3.15) with respect to γ_k to zero and solving for γ_k , yields γ_k^* for each $k \in \mathcal{K}$ which is given by

$$\gamma_k^* = \frac{|\mathbf{v}_k^H \mathbf{h}_{UL,k}|^2 p_k}{\sum_{i=1, i \neq k}^K |\mathbf{v}_k^H \mathbf{h}_{UL,i}|^2 p_i + \sigma_{UL}^2}. \quad (3.16)$$

By applying the quadratic transform and proceeding similar to [94], the power control update for each user k is obtained as

$$p_k^* = \min \left\{ P_{max}^{UL}, \frac{\chi_k^2 \varepsilon_k^{UL} (1 + \gamma_k) |\mathbf{v}_k^H \mathbf{h}_{UL,k}|^2}{\left(\sum_{i=1}^K \chi_i^2 |\mathbf{v}_i^H \mathbf{h}_{UL,k}|^2 \right)^2} \right\}, \quad (3.17)$$

where γ_k is obtained using (3.16) and the auxiliary variables $\{\chi_k\}$ are updated as:

$$\chi_k^* = \frac{\sqrt{\varepsilon_k^{UL} (1 + \gamma_k) |\mathbf{v}_k^H \mathbf{h}_{UL,k}|^2 p_k}}{\sum_{i=1}^K |\mathbf{v}_k^H \mathbf{h}_{UL,i}|^2 p_i + \sigma_{UL}^2}. \quad (3.18)$$

Iterating over the updates in (3.18), (3.16) and (3.17) yields a stationary point of the power control problem in (3.15) as established in [95].

3.4.2 Updating the RIS configuration for fixed beamforming vectors and uplink power control

For fixed $\{\mathbf{w}_k\}_{k=1}^K$, $\{\mathbf{v}_k\}_{k=1}^K$ and $\{p_k\}_{k=1}^K$, and ignoring terms that don't depend on $\boldsymbol{\theta}$, the WSP in (3.11) is equivalent to:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & J_{MO}(\boldsymbol{\theta}) = \alpha \beta \sum_{k=1}^K \varepsilon_k^{DL} R_{MO,k}^{DL}(\boldsymbol{\theta}) + (1 - \alpha) (1 - \beta) \sum_{k=1}^K \varepsilon_k^{UL} R_{MO,k}^{UL}(\boldsymbol{\theta}) \\ \text{subject to} \quad & |\theta_n| = 1, \quad n = 1, \dots, N \end{aligned} \quad (3.19)$$

Here for $t \in \{DL, UL\}$, $R_{MO,k}^t(\boldsymbol{\theta})$ is given as

$$R_{MO,k}^t(\boldsymbol{\theta}) = \log \left(1 + \frac{|\boldsymbol{\theta}^H \boldsymbol{\lambda}_{t,k,k} + \mu_{t,k,k}|^2}{\sum_{i=1, i \neq k}^K |\boldsymbol{\theta}^H \boldsymbol{\lambda}_{t,i,k} + \mu_{t,i,k}|^2 + \sigma_{k,t}^2} \right), \quad (3.20)$$

where

$$\boldsymbol{\lambda}_{DL,i,k} = \text{diag}(\mathbf{h}_{DL,r,k}^H) \mathbf{G}_{DL}^H \mathbf{w}_i, \quad \text{and} \quad \boldsymbol{\lambda}_{UL,i,k} = \sqrt{p_i} \text{diag}(\mathbf{h}_{UL,r,i}^H) \mathbf{G}_{UL}^H \mathbf{v}_k, \quad (3.21a)$$

$$\mu_{DL,i,k} = \mathbf{h}_{DL,d,k}^H \mathbf{w}_i, \quad \text{and} \quad \mu_{UL,i,k} = \sqrt{p_i} \mathbf{h}_{UL,d,i}^H \mathbf{v}_k. \quad (3.21b)$$

The $|\theta_n| = 1$ constraint of the optimization problem in (3.19) defines a search space characterized by a product of N complex circles, thus representing a Riemannian submanifold of $\mathbb{C}^N$. Moreover, the function $J_{MO}(\boldsymbol{\theta})$ is differentiable. Consequently, a stationary point of the problem in (3.19) can be obtained using the Riemannian conjugate gradient (RCG) algorithm [39, 121].

For any point $\boldsymbol{\theta}_j$ on a manifold $\mathcal{M}$, the tangent space denoted as $\mathcal{T}_{\boldsymbol{\theta}_j} \mathcal{M}$ comprises all the tangent vectors, each of which defines a search direction that can be used to optimize an objective function. The Riemannian gradient, denoted as $\text{grad}_{\boldsymbol{\theta}_j} J_{MO}$, represents the direction along which the objective function experiences the steepest increase and in the case of the complex circle manifold (CCM) of (3.19), is given by

$$\text{grad}_{\boldsymbol{\theta}_j} J_{MO} = \nabla J_{MO} - \mathcal{R}\{\nabla J_{MO} \circ \boldsymbol{\theta}_j^*\} \circ \boldsymbol{\theta}_j, \quad (3.22)$$

with the Euclidean gradient ∇J_{MO} given as

$$\nabla J_{MO} = \alpha \beta \sum_{k=1}^K 2\varepsilon_k^{DL} \boldsymbol{\Lambda}_{DL,k} + (1 - \alpha) (1 - \beta) \sum_{k=1}^K 2\varepsilon_k^{UL} \boldsymbol{\Lambda}_{UL,k}, \quad (3.23)$$

where for $t \in \{DL, UL\}$, $\Lambda_{t,k}$ is given by

$$\Lambda_{t,k} = \frac{\sum_{i=1}^K \lambda_{t,i,k} \lambda_{t,i,k}^H \boldsymbol{\theta} + \sum_{i=1}^K \lambda_{t,i,k} \mu_{t,i,k}^*}{\sum_{i=1}^K |\boldsymbol{\theta}^H \lambda_{t,i,k} + \mu_{t,i,k}|^2 + \sigma_{t,k}^2} - \frac{\sum_{i=1, i \neq k}^K \lambda_{t,i,k} \lambda_{t,i,k}^H \boldsymbol{\theta} + \sum_{i=1, i \neq k}^K \lambda_{t,i,k} \mu_{t,i,k}^*}{\sum_{i=1, i \neq k}^K |\boldsymbol{\theta}^H \lambda_{t,i,k} + \mu_{t,i,k}|^2 + \sigma_{t,k}^2}. \quad (3.24)$$

For each iteration j , the RCG algorithm involves iterating over three updates:

- 1- Performing retraction by projecting the search direction $\boldsymbol{\eta}_j$ to the CCM using

$$\boldsymbol{\theta}_{j+1} = \text{ret}_{CM}(\boldsymbol{\theta}_j + \alpha_j \boldsymbol{\eta}_j), \quad (3.25)$$

where α_j is an Armijo backtracking line search step size and for a vector $\mathbf{x} \in \mathbb{C}^N$ $\text{ret}_{CM}(\mathbf{x})$ is defined as

$$\text{ret}_{CM}(\mathbf{x}) = \left[\frac{x_1}{|x_1|}, \dots, \frac{x_N}{|x_N|} \right].$$

- 2- Finding the Riemannian Gradient $\text{grad}_{\boldsymbol{\theta}_{j+1}} J_{MO}$ according to (3.22)
- 3- Finding the search direction $\boldsymbol{\eta}_{j+1}$ by finding a tangent vector that is conjugate to $\text{grad}_{\boldsymbol{\theta}_{j+1}} J_{MO}$, given as

$$\boldsymbol{\eta}_{j+1} = -\text{grad}_{\boldsymbol{\theta}_{j+1}} J_{MO} + \tau_j \mathcal{T}_{j,j+1}(\boldsymbol{\eta}_j), \quad (3.26)$$

where τ_j is the Polak-Ribiere parameter obtained as in [121, 8] and the vector transport function $\mathcal{T}_{j,j+1}$ maps the tangent vector space at $\boldsymbol{\theta}_j$ to the tangent vector space at $\boldsymbol{\theta}_{j+1}$.

For the CCM of (3.19), $\mathcal{T}_{j,j+1}$ is given by

$$\mathcal{T}_{j,j+1}(\boldsymbol{\eta}_j) = \boldsymbol{\eta}_j - \mathcal{R}\{\boldsymbol{\eta}_j \circ \boldsymbol{\theta}_{j+1}^*\} \circ \boldsymbol{\theta}_{j+1} \quad (3.27)$$

The convergence of the RCG algorithm described above to a stationary point of the optimization problem in (3.19) is established in [8].

Algorithm 1 summarizes the BCD algorithm used to obtain the joint uplink-downlink design.

Algorithm 1 BCD algorithm

- 1: Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, $\{\mathbf{v}_k\}_{k=1}^K$, $\{p_k\}_{k=1}^K$, $\{\mathbf{w}_k\}_{k=1}^K$ and iteration $j = 1$
 - 2: **repeat**
 - 3: Update power $\{p_k\}_{k=1}^K$ using iterative algorithm (3.16), (3.17) and (3.18).
 - 4: Update $\{\mathbf{v}_k\}_{k=1}^K$ using (3.14).
 - 5: Update $\{\mathbf{w}_k\}_{k=1}^K$ using iterative algorithm in (3.13a), (3.13b) and (3.13c).
 - 6: Update $\boldsymbol{\theta}_j$ using RCG algorithm in (3.25), (3.26) and (3.27).
 - 7: **until** $\|J_{WSP,j} - J_{WSP,j-1}\| < \epsilon$, $\epsilon > 0$ is a threshold
-

3.4.3 Structure of beamforming vectors and channel reciprocity in RIS systems

The joint RIS problem, optimized using the developed BCD algorithm, involves configuring the RIS phase shifts based on the downlink transmit beamforming and the uplink receive beamforming. Consequently, a joint RIS design would benefit from any similarity in the structure of the uplink and downlink beamforming vectors. Indeed, the similarity of the structure can be seen from the uplink beamforming vectors in (3.14) and the downlink beamforming vectors in (3.13), as both beamforming vectors arise due to an MMSE filter with different weights and powers (see [13]).

In TDD, channel reciprocity holds for the direct link between the BS and the users as these links are independent of the RIS and corresponds to a classical (non-RIS) system. Recently, experimental results in [85] and [102] have confirmed channel reciprocity for the cascaded link through the RIS. Hence, for TDD operation, the joint RIS design would also benefit from channel reciprocity and thus any similarity in the uplink and downlink beamforming vectors. Subsequently, the performance gap due to a joint design compared to individual design may be narrower for TDD compared to FDD. Indeed, as will be shown in Section 3.5.2, for equal and proportional-fair weights, the joint design almost achieves the individual design bound in TDD, but not in FDD.

3.4.4 Alternate Schemes

For an RIS with N elements, the spectral efficiency performance of the joint design in TDD/FDD systems is compared with optimizing the RIS for UL and DL transmissions individually and then the two RIS configurations are time-shared, which we refer to as

Individual design. Such a scheme can be implemented in TDD but at the cost of increased overhead and configuration periods as explained before. However, this scheme is unrealistic for FDD communications as both transmissions occur simultaneously.

Moreover, the performance of the joint design is compared to two other designs called *fixed-downlink* and *fixed-uplink* design. In fixed-downlink design, the weighted sum rate (WSR) in (3.9) is first obtained by jointly optimizing the RIS configuration Θ and the transmit beamforming vectors $\{\mathbf{w}_k\}_{k=1}^K$ using WMMSE and RCG. Then, the receive combining vectors $\{\mathbf{v}_k\}_{k=1}^K$ are found from (3.14) using the effective channel configured by the RIS phase shifts in (3.1b). In fixed-uplink design, the RIS configuration Θ and the receive combining vectors $\{\mathbf{v}_k\}_{k=1}^K$ are first jointly optimized according to (3.10) using MMSE, FP and RCG. With this RIS configured for uplink transmission, the transmit beamforming vectors $\{\mathbf{w}_k\}_{k=1}^K$ are then found by applying a downlink beamforming strategy for the effective channel in (3.1a).

In addition, the performance of the joint design is also compared to two slicing benchmarks, namely: *slicing-with-interference* and *slicing-without-interference*. An N -element RIS is sliced by optimizing $\frac{N}{2}$ RIS elements to exclusively assist each of the UL and DL transmissions. The slicing-without-interference benchmark assumes that the downlink (respectively, uplink) transmissions don't affect the RIS elements dedicated to assist the uplink (respectively, downlink) transmissions. This is idealistic as the uplink and downlink signals are reflected by all of the RIS elements and hence the phase shift associated with each element affects both uplink and downlink transmissions. Conversely, the slicing-with-interference benchmark then computes the achievable DL and UL rates using the entire RIS configuration. This in turn incorporates the interference terms due to slicing the RIS.

3.5 Simulation Results

3.5.1 System setup and parameters

The RIS-assisted communication system is comprised of an M -antenna BS, an N -element RIS, and K single-antenna users placed in a 3D coordinate system. Following Fig. 3.2b, the BS is located at $(x_{BS}, y_{BS}, h_{BS}) = (0, 25, 25)$ m and the RIS is placed at $(x_{RIS}, y_{RIS}, h_{RIS}) = (300, 0, 15)$ m. Moreover, the K users are placed at a height of $h_c = 1.5$ m while being distributed within an $x - y$ plane disc which is centered at $(x_c, y_c) = (300, 25)$ m with radius $R = 20$ m.

A uniform rectangular array (URA) is used at the BS with M elements such that $M_y \times M_z = M$ where M_y and M_z correspond to the number of elements aligned with

the y -axis and z -axis. Similarly, a URA is used at the RIS with N elements such that $N_x \times N_z = N$ where N_x and N_z correspond to the number of elements aligned with the x -axis and z -axis.

For TDD, the uplink/downlink carrier frequency is 1.95 GHz. In FDD, the uplink carrier frequency is $f_{UL} = 1.95$ GHz and the downlink carrier frequency is $f_{DL} = 2.14$ GHz. In both TDD and FDD, the antennas at the BS and the RIS are spaced by $\frac{1}{2} \times \frac{c}{f_{UL}}$ where c is the speed of light in free space. Let $P_1 = 30$ dBm, $P_2 = 17$ dBm. In TDD, $P_{max}^{DL} = P_1$ and $P_{max}^{UL} = P_2$ and in FDD, $P_{max}^{DL} = \alpha P_1$ and $P_{max}^{UL} = (1 - \alpha)P_2$. Moreover, $N_0 = N_i \times NF$ where $N_i = -170$ dBm/Hz and the noise figure in downlink is $NF = 9$ dB and in uplink $NF = 7$ dB. For TDD, the bandwidth is 20 MHz. In FDD, the total bandwidth is 20 MHz, shared between uplink and downlink transmissions.

3GPP path loss models are used [1] with the line-of-sight (LOS) and non-line-of-sight (NLOS) models are given by

$$PL^{LOS}(d_{3d}, f_c) = 28 + 22 \log_{10}(d_{3D}) + 20 \log_{10}(f_c), \quad (3.28a)$$

$$PL^{NLOS}(d_{3d}, f_c) = 13.54 + 39.08 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) - 0.6(h_c - 1.5), \quad (3.28b)$$

where d_{3d} is the 3D distance between two points and f_c is the carrier frequency. For each user $k \in \mathcal{K}$, the RIS-user reflected channel $\mathbf{h}_{t,r,k}$ and the BS-RIS channel $\mathbf{G}_t$ are modeled using a Rician fading model for $t \in \{UL, DL\}$. In particular,

$$\mathbf{G}_t = \sqrt{PL^{LOS}(d_{BS-RIS}, f_t)} \left(\sqrt{\frac{\kappa}{1 + \kappa}} \mathbf{G}_t^{LOS} + \sqrt{\frac{1}{1 + \kappa}} \mathbf{G}_t^{NLOS} \right), \quad (3.29a)$$

$$\mathbf{h}_{t,r,k} = \sqrt{PL^{LOS}(d_{RIS-k}, f_t)} \left(\sqrt{\frac{\kappa_k}{1 + \kappa_k}} \mathbf{h}_{t,r,k}^{LOS} + \sqrt{\frac{1}{1 + \kappa_k}} \mathbf{h}_{t,r,k}^{NLOS} \right), \quad (3.29b)$$

where $PL^{LOS}(d_{3d}, f_c)$ is given in (3.28a). Also, d_{BS-RIS} denotes the 3D distance between the BS and the RIS and d_{BS-k} denotes the 3D distance between the RIS and user k . The Rician factors κ and $\{\kappa_k\}_{k=1}^K$ capture the relative strength of the LOS component compared to the NLOS components between the BS and the RIS and between the RIS and user k , respectively. The matrix $\mathbf{G}_t^{LOS} \in \mathbb{C}^{M \times N}$ contains the phases for the LOS components which were computed using the underlying geometry. In particular, for the m^{th} element of the BS and the n^{th} element of the RIS, separated by 3D distance of $d_{3d}^{m,n}$, the corresponding phase shift is given by $[\mathbf{G}_t^{LOS}]_{m,n} = \exp\left(\frac{2\pi j d_{3d}^{m,n}}{\lambda_t}\right)$ where λ_t is the wavelength associated with the corresponding carrier frequency. Similarly, the vector $\mathbf{h}_{t,r,k}^{LOS}$ captures the N phase shifts associated with LOS component between user k and the RIS. On the other hand, $\mathbf{G}_t^{NLOS}$ and $\mathbf{h}_{t,r,k}^{NLOS}$ capture the NLOS components associated with BS-RIS and RIS-user

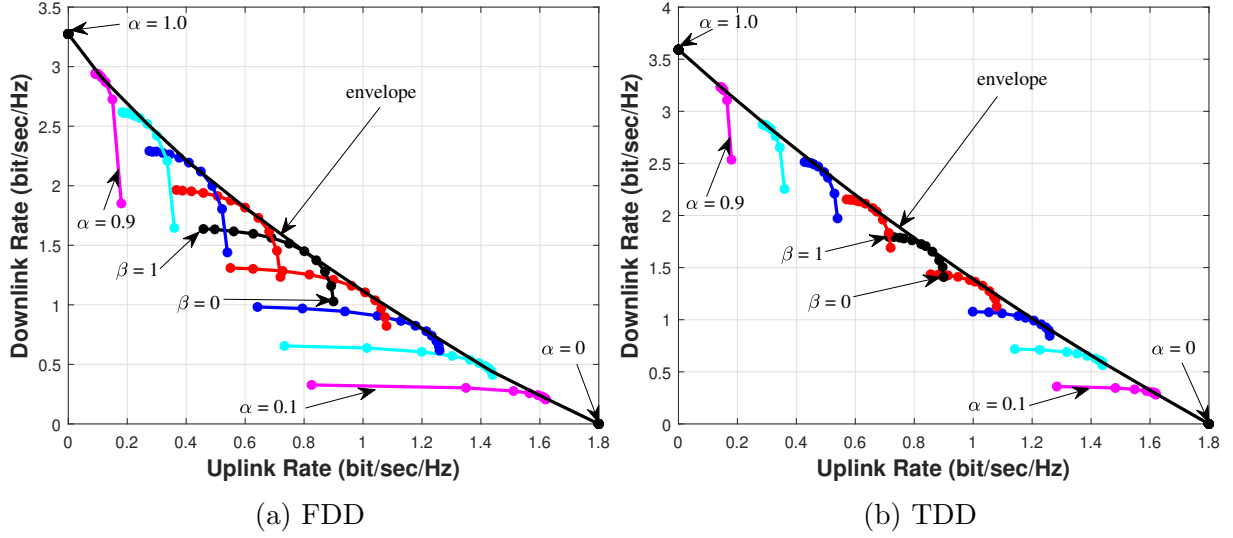


Figure 3.3: Uplink-Downlink trade-off region for $\alpha = 0, 0.1, \dots, 1$, where for a fixed α , $\beta = 0, 0.1, \dots, 1$, with $N = 200$ and under independent weights.

k propagation channels, respectively. Moreover, the Rayleigh fading components $\mathbf{G}_t^{NLOS}$ and $\mathbf{h}_{t,r,k}^{NLOS}$ are sampled with each entry i.i.d from $\mathcal{CN}(0, 1)$.

For each user k , the BS-user channel $\mathbf{h}_{t,d,k}$ is modeled using Rayleigh fading. Particularly,

$$\mathbf{h}_{t,d,k} = \sqrt{PL^{NLOS}} \mathbf{h}_{t,d,k}^{NLOS}, \quad (3.30)$$

where PL^{NLOS} is given in (3.28b) and the components of $\mathbf{h}_{t,d,k}^{NLOS}$ are sampled i.i.d. from $\mathcal{CN}(0, 1)$. Unless otherwise specified, the Rician factors are $\kappa = 6$ dB and $\{\kappa_k\}_{k=1}^K = 8$ dB, and the number of antennas at the BS and RIS are $M = 4 \times 2 = 8$ and $N = 20 \times 10 = 200$ serving $K = 4$ users. The numerical results were obtained by averaging the results over 100 independent channel realizations.

3.5.2 Uplink-Downlink Trade-off Regions

For a fixed time/frequency resource allocation defined by a fixed α , the relative priority of uplink vs downlink is defined by the weight β . In order to first explore the rate region for a fixed α , the problem in (3.11) is optimized using the BCD algorithm for different values of β from 0 to 1 under the independent weighting strategy. Figs. 3.3a and 3.3b show the trade-off between average uplink weighted rates, given by (3.10a) and (3.8b), versus average downlink weighted rates, given by (3.9a) and (3.8a), for $\alpha = 0, 0.1, \dots, 1$ and both FDD and TDD systems.

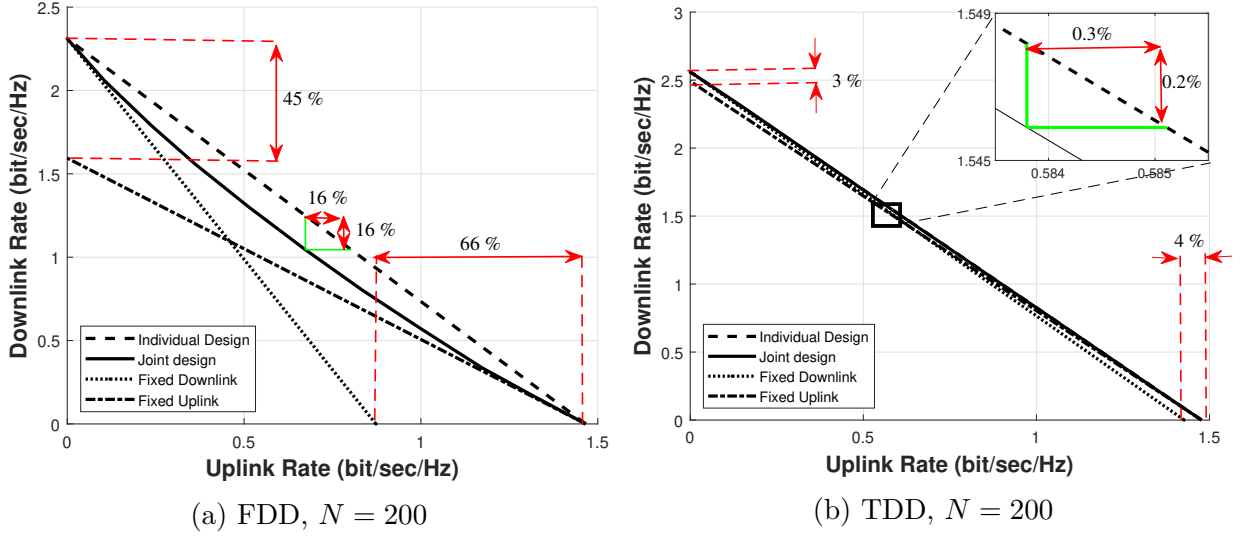


Figure 3.4: Uplink-Downlink Rate region for $N = 200$ and equal weights.

By optimizing the fraction of time or band spent on uplink vs downlink, i.e. optimizing α , the trade-off between UL-rate and DL-rate achieved by the BCD algorithm is found and is given by the envelope of the curves in Figs. 3.3a and 3.3b, and shown in solid black. This envelope represents the trade-off between UL weighted rates and DL weighted rates under the joint design and UL/DL resource optimization.

The performance of the joint design envelope with respect to the individual design and the fixed-downlink/fixed-uplink schemes are shown in Figs. 3.4-3.6 for the three user weighting schemes. Four metrics are used to compare performance: the maximum DL gain of joint design over fixed-uplink design, the maximum UL gain of joint design over fixed-downlink design, and both the maximum DL loss and the maximum UL loss of joint design compared to individual design. The two maximum gain metrics correspond to the maximum improvement yielded by employing the joint design vs employing a fixed-downlink/fixed-uplink design. Moreover, the loss of performance due to joint design compared with the individual design is quantified using the maximum DL/UL losses.

Figs. 3.4a and 3.4b show the trade-off curves with equal weighting strategies for FDD and TDD systems with $N = 200$ along with the 4 performance metrics described above. Figs. 3.5a and 3.5b show the trade-off curves associated with PF weighting under FDD and TDD systems, where the PF weights were obtained using the approximation in Section 3.3.2 for a large number of time slots $s = 100$. Similarly, Figs. 3.6a and 3.6b show the trade-off curves with independent user weights, again for $N = 200$. From the figures, the

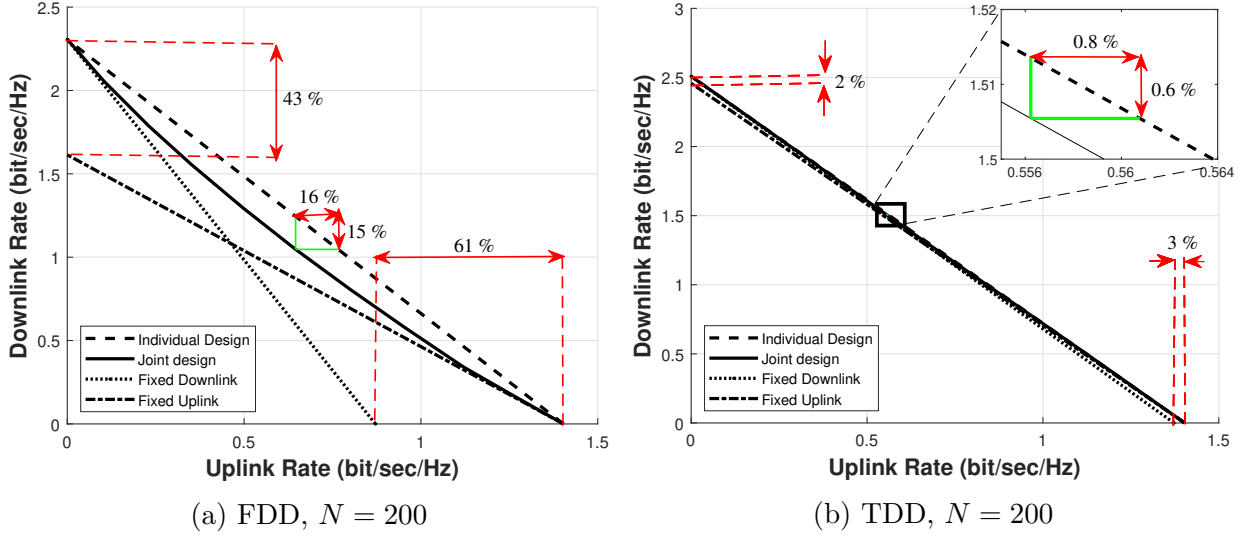


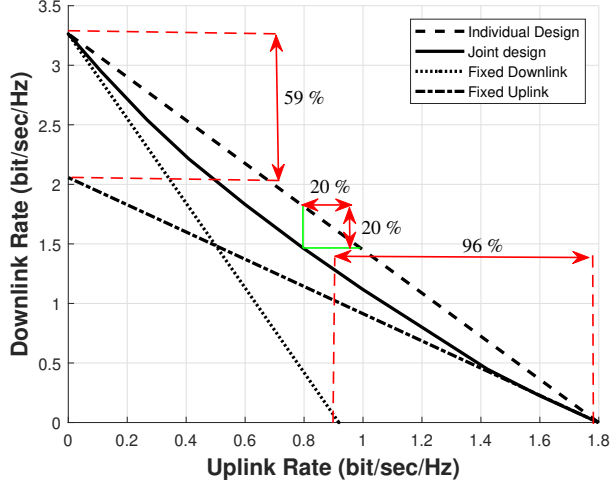
Figure 3.5: Uplink-Downlink Rate region for $N = 200$ and PF weights.

improvement due to joint design compared with the fixed-downlink/uplink designs is more significant as we deviate towards independent weights and as more flexibility in resource allocation is needed.

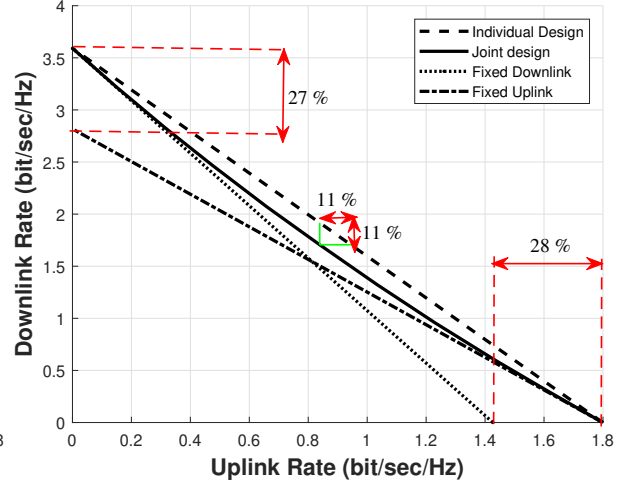
Indeed, as shown in Fig. 3.6, the maximum UL gain due to joint design is 96% and the maximum DL gain is 59% when independent weights are used under the FDD scenario. This is compared to 61% and 43% when PF weights are used (Fig. 3.5a) and compared to 66% and 45% when equal weights are used (Fig. 3.4a). The same pattern applies for TDD where the maximum UL and DL gains are 28% and 27% when independent weights are used. This is in contrast with max-UL and max-DL gains of 3% and 2% when PF weights are used and 4% and 3% when equal weight are used as shown in Figs. 3.5b and 3.4b, respectively.

As outlined in Section 3.4.3, TDD benefits from channel reciprocity and hence the max-UL and max-DL losses, of joint over individual design, are expected to be lower for TDD compared to FDD. Under TDD, the max-UL loss is 0.3%, 0.8%, and 11% for equal, PF, and independent weights, respectively. This is compared to 16%, 16%, and 20% for FDD. Similarly, for TDD the max-DL loss is 0.2%, 0.6% and 11% for equal, PF, and independent weights, respectively. This is in contrast with the higher losses in FDD of 16%, 15%, and 20%.

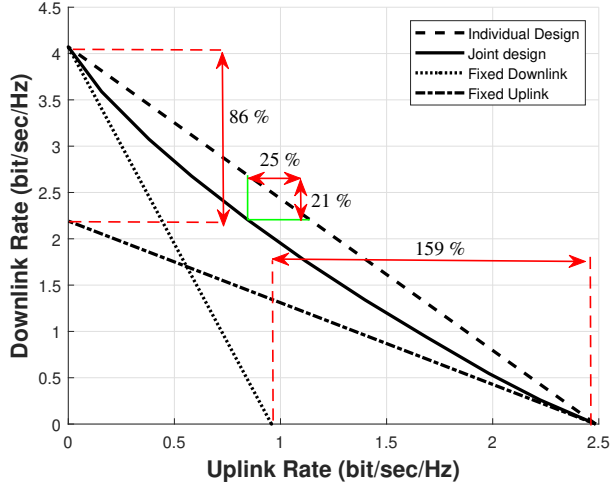
As shown in Figs. 3.6c and 3.6d compared to Figs. 3.6a and 3.6b, the joint design provides greater benefits for larger N . The max-DL and max-UL gains increase from 59%



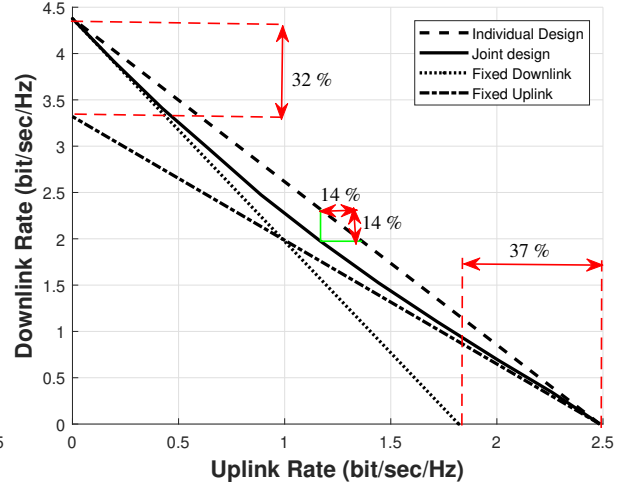
(a) FDD, $N = 200$



(b) TDD, $N = 200$



(c) FDD, $N = 400$



(d) TDD, $N = 400$

Figure 3.6: Uplink-Downlink region for $N = 200$ and $N = 400$, under independent user weights.

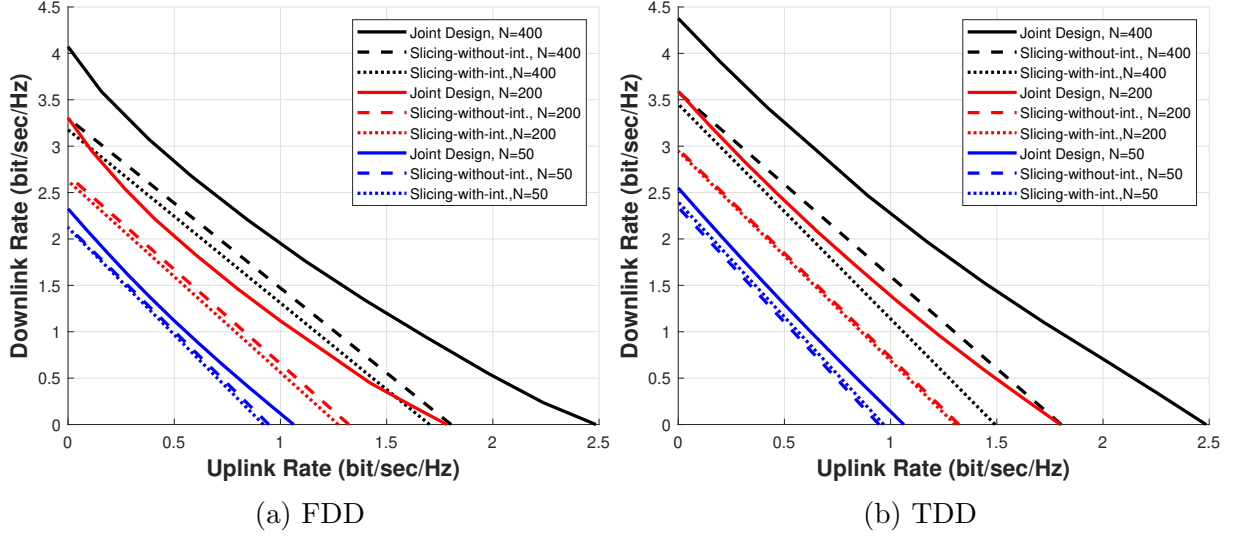


Figure 3.7: Uplink-Downlink Rate region achieved by joint design and slicing benchmarks for different N and independent weights.

and 96% for $N = 200$ to 86% and 159% for $N = 400$ in the case of FDD configuration. We attribute this to the fact that as N increases, performance is more dependent on the RIS. Hence, a joint RIS design with optimized phase shifts becomes more beneficial compared to fixed-downlink/fixed-uplink designs.

Fig. 3.7 shows, for independent weights, the uplink-downlink trade-off regions achieved by the joint design and the two slicing benchmarks (slicing-with-interference and slicing-without-interference) for different values of N under both FDD and TDD. The figure reveals that for both FDD and TDD, the joint design outperforms the slicing-with-interference and the slicing-without-interference benchmarks. This follows since for all values of N in Fig. 3.7, the UL-DL region achieved by the joint design strictly contains the UL-DL regions associated with the two slicing benchmarks. Furthermore, since slicing-without-interference underperforms joint design, Fig. 3.7 also demonstrates that the interference alone is not responsible for the inferior performance achieved by a sliced RIS.

Fig. 3.8 shows, for independent weights, the uplink-downlink trade-off regions achieved by the joint design, individual design as well as fixed designs, all obtained using the BCD algorithm 1. Moreover, Fig. 3.8 also shows (in blue) the uplink-downlink trade-off region obtained by employing an individual design where the uplink and downlink problems are optimized using Couenne numerical solver ([11]) with a time limit of 2.5 hours per channel realization. The figure demonstrates the efficacy of the proposed BCD algorithm 1 in

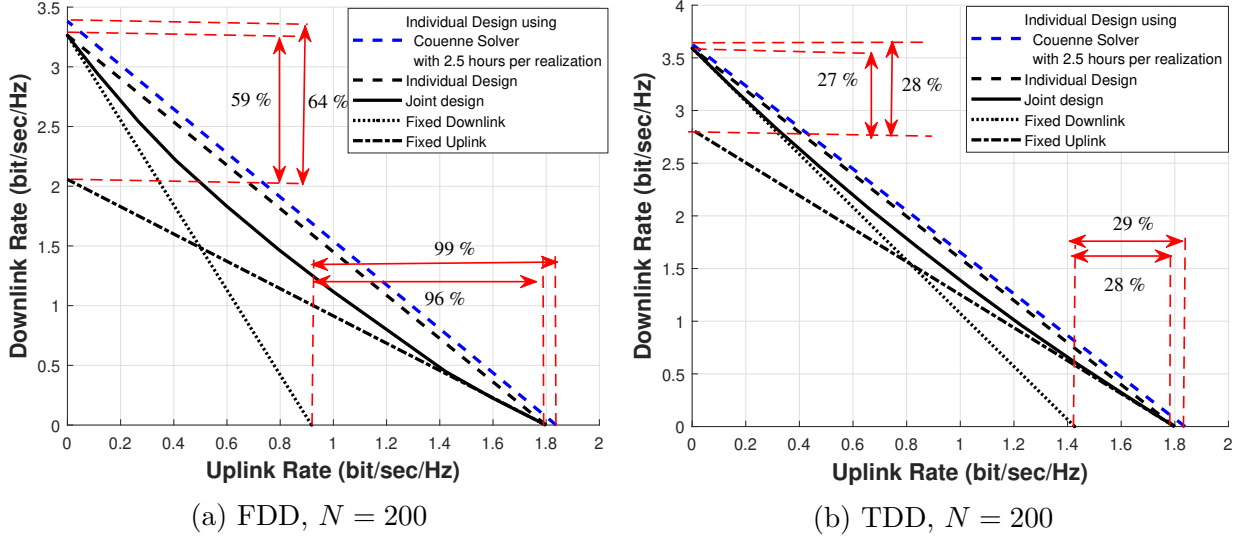


Figure 3.8: Uplink-Downlink Rate region achieved by joint design and alternative schemes obtained using BCD algorithm 1 as well as individual design obtained using Couenne numerical solver. Here, $N = 200$ and independent weights are employed.

efficiently obtaining uplink-downlink trade-off region. For instance, in FDD, the maximum DL gain due to joint design (using the BCD algorithm) over fixed designs is 59% and the maximum DL gain due to individual design (using numerical solver) over fixed designs is 64%. Subsequently, the use of a time-intensive numerical solver (with up to 2.5 hours per realization) achieves at most a 5% gain in terms of maximum DL or UL rates improvement over a fixed design, compared to the BCD algorithm.

Fig. 3.9 shows the impact of the number of antennas at the BS on the max-UL and max-DL gains of joint design over fixed-uplink/fixed-downlink designs as well as the max-UL and max-DL losses of joint design over the individual design, under independent weights. From the figure, as the ratio of passive beamforming elements to active beamforming elements is increased, the performance of the system becomes more dependent on the passive beamforming at the RIS and hence the performance improvement due to joint RIS design is more substantial. Specifically, Figs. 3.9a and 3.9b show the uplink-downlink trade-off regions for $M = 4$ and $M = 16$ under FDD with an RIS of $N = 200$ elements. Comparing Figs. 3.9a and 3.9b with the case of $M = 8$ in Fig. 3.6a, the max-DL gain increases from 40% to 59% to 84% as the antennas at the BS decrease from $M = 16$ to $M = 8$ to $M = 4$. In a similar manner, the max-UL gain increases from 63% to 96% to 142% as the antennas at the BS decrease from $M = 16$ to $M = 8$ to $M = 4$. The same

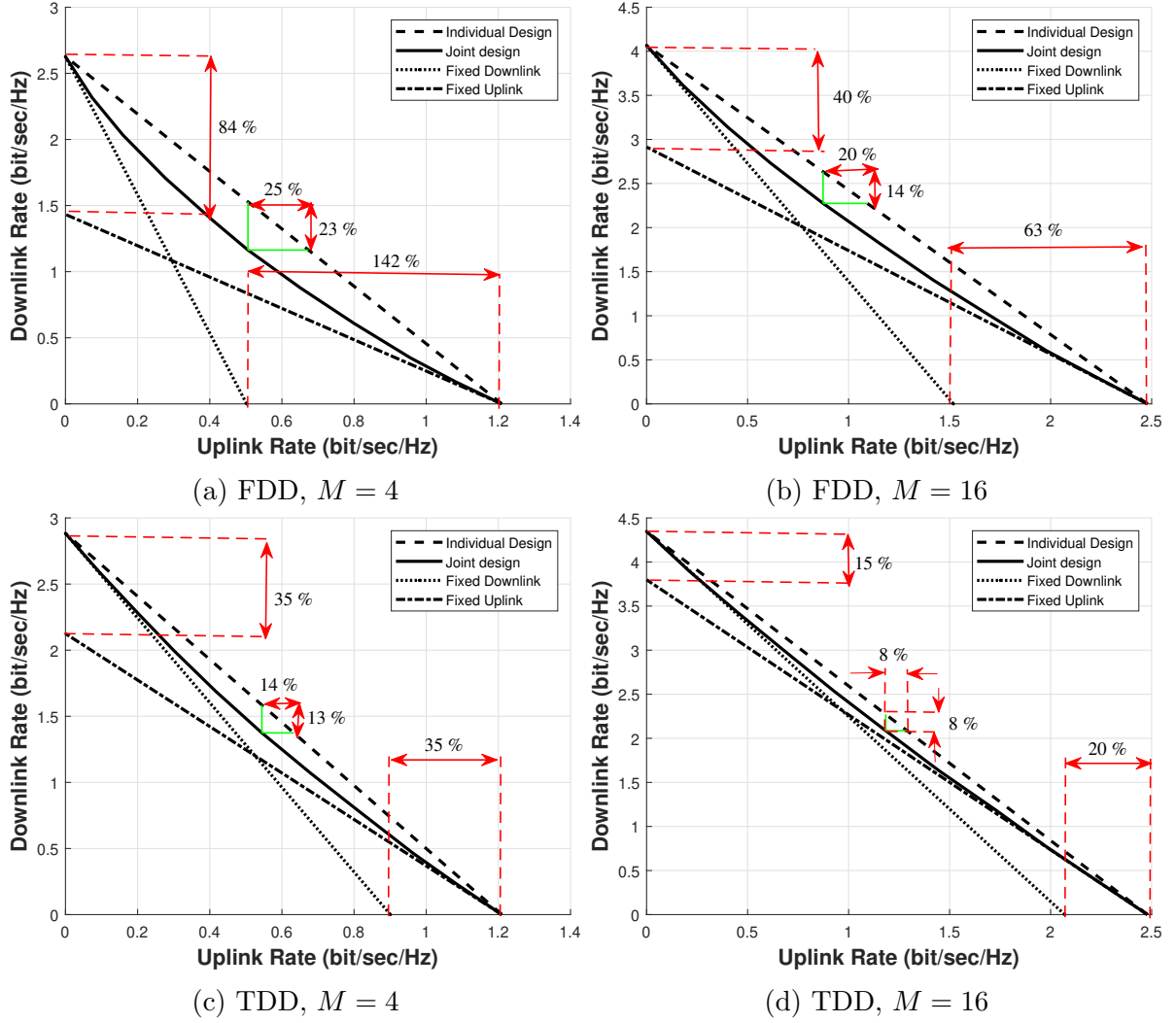


Figure 3.9: Uplink-Downlink Trade-off region for $N = 200$ under FDD/TDD for different M with independent weights.

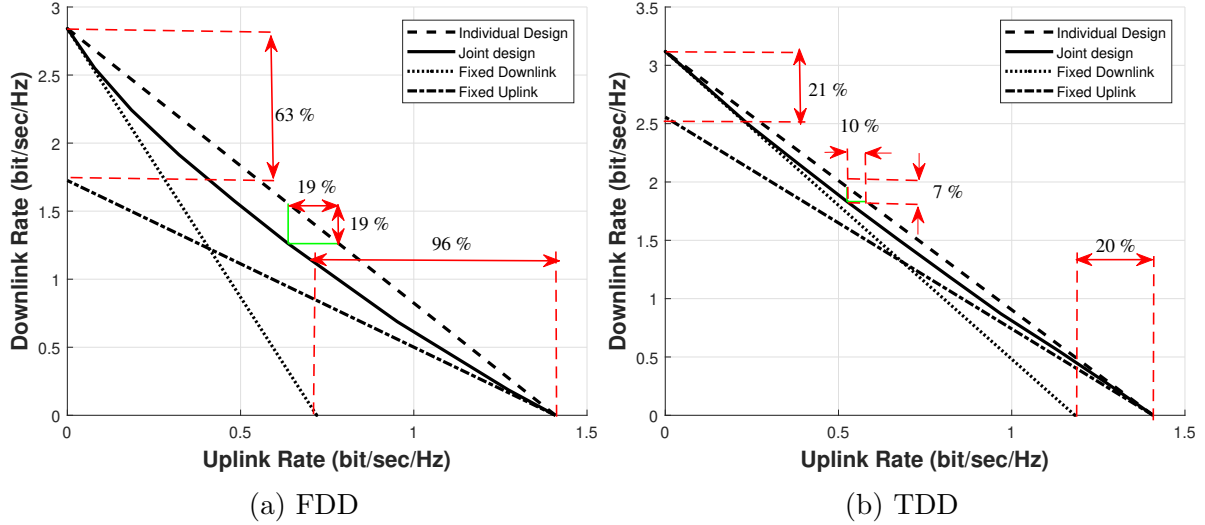


Figure 3.10: Uplink-Downlink Rate region for $N = 200$ under ZF beamforming and independent weights.

behavior was observed in TDD as demonstrated by Figs. 3.9c and 3.9d compared with Fig. 3.6b as the max-UL gain increases from 20% to 28% to 35% and the max-DL gain increases from 20% to 27% to 35% as the antennas at the BS decrease.

Zero forcing (ZF) beamforming/combining can also be used as a heuristic in order to obtain beamforming/combining vectors with less complexity [13]. Fig. 3.10 shows the UL-DL trade-off regions when ZF beamforming/combining is used for FDD and TDD with $N = 200$. For FDD, compared to the results of Fig. 3.6a, the rate regions shrink when ZF is used as the maximum achievable downlink rate is reduced from 3.26 bit/sec/Hz to 2.84 bit/sec/Hz and thus WMMSE yields an improvement of 15% over ZF. Similarly, MMSE yields an improvement of 28% in terms of the maximum achievable uplink rate. The max-DL gain increases slightly from 59% under WMMSE to 63% for ZF beamforming. The max-UL gain of 96% is the same under MMSE and ZF combining. Hence, for ZF, the improvement in performance due to joint design compared with the uplink/downlink fixed design is similar to MMSE/WMMSE. Similarly for TDD, the rate regions shrink when ZF is used as the maximum achievable downlink rate is reduced from 3.59 bit/sec/Hz to 3.12 bit/sec/Hz and thus WMMSE yields an improvement of 16% over ZF. The maximum achievable uplink rate shows an improvement of 28% when MMSE is adopted compared to ZF. Furthermore, under ZF the max-UL and max-DL gains are similar to those when MMSE/WMMSE is adopted. Consequently, while ZF has lower complexity and performance compared to MMSE/WMMSE, the relative performance gains of joint-design are

about the same as MMSE/WMMSE.

3.5.3 Complexity Analysis

In this section, we present the computational complexity associated with the BCD algorithm of Section 3.4. Moreover, we compare the overhead and computational complexity associated with the BCD optimized joint RIS design and the fixed uplink/downlink heuristic designs.

Computational Complexity

The BCD algorithm in Section 3.4 iteratively updates the downlink beamforming vectors, uplink combining vectors, power control and RIS configuration blocks. We present the per-iteration complexity associated with updating the BCD blocks.

Let M be the number of antennas at the BS, N be the number of RIS elements and K be the number of users. To update the BCD blocks, the effective downlink and uplink channels associated with user k need to be computed as given in (3.1a) and (3.1b). For $t \in \{UL, DL\}$, computing the k^{th} effective downlink/uplink channel involves N multiplication operations to multiply the diagonal matrix Θ_t by the reflected channel $\mathbf{h}_{t,r,k}$ and another MN multiplication operations to multiply the aforementioned product by $\mathbf{G}_t$. Moreover, $M(N-1)$ and M addition operations are needed to compute the expression in (3.1a) or (3.1b). Subsequently, $2MN + N$ operations are needed to obtain the effective channel for each user k and hence the complexity of computing the K effective channels is $\mathcal{O}(KMN)$.

For given effective downlink channels $\{\mathbf{h}_{DL,k}\}_{k=1}^K$, obtained with complexity $\mathcal{O}(KMN)$, the downlink beamforming vectors $\{\mathbf{w}_k\}_{k=1}^K$ can be updated either using the WMMSE algorithm or zero-forcing (ZF). For given effective channels, each (inner) iteration of the WMMSE algorithm, presented in (3.13), is dominated by the matrix inversions and computing the products $\mathbf{h}_{DL,k}^T \mathbf{w}_i$ as well as the search for the optimal Lagrange multiplier ν in each iteration of WMMSE. Hence, the *per BCD-iteration* complexity is $\mathcal{O}(I_B K^2 M + I_B I_\nu K M^2 + I_B I_\nu M^3)$, where I_B is number of (inner) iterations needed for updating the WMMSE algorithm and I_ν is the number of iterations to search for ν . Consequently, when WMMSE is adopted, the complexity of computing the effective channels and updating the downlink beamforming vectors is $\mathcal{O}(C_{DL})$ where $C_{DL} = I_B K^2 M + I_B I_\nu K M^2 + I_B I_\nu M^3 + KMN$. Alternatively, when ZF is used to update $\{\mathbf{w}_k\}_{k=1}^K$, *one* update is needed and is dominated by computing an $M \times M$ matrix inversion and K matrix multiplications of the inverted matrix by an M -length vector. Consequently, when ZF is adopted, the complex-

ity of computing the effective channels and updating the downlink beamforming vectors is $\mathcal{O}(C_{DL})$ where $C_{DL} = M^3 + KM^2 + KMN$.

The (same) effective uplink channels $\{\mathbf{h}_{UL,k}\}_{k=1}^K$, obtained with complexity $\mathcal{O}(KMN)$, are needed to update *both* the uplink combining vectors $\{\mathbf{v}_k\}_{k=1}^K$ and power control $\{p_k\}_{k=1}^K$. For given effective uplink channels, updating the K combining vectors using the MMSE receiver incurs a complexity of $\mathcal{O}(M^3 + KM^2)$. This follows since computing the K combining vectors in (3.14) is dominated by an $M \times M$ matrix inversion and K multiplications of the inverted matrix by an M -length vector. Moreover, for given effective channels, the fractional programming (FP) algorithm in (3.18), (3.16) and (3.17) is used to update the uplink power $\{p_k\}_{k=1}^K$. This incurs a complexity of $\mathcal{O}(K^2M + I_F K^2)$, where I_F is the number of iterations needed to update the power control. Subsequently, the complexity of computing the effective channels and updating the uplink combining vectors and power control is $\mathcal{O}(C_{UL})$ where $C_{UL} = M^3 + KM^2 + K^2M + I_F K^2 + KMN$.

To update the RIS configuration, the $\lambda_{t,i,k}$ and $\mu_{t,i,k}$ expressions in (3.21a) and (3.21b) need to be computed for $t \in \{UL, DL\}$ and for $i, k \in \mathcal{K}$. This incurs a complexity of $\mathcal{O}(PKMN)$, where $P = 2K$ corresponds to the K (downlink) beamforming vectors and K (uplink) combining vectors. When the expressions in (3.21a) and (3.21b) are given, each (inner) iteration of the RCG is dominated by computing the objective in the line search and hence incurs a complexity of $\mathcal{O}(I_L I_R PKN)$, where I_R is the number of (inner) iterations associated with the RCG algorithm and I_L is the number of line search iterations. Consequently, the complexity of computing the necessary expressions in (3.21a) and (3.21b) and updating the RIS configuration is $\mathcal{O}(C_{RCG})$ where $C_{RCG} = I_L I_R PKN + PKMN$.

For the scenario illustrated in Fig. 3.3, the BCD algorithm required an average runtime of 9 ms per realization to optimize the joint UL-DL design for a given pair of α and β . This result was obtained using a MATLAB implementation executed on an AMD Ryzen 7 5800X 8-core processor at 3.80 GHz.

Complexity Comparisons

We first compare the three proposed designs as well as the individual design in terms of feedback overhead and configuration periods. In FDD, the uplink and downlink transmissions occur concurrently and hence an individual design is not achievable. In TDD, an individual RIS design would constitute sending two sets of configurations to the RIS, one for downlink and one for uplink. Conversely, the optimized joint design as well as the fixed-uplink and fixed-downlink design involve sending *one* RIS configuration. Thus, these designs halve the feedback overhead compared to the individual design. In practice, the

individual design would also involve switching the RIS configurations between downlink and uplink and hence the joint design and fixed-uplink/fixed-downlink designs also provide some reductions in power consumption and the duration of the period where the RIS system can not be used for communication.

From the computational complexity analysis, RCG for fixed-uplink and fixed-downlink requires only $P = K$ terms in computing the objective, reducing the computational complexity of this dominant term by a factor of 2 compared to a joint design. In addition, the fixed-downlink (respectively, fixed-uplink) design iteratively optimizes the RIS configuration using RCG and WMMSE (respectively, MMSE) algorithms. Then for a fixed RIS configuration, the uplink combining vectors (respectively, downlink beamforming vectors) are computed by applying the MMSE (respectively, WMMSE) *once* to the effective uplink (respectively, downlink) channels. Hence, compared to an optimized joint design, the complexity of a fixed-downlink design is reduced by $I_O - 1$ iterations of complexity $\mathcal{O}(C_{UL})$, where I_O is the number of outer-loop iterations and C_{UL} indicates the number of operations needed to update the uplink combining vectors and power control. Similarly, compared to an optimized joint design, the complexity of a fixed-uplink design is reduced by $I_O - 1$ iterations of complexity $\mathcal{O}(C_{DL})$, where C_{DL} indicates the number of operations needed to update the downlink beamforming vectors.

Furthermore, compared to an individual design, the complexity of a fixed-downlink design is reduced by $I_O - 1$ iterations of complexity $\mathcal{O}(C_{UL})$ and the operations needed to optimize the uplink RIS configuration are eliminated. Similarly, a fixed-uplink design incurs a lower complexity compared to an individual design by eliminating the need for optimizing the downlink RIS configuration and reducing the complexity of beamforming design by $I_O - 1$ iterations of complexity $\mathcal{O}(C_{DL})$.

The aforementioned complexity comparisons, when combined with the trade-off regions presented in Section 3.5.2, demonstrate the flexibility offered by the proposed joint design as well as fixed-uplink and fixed-downlink designs in trading off spectral efficiency performance for reductions in computational complexity, overhead and configuration periods. These trade-offs may be particularly appealing when the reductions in overhead and computational complexity can be achieved in exchange for marginal losses in spectral efficiency. For instance, as demonstrated in Section 3.5.2, the fixed-uplink and fixed-downlink designs perform almost as well as the individual design in TDD scenarios under equal and proportional fair user-weighting strategies (but not independent weights). Thus, under these scenarios, the fixed designs almost achieve the performance of the individual design but with lower complexity than the BCD optimized joint design and individual design, and lower overhead/power consumption than the individual design.

On the other hand, in the FDD scenarios considered in Section 3.5.2, the proposed joint RIS design performs substantially better than the fixed-uplink and fixed-downlink heuristic designs and diminishes the gap to the individual design. In TDD, provided that the user weights are independent, the joint design substantially improves the spectral efficiency performance compared with the fixed-uplink and fixed-downlink heuristic designs. Moreover, the relative performance improvement of the joint design over the fixed designs increases as the ratio of RIS to BS elements increases. Hence, for these scenarios, the complexity comparisons along with the trade-off regions allow the designer to trade-off spectral efficiency with complexity and overhead when adopting an RIS design. For example, if complexity is not a constraint, then the optimized joint design provides the highest spectral efficiency in FDD.

3.6 Conclusion

This chapter has investigated jointly configured uplink-downlink RIS design for both FDD and TDD multi-user systems. Trade-off regions between uplink and downlink rates achieved by a jointly configured RIS design were obtained by formulating and optimizing a WSP using a BCD algorithm. A joint design is essential to the operation of FDD systems as uplink and downlink transmissions occur simultaneously and hence an individual design is not physically realizable. In TDD, a joint design can substantially reduce signalling overhead as well as reduce power consumption and configuration periods.

With the system configuration optimized by the proposed BCD algorithm, the effect of different parameters on the improvement in spectral efficiency due to a jointly optimized design compared to the fixed-downlink/fixed-uplink designs was investigated. When the user weights become less uniform and as more flexibility in resource allocation is needed, a joint design provides significant gains compared to the two heuristic designs (i.e., fixed-uplink and fixed-downlink design).

Further, the improvement in performance due to joint design is similar under both MMSE/WMMSE and ZF beamforming strategies. In addition, the optimized joint design was compared to two slicing benchmarks, slicing-with-interference and slicing-without-interference. This comparison revealed that the joint design significantly outperformed the two slicing benchmarks and that the interference between the slices alone was not responsible for the inferior performance achieved by a sliced RIS.

A complexity analysis of the joint and fixed designs was used along with the computed trade-off regions to gain insights into scenarios where a marginal loss in spectral efficiency

can be traded off for reductions in computational complexity, configuration periods, and overhead. In FDD, the gains achieved by the joint design compared with the alternative schemes were significant in all considered scenarios and thus the joint design has significant benefits for FDD in all scenarios. For TDD, the joint design yields substantial benefits compared to fixed-downlink/fixed-uplink under independent weights. Moreover, the relative performance improvement of the joint design over the fixed designs increases as the ratio of RIS to BS elements increases. Otherwise, in TDD, the fixed-uplink (i.e., using the optimized UL RIS design for both UL and DL transmissions) and fixed-downlink designs (i.e., using the optimized DL RIS design for both UL and DL transmissions) achieve almost the same performance as the individual design (separately optimized RIS configurations for UL and DL) but with less complexity than the jointly optimized design, or the individual design.

Chapter 4

A Hypernetwork Framework for Learning Adaptive Beamforming Schemes in RIS Systems

4.1 Introduction

Channel estimation and physical layer-based resource allocation, including RIS phase shifts and active beamforming configurations, are typically treated as two separate problems. First, user pilots are employed to estimate CSI based on specific criteria (e.g., MSE), as explained in Section 2.3. Then, given the estimated CSI, the RIS reflection coefficients are optimized using mathematical optimization methods such as those outlined in Section 2.2.1. Alternatively, data-driven methods offer a tractable approach to bypass explicit CSI estimation and directly optimize the RIS beamforming configuration from noisy pilots, resulting in a form of implicit channel estimation [129, 52].

In [129], an implicit channel estimation method was proposed for user scheduling problems. Moreover, in [52], neural networks were trained to map pilot information to the RIS beamforming configuration that optimizes the downlink *equal* weighted sum-rate, demonstrating the significant potential of implicit channel estimation. In practical systems, fairness requirements, service priorities, and long-term channel conditions are typically accounted for through user weights. Therefore, the goal of this chapter is to efficiently learn RIS beamforming configuration schemes that directly exploit noisy pilots to optimize the weighted sum-rate (WSR).

To accommodate different sets of user weights, one approach would be to train a separate model for each set of user weights. However, this approach is intractable in practical systems as even a coarse quantization of the user weights would require training a prohibitive number of models. A second approach is to directly extend existing learning methods by including the user weights as an input along with the pilot information. Then, a single neural network is trained to work for every possible set of user weights. This is referred to as the *static learning* approach.

Neural networks are theoretically known to be universal approximators under certain asymptotic conditions [44]. However, a single network with a practical number of layers and hidden units, trained using stochastic gradient descent methods, can be limited in how well it can approximate the function that maps the available information to the optimized beamforming configuration. Additionally, the *static learning* approach has a fundamental limitation as the deep learning model, once it is trained, would have fixed learning parameters that cannot be tuned to reflect changes in user weights (or other available information). Consequently, this can result in a significant loss in spectral efficiency, as demonstrated by the numerical experiments in Section 4.6.

To address the static learning approach limitations and provide a better approximation to the beamforming problem, we propose a novel hypernetwork-based beamforming (HNB) framework. Particularly, a neural network, called beamforming network (BFN), is designed to leverage all available information including pilot information to generate the RIS beamforming configuration that maximizes the downlink WSR. Under the proposed HNB framework, instead of directly learning a single BFN to generate beamforming configurations, another network (called hypernetwork) is trained to dynamically generate a subset of the BFN learning parameters based on an input conditioning vector. Consequently, a trained HNB composed of a BFN and its tuner (i.e., hypernetwork) represents a family of beamforming networks parameterized by the conditioning vector.

Hence, if the conditioning vector is chosen to be the user weights, the trained HNB scheme is able to tune (adapt) the learning parameters of the BFN to the user weights. This effectively allows the HNB, during deployment, to generate a different BFN for each input condition (e.g., user weights) without having to retrain for the new input condition. As demonstrated in Section 4.6, over a wide range of practical scenarios, tuning the BFN to the user weights enables the proposed HNB to operate close to an optimistic BCD algorithm with perfect CSI benchmark. Moreover, the HNB yields substantial performance gains compared to static learning schemes that directly extend existing learning methods.

When location information is available, the hypernetwork’s conditioning vector can also be exploited to tune the BFN learning parameters to the user location information

(in addition to user weights). As illustrated in Section 4.6, by tuning the BFN to the user weights and locations, an HNB scheme significantly outperforms static learning schemes and considerably reduces the number of pilots needed to provide performance close to the BCD with perfect CSI benchmark.

The work in [41] first introduced hypernetworks for applications in machine translation and image processing. In [79], hypernetworks were leveraged to investigate the Pareto front associated with multi-class classification and multi-task regression problems. Moreover, [125] employed hypernetworks to design adaptive MIMO detectors. The work in [67] exploited hypernetworks to transfer channel features from uplink to downlink in an FDD MIMO system and thus enable more efficient CSI estimation in FDD massive MIMO systems.

4.1.1 Contributions and Key Findings

In this chapter, we employ hypernetworks to capture a family of beamforming networks that can be tuned to user weights and, if available, user location information. Subsequently, hypernetwork-based beamforming is shown to effectively approximate solutions to RIS beamforming problems, significantly reducing the number of pilots needed to configure the RIS and thereby addressing a key challenge in integrating RIS systems into future communication networks. The contributions and key findings of this chapter can be categorized as follows:

i) **problem formulation and analysis:**

- We consider beamforming schemes that directly leverage noisy uplink pilots to maximize the downlink weighted sum-rate in multi-user RIS TDD systems. For this, an adaptive beamforming configuration problem is formulated to generate the beamforming configuration that optimizes the weighted sum-rate, parameterized by pilot information and user weights.
- Under some regularity conditions, this adaptive beamforming configuration problem is shown to attain a maximum and its solution optimizes a family of beamforming configuration problems.

ii) **dynamic approximation via HNB framework:**

- To approximate solutions to the formulated problem, a hypernetwork-based beamforming (HNB) framework is proposed. Particularly, a beamforming network (BFN) is designed to leverage available information (i.e., pilot information and user weights) to

generate optimized beamforming configurations. Rather than directly training one set of BFN learning parameters, another network (called hypernetwork) is trained to tune the BFN by dynamically generating a subset of the BFN learning parameters based on an input conditioning vector (e.g., user weights) fed to the hypernetwork.

- During deployment, the family of beamforming networks captured by the trained HNB can be used to generate different BFN learning parameters for each input condition without having to retrain for it.

iii) **numerical results and key findings:**

- Numerical experiments show that, even with limited training, the proposed HNB schemes can achieve significant performance improvements over schemes that directly extend *existing* methods (i.e., *static-learning benchmarks*) and bridge the gap to an optimistic block-coordinate descent with perfect CSI (denoted BCD-perfect-CSI) benchmark.
- Moreover, by tuning the BFN to both the user weights and locations, an HNB scheme is shown to considerably reduce the number of pilots needed to provide performance close to (within 10% of) the optimistic BCD-perfect-CSI benchmark.
- Furthermore, the numerical results show that the improvements introduced by the HNB-based schemes over existing methods persist for different scenarios characterized by different user weighting strategies, number of user pilots, noise considerations, Rician factors, and user location availability.

4.2 System Model and Signal Transmissions

In this chapter, a downlink time-division-duplexing (TDD) RIS system is considered where an N -element RIS assists an M -antenna base station (BS) in serving K users. Moreover, to acquire the necessary CSI, uplink pilots and training RIS reflection coefficients are employed.

The channel between the BS and RIS is denoted by $\mathbf{G}^t \in \mathbb{C}^{M \times N}$ and the channel between the RIS and the k^{th} user is $\mathbf{h}_{r,k}^t \in \mathbb{C}^{N \times 1}$, where $t \in \{DL, UL\}$ indicates whether the downlink or uplink channel is considered. Further, the direct channel between the BS and user k is given by $\mathbf{h}_{d,k}^t \in \mathbb{C}^{M \times 1}$. Here, quasi-static flat fading is assumed for all channels. An overview of the considered system is shown in Figure 4.1.

4.2.1 Downlink Data Transmission

The BS transmits a precoded signal $\mathbf{x}^{DL} = \sum_{k=1}^K \mathbf{w}_k s_k$, where the downlink information

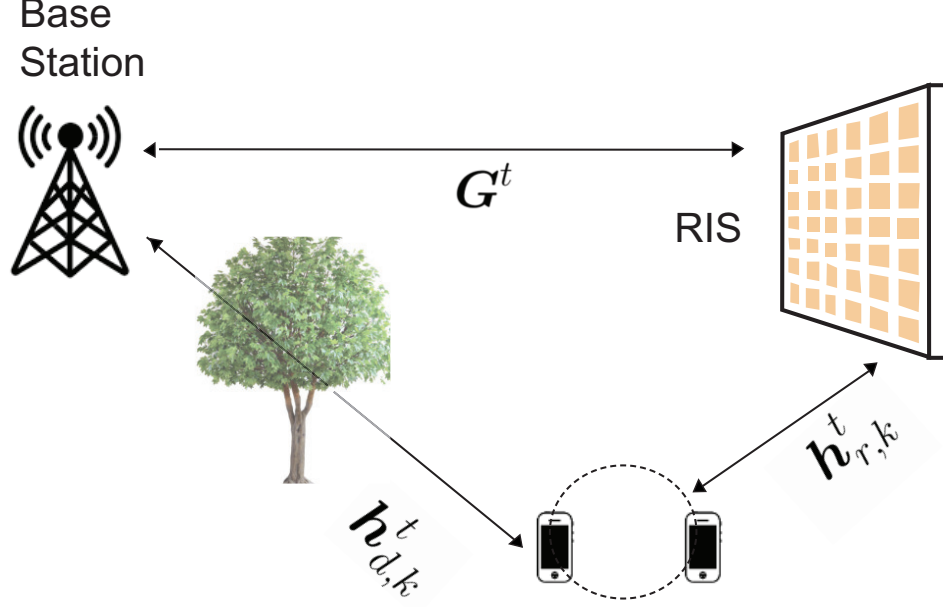


Figure 4.1: Considered multi-user RIS system where $t \in \{DL, UL\}$, indicates the downlink or uplink channel.

symbols $s_1, \dots, s_K$ are i.i.d. with zero-mean and unit power, and the column vector $\mathbf{w}_k \in \mathbb{C}^M$ is the transmit beamforming vectors of user k . The RIS elements are indexed as $n \in \{1, \dots, N\}$ and each RIS element is associated with a reflection coefficient $\theta_n = e^{-j\vartheta_n}$. Here, $\vartheta_n \in [0, 2\pi)$ is the phase shift introduced by the n^{th} RIS element. The RIS is designed to assist signal transmissions in downlink by passively beamforming using the RIS reflection coefficients $\mathbf{\Theta} = [\theta_1, \dots, \theta_N]$. Subsequently, the effective downlink channel between the BS and k^{th} user is given as $\mathbf{h}_k^{DL} = \mathbf{h}_{d,k}^{DL} + \mathbf{G}^{DL} \text{diag}(\mathbf{\Theta}) \mathbf{h}_{r,k}^{DL}$.

In downlink, the signal received by user $k \in \mathcal{K}$ is expressed as

$$y_k^{DL} = \underbrace{(\mathbf{h}_k^{DL})^T \mathbf{w}_k s_k}_{\text{Intended signal}} + \underbrace{\sum_{i \in \mathcal{K}, i \neq k} (\mathbf{h}_k^{DL})^T \mathbf{w}_i s_i}_{\text{MU-interference}} + \underbrace{n_k^{DL}}_{\text{AWGN}},$$

where $n_k^{DL} \sim \mathcal{CN}(0, \sigma_{DL}^2)$ is the circularly symmetric complex additive white Gaussian noise (AWGN) associated with user k . Thus, for user $k \in \mathcal{K}$, an achievable downlink rate

in bit/sec/Hz is given by

$$R_k = \log_2 \left(1 + \frac{|(\mathbf{h}_k^{DL})^T \mathbf{w}_k|^2}{\sum_{i \in \mathcal{K}, i \neq k} |(\mathbf{h}_k^{DL})^T \mathbf{w}_i|^2 + \sigma_{DL}^2} \right). \quad (4.1)$$

4.2.2 Uplink Pilot-based CSI Acquisition

For a channel training period of I slots, uplink pilot symbol transmissions are employed to acquire the CSI necessary for downlink data transmissions. In particular, in each slot $i \in \mathcal{I} = \{1, \dots, I\}$, the RIS reflection coefficients are set to a training configuration Θ_i^{UL} . Moreover, for user $k \in \mathcal{K}$ and training slot $i \in \mathcal{I}$, each user transmits a pilot symbol $x_{i,k}^{UL} = \sqrt{p_k^i} s_{i,k}^{UL}$. Here, $s_{i,k}^{UL} \in \mathbb{C}$ and $p_k^i > 0$ are the pilot symbol and the uplink transmission power associated with user k , respectively.

In training slot $i \in \mathcal{I}$, for a given Θ_i^{UL} , the signal received by the BS is given by

$$\mathbf{y}_i^{UL} = \sum_{k=1}^K (\mathbf{h}_{d,k}^{UL} + \mathbf{G}^{UL} \text{diag}(\Theta_i^{UL}) \mathbf{h}_{r,k}^{UL}) x_{i,k}^{UL} + \mathbf{n}_i^{UL}, \quad (4.2)$$

where $\mathbf{n}_i^{UL} \sim \mathcal{CN}(0, \sigma_{UL}^2 \mathbb{I}_M)$ is the AWGN at the BS. The pilots received over the training period are captured by the matrix $\mathbf{Y} = [\mathbf{y}_1^{UL}, \dots, \mathbf{y}_I^{UL}]$. The pilot matrix $\mathbf{Y}$ is subsequently leveraged to generate the beamforming configuration (i.e., transmit beamforming vectors and RIS reflection coefficients) that maximize the weighted sum-rate of the considered multi-user RIS system. For the considered TDD system, channel reciprocity as confirmed in works such as [85], is assumed.

4.3 Problem Formulation

In this section, we first formulate an adaptive beamforming configuration problem that leverages noisy pilots and user weights to generate the beamforming configuration that optimizes the weighted sum-rate associated with an RIS system. Subsequently, under mild regularity conditions, we show that this adaptive beamforming configuration problem attains a maximum and its solution optimizes a family of beamforming configuration problems.

4.3.1 Adaptive beamforming configuration problem formulation

In this work, no prior CSI knowledge is assumed. The user weights, capturing the relative priority when allocating resources, are denoted by the vector $\mathbf{\Gamma} = [\Gamma_1, \dots, \Gamma_K]$ such that $0 \leq \Gamma_k \leq 1$ for all $k \in \mathcal{K}$. The goal is to directly exploit pilot information $\mathbf{Y}$ and user weights $\mathbf{\Gamma}$ to optimize the transmit beamforming vectors $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_K]$ and the RIS reflection coefficients $\mathbf{\Theta}$. The pair $(\mathbf{W}, \mathbf{\Theta})$ are collectively referred to as the beamforming configuration of the RIS system.

Since the user pilots are noisy and CSI is therefore not perfectly known, there is uncertainty as to what rate (i.e., reward) will be achieved. Subsequently, for $\mathbf{Z} \triangleq (\mathbf{Y}, \mathbf{\Gamma})$ and following a stochastic optimization approach [12], the objective is to find a policy (i.e., a mapping) $\mathcal{F}(\cdot)$ such that $(\mathbf{W}, \mathbf{\Theta}) = \mathcal{F}(\mathbf{Z})$ optimizes the expected downlink weighted sum-rate (WSR) over the joint distribution of channels, user weights and noisy pilots. Towards this goal, an adaptive beamforming configuration problem is formulated as

$$\sup_{\mathcal{F}(\mathbf{Z})} \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \mathbf{\Theta}) \right] \quad (4.3a)$$

$$\text{subject to} \quad (\mathbf{W}, \mathbf{\Theta}) = \mathcal{F}(\mathbf{Z}) \quad (4.3b)$$

$$\|\mathbf{W}\|_F^2 \leq P_{max}^{DL} \quad (4.3c)$$

$$|[\mathbf{\Theta}]_n| = 1, n \in \mathcal{N}, \quad (4.3d)$$

where $[\mathbf{\Theta}]_n$ is the n^{th} component of $\mathbf{\Theta}$ and the sup in (4.3) is taken over all measurable mappings $\mathcal{F}$. Moreover, the downlink rate R_k is as given in (4.1) and the expectation is over the distribution of the downlink channel components $\mathbf{G}^{DL}$ and $\{\mathbf{h}_{d,k}^{DL}, \mathbf{h}_{r,k}^{DL}\}_{k \in \mathcal{K}}$, user weights $\mathbf{\Gamma}$ and received uplink pilot information $\mathbf{Y}$. As indicated by (4.3c), the transmit beamforming vectors $\mathbf{W}$ are subject to the sum-power constraint with P_{max}^{DL} indicating the maximum sum power at the BS.

We note that problem (4.3) was formulated with sup to emphasize that in general, for problems such as (4.3), it is not guaranteed that there exists a measurable mapping that attains the maximum [12, Chap. 10].

4.3.2 Analyzing the adaptive beamforming configuration problem

Solutions (and approximate solutions) to the adaptive beamforming configuration problem of (4.3) are of interest as they characterize a family of beamforming schemes parameterized

by $\mathbf{Z} = (\mathbf{Y}, \mathbf{\Gamma})$. Once obtained, such solutions can be employed to adaptively generate $\mathbf{W}$ and $\mathbf{\Theta}$ that optimize the WSR for different user weights and pilot information.

In this section, under some regularity conditions, we show that the adaptive beamforming configuration problem of (4.3) attains a maximum and its solution optimizes a family of beamforming configuration problems. To show this result, we first introduce, for each user, the notion of a capacity bound U_k , and an associated regularity condition.

Specifically, the downlink achievable rate of user k , $R_k(\mathbf{W}, \mathbf{\Theta})$, can be upper-bounded as follows,

$$\begin{aligned}
& \log_2 \left(1 + \frac{|(\mathbf{h}_{d,k}^{DL} + \mathbf{G}^{DL} \text{diag}(\mathbf{\Theta}) \mathbf{h}_{r,k}^{DL})^T \mathbf{w}_k|^2}{\sum_{\substack{i \in \mathcal{K} \\ i \neq k}} |(\mathbf{h}_{d,i}^{DL} + \mathbf{G}^{DL} \text{diag}(\mathbf{\Theta}) \mathbf{h}_{r,i}^{DL})^T \mathbf{w}_i|^2 + \sigma_{DL}^2} \right) \\
& \leq \log_2 \left(1 + \frac{|(\mathbf{h}_{d,k}^{DL} + \mathbf{G}^{DL} \text{diag}(\mathbf{\Theta}) \mathbf{h}_{r,k}^{DL})^H \bar{\mathbf{w}}_k|^2}{\sigma_{DL}^2} \right) \\
& \stackrel{(a)}{\leq} \log_2 \left(1 + \frac{(\|\mathbf{h}_{d,k}^{DL}\| + \sum_{n=1}^N |h_{n,r,k}^{DL}| \|\mathbf{g}_n^{DL}\|)^2 P_{max}^{DL}}{\sigma_{DL}^2} \right) \\
& \triangleq U_k,
\end{aligned} \tag{4.4}$$

where $\bar{\mathbf{w}}_k$ is the elementwise complex conjugate of $\mathbf{w}_k$. Moreover, $\mathbf{g}_n^{DL}$ is the n^{th} column of $\mathbf{G}^{DL}$, and (a) follows as

$$\begin{aligned}
& |(\mathbf{h}_{d,k}^{DL} + \mathbf{G}^{DL} \text{diag}(\mathbf{\Theta}) \mathbf{h}_{r,k}^{DL})^H \bar{\mathbf{w}}_k| \\
& \leq |(\mathbf{h}_{d,k}^{DL})^H \bar{\mathbf{w}}_k| + |(\mathbf{G}^{DL} \text{diag}(\mathbf{\Theta}) \mathbf{h}_{r,k}^{DL})^H \bar{\mathbf{w}}_k| \\
& \leq \|\mathbf{h}_{d,k}^{DL}\| \|\mathbf{w}_k\| + \|\mathbf{w}_k\| \sum_{n=1}^N |h_{n,r,k}^{DL}| \|\mathbf{g}_n^{DL}\|.
\end{aligned}$$

We note that the bound in (4.4) does not depend on the choice of $\mathbf{\Theta}$ or $\mathbf{W}$. Next, we introduce a family of beamforming configuration problems by conditioning on the pilot information $\mathbf{Y} = \mathbf{y}$ and user weights vector $\mathbf{\Gamma} = \boldsymbol{\gamma}$. Particularly, for each given $\mathbf{z} = (\mathbf{y}, \boldsymbol{\gamma})$, this conditional beamforming configuration problem can be expressed as

$$\sup_{\mathbf{\Theta}, \mathbf{W}} \quad \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \mathbf{\Theta}) \mid \mathbf{Z} = \mathbf{z} \right] \tag{4.5a}$$

$$\text{subject to} \quad \|\mathbf{W}\|_F^2 \leq P_{max}^{DL} \tag{4.5b}$$

$$|[\mathbf{\Theta}]_n| = 1, n \in \mathcal{N}. \tag{4.5c}$$

Here, the expectation is over the downlink channels and conditioned on realizations $\mathbf{z} = (\mathbf{y}, \gamma)$ of $\mathbf{Z} = (\mathbf{Y}, \Gamma)$. Proposition 4.1 below shows that under mild assumptions, the adaptive beamforming configuration problem of (4.3) attains a maximum and an optimizer of Problem (4.3) also maximizes the conditional problem in (4.5) for almost all $\mathbf{y}$ and γ . As shown following Lemma 1 of Appendix A.3, common fading models such as Rician, Rayleigh, and Nakagami satisfy the $\mathbb{E}[U_k] < \infty$ assumption of Proposition 4.1.

Proposition 4.1. *If $\mathbb{E}[U_k] < \infty$ for all $k \in \mathcal{K}$ and the joint distribution on the channels, pilots and user weights does not depend on the choice of $\mathcal{F}(\mathbf{Z})$, then:*

- (a) *The adaptive beamforming configuration problem in (4.3) attains a maximum that can be achieved by a family of solutions $\tilde{\mathcal{F}}(\mathbf{Z})$ obtained from the conditional problem in (4.5).*
- (b) *Suppose that $\mathcal{F}^*(\mathbf{Z})$ optimizes the problem in (4.3), then for $\tilde{\mathcal{F}}(\mathbf{Z})$ of part (a) above,*

$$\begin{aligned} \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\mathcal{F}^*(\mathbf{Z})) | \mathbf{Z} = \mathbf{z} \right] \\ = \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z})) | \mathbf{Z} = \mathbf{z} \right] \end{aligned} \quad (4.6)$$

for almost all $\mathbf{z}$.

The proof is given in Appendix A.3.

To dynamically approximate solutions to the computationally intractable optimization problem in (4.3), in the next section, a novel HNB framework is presented. Under the HNB framework, a BFN is structured to leverage $\mathbf{Z} = (\mathbf{Y}, \Gamma)$ to generate the beamforming configuration (comprised of transmit beamforming vectors $\mathbf{W}$ and RIS reflection coefficients Θ) that optimize the WSR. Rather than directly learning the BFN, another network (called hypernetwork) is trained to dynamically generate a subset of the BFN learning parameters based on an input conditioning vector Ξ . Thus, by choosing $\Xi = \Gamma$, the trained HNB represents a family of BFNs, parameterized by the user weights. During the deployment stage, the HNB is expected to provide a better approximation to the optimal beamforming solution compared to the case where a single BFN is directly trained to optimize beamforming configurations for all user weights. This follows as the trained HNB can tune the learning parameters of the BFN to the user weights and any additional information provided by Ξ .

4.4 Hypernetwork-based Beamforming Framework

4.4.1 Hypernetwork-based beamforming

To optimize the WSR objective in (4.3), a BFN, parameterized by learning parameters $\mathbf{W}$, is structured to map the pilot information $\mathbf{Y}$ into an optimized beamforming configuration comprised of transmit beamforming vectors $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_K]$ and RIS reflection coefficients Θ . In addition to the pilot information $\mathbf{Y}$, the BFN exploits other available information η , including the user weights Γ and (if available) the user locations $\mathbf{p}_1, \dots, \mathbf{p}_K$. The latter may improve the performance achieved by the generated beamforming configuration. Subsequently, the BFN can be expressed as a mapping $g(\cdot; \mathbf{W})$ such that

$$(\mathbf{W}, \Theta) = g(\mathbf{Y}, \eta; \mathbf{W}), \quad (4.7)$$

where $\mathbf{W}$ denotes the BFN learning parameters. Rather than statically learning one set of $\mathbf{W}$, in an HNB framework, the learning parameters $\mathbf{W}$ are partitioned into 2 components $\mathbf{W} = (\mathbf{W}_S, \mathbf{W}_D)$. Here, the $\mathbf{W}_S$ component remains statically trained while the learning parameters $\mathbf{W}_D$ are dynamically generated by another neural network (called hypernetwork) based on an input conditioning vector Ξ . Particularly, the hypernetwork is a mapping

$$\mathbf{W}_D = f(\Xi; \mathbf{W}_H), \quad (4.8)$$

where the input conditioning vector to the hypernetwork is Ξ and $\mathbf{W}_H$ denotes the learning parameters associated with the hypernetwork. Unless otherwise stated, the user weights Γ are fed to the hypernetwork, i.e., $\Xi = \Gamma$, and the HNB scheme (comprised of the hypernetwork and the beamforming network) represents a family of BFNs, where $\mathbf{W}_D$ is parameterized by Γ . Under the proposed framework, the hypernetwork is trained to generate the $\mathbf{W}_D$ parameters of the BFN that optimize the WSR, corresponding to the following hypernetwork-based beamforming learning problem:

$$\max_{\mathbf{W}_H, \mathbf{W}_S} \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \Theta) \right] \quad (4.9a)$$

$$\text{subject to} \quad (\mathbf{W}, \Theta) = g(\mathbf{Y}, \eta; \mathbf{W}_S, \mathbf{W}_D) \quad (4.9b)$$

$$\mathbf{W}_D = f(\Xi; \mathbf{W}_H) \quad (4.9c)$$

$$\|\mathbf{W}\|_F^2 \leq P_{max}^{DL}, \quad (4.9d)$$

$$|[\Theta]_n| = 1, n \in \mathcal{N}. \quad (4.9e)$$

The expectation in (4.9) is over the distribution of user weights, downlink channels and received uplink user pilots.

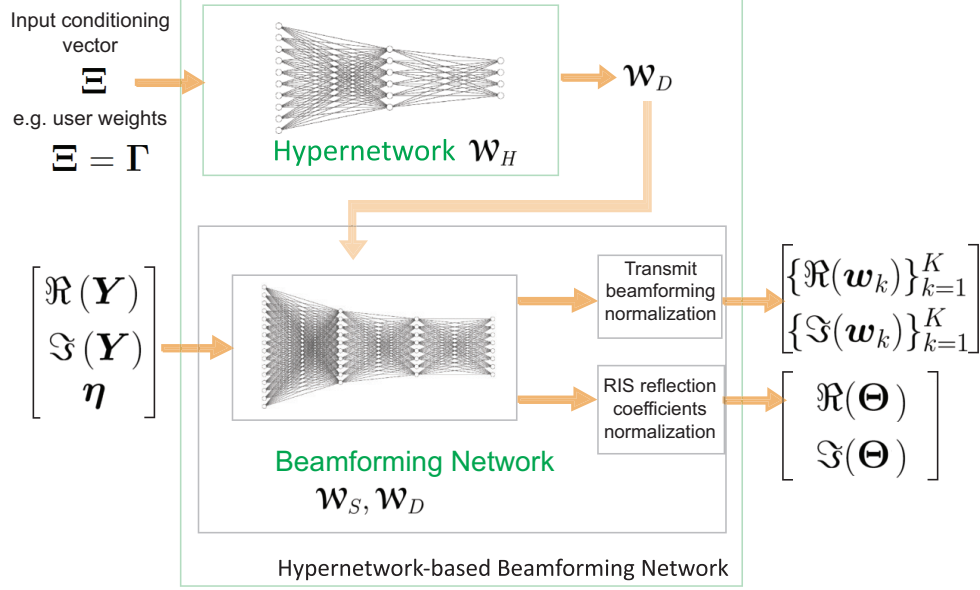


Figure 4.2: Hypernetwork-based Beamforming architecture. A hypernetwork is trained to output the learning parameters $\mathcal{W}_D$ of the beamforming network.

4.4.2 HNB Architecture

The architecture of the hypernetwork-based beamformer is shown in Fig. 4.2. The beamforming network is implemented using a deep neural network with learning parameters $\mathcal{W}_S$ and $\mathcal{W}_D$. Moreover, the hypernetwork, implemented using a deep neural network with learning parameters $\mathcal{W}_H$, is trained to generate a subset $\mathcal{W}_D$ of the learning parameters of the BFN based on an input conditioning vector Ξ (e.g., $\Xi = \Gamma$). In particular, for $(\mathbf{W}, \boldsymbol{\Theta}) = g(\mathbf{Y}, \boldsymbol{\eta}; \mathcal{W}_S, \mathcal{W}_D)$ and $\mathcal{W}_D = f(\Xi; \mathcal{W}_H)$, the loss function $-\mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \boldsymbol{\Theta}) \right]$ is used to update the learning parameters $\mathcal{W}_H$ and $\mathcal{W}_S$.

We emphasize that the proposed HNB framework can also be employed to tune the learning parameters of the BFN to the user locations, in addition to the user weights¹. In this case, $\Xi = (\Gamma, \{\mathbf{p}_k\}_{k=1}^K)$ and as will be seen in Section 4.6, the number of pilots required by the proposed HNB to design the beamforming configuration is significantly reduced.

In this work, the hypernetwork is implemented using a fully-connected deep neural network. The beamforming network is designed using either a fully connected neural

¹Unless otherwise specified, the beamforming learning schemes do not leverage location information

network (FCNN) or a graph neural network (GNN) architecture:

BFN design- FCNN Architecture

The BFN is comprised of an input processing layer, P hidden layers, and an output layer. As illustrated in Fig. 4.2, normalization layers are used to impose the constraints on the generated transmit beamforming vectors $\mathbf{W}$ and RIS reflection coefficients Θ : the function $\sigma_W(\mathbf{W}) = \sqrt{P_{max}^{DL}} \mathbf{W} / \|\mathbf{W}\|_F$ enforces a sum-power constraint on $\mathbf{W}$ and the function $\sigma_\theta(\theta_n) = \theta_n / |\theta_n|$ enforces the unit-modulus constraints on the Θ .

BFN design - GNN Architecture

In this case, the BFN is comprised of an input processing layer, L_G GNN layers indexed by $l \in \mathcal{L}_G = \{1, \dots, L_G\}$ and an output layer with normalization. Each of the L_G GNN layers is comprised of N_G nodes $\mathcal{N}_G = \{1, \dots, N_G\}$ [97]. For each layer $l \in \mathcal{L}_G$, a node $i \in \mathcal{N}_G$ employs aggregation networks and combination networks to update its representation $\mathbf{d}_i^{(l)}$ using information from previous layers and its neighboring nodes [115].

The aggregated information associated with node i at layer l is computed as $\mathbf{a}_i^{(l)} = \Psi_a \left(\{\mathbf{q}_{a,i}^{(l)}(\mathbf{d}_i^{(l-1)})\}_{i \in \mathcal{N}_i} \right)$, where $\mathbf{q}_{a,i}^{(l)}$ is an aggregation network, Ψ_a is a permutation invariant pooling function, and $\mathcal{N}_i$ is the set of neighbors associated with node i . Next, $\mathbf{d}_i^{(l)}$ is computed by combining the aggregated information $\mathbf{a}_i^{(l)}$ with $\mathbf{d}_i^{(l-1)}$ from the previous layer. Specifically, the representation is updated as $\mathbf{d}_i^{(l)} = \mathbf{q}_{c,i}^{(l)} \left(\mathbf{d}_i^{(l-1)}, \mathbf{a}_i^{(l)} \right)$, where $\mathbf{q}_{c,i}^{(l)}$ is a combination network. The node representations associated with the final GNN layer are processed to generate the RIS configuration and beamforming vectors. Similar to the BFN-FCNN architecture, the constraints on the generated $\mathbf{W}$ and Θ are enforced using normalization layers.

4.5 Proposed HNB Schemes and Benchmarks

Depending on whether a FCNN or a GNN architecture is adopted for the BFN design, two HNB schemes, denoted HNB-FCNN and HNB-GNN, are proposed to approximate solutions to the beamforming learning problem (4.3). In this section, we present the implementation details of the proposed HNB schemes along with four benchmarks: an optimistic BCD benchmark with perfect CSI and three alternative static deep learning beamforming schemes.

Table 4.1: BFN Architectures

A. BFN-FCNN Architecture				
Network	Parameterization		Number of parameters	
$q_{\text{BFN}}^{\text{FCNN}}$	$2MKT \times 512 \times 512 \times 512 \times (2MK + 2N)$		$\rho = \rho_i + \rho_h + \rho_o$ $\rho_i = (2MKT) 512$ $\rho_h = 2 (512) (512)$ $\rho_o = (512) (2MK + 2N)$	
B. BFN-GNN2 Architecture			C. BFN-GNN3 Architecture	
Network	Parameterization	Number of parameters	Parameterization	Number of parameters
$q_{\text{in}}^{\text{User}}$	$2MT \times 1024 \times 512$	$\chi_{i,1}$	$2MT \times 1024 \times 512 \times 512$	$\beta_{i,1}$
$q_{\text{in}}^{\text{RIS}}$	$512 \times 1024 \times 512$	$\chi_{i,2}$	$512 \times 1024 \times 512 \times 512$	$\beta_{i,2}$
$q_{a,I}^{(l)}$	$512 \times 512 \times 512$	$\chi_{a,1}$	$512 \times 512 \times 512 \times 512$	$\beta_{a,1}$
$q_{a,I,U}^{(l)}$	$512 \times 512 \times 512$	$\chi_{a,2}$	$512 \times 512 \times 512 \times 512$	$\beta_{a,2}$
$q_{c,I}^{(l)}$	$512 \times 512 \times 512$	$\chi_{c,1}$	$512 \times 512 \times 512 \times 512$	$\beta_{c,1}$
$q_{a,U,U}^{(l)}$	$512 \times 512 \times 512$	$\chi_{a,3}$	$512 \times 512 \times 512 \times 512$	$\beta_{a,3}$
$q_{c,U}^{(l)}$	$512 \times 512 \times 512$	$\chi_{c,2}$	$512 \times 512 \times 512 \times 512$	$\beta_{c,2}$
$q_o^{(U)}$	$512 \times 2M$	$\chi_{o,1}$	$512 \times 2M$	$\beta_{o,1}$
$q_o^{(I)}$	$512 \times 2N$	$\chi_{o,2}$	$512 \times 2N$	$\beta_{o,2}$

4.5.1 Benchmarks

In order to evaluate the performance of the proposed HNB schemes, an optimistic benchmark and 3 alternate static deep learning beamforming schemes are presented:

BCD-perfect-CSI benchmark

Under this optimistic benchmark, perfect CSI is assumed, and a block-coordinate descent (BCD) algorithm iteratively employs manifold optimization and weighted MMSE to optimize the beamforming configuration of the RIS system [40].

Static-FCNN benchmark

The BFN, denoted $q_{\text{BFN}}^{\text{FCNN}}$, is implemented with an input layer, $P = 3$ hidden layers of 512 units and an output layer with normalization as described in part A of Table 4.1. Under this static approach, the BFN is trained to work for every possible set of user weights. In particular, the BFN is directly trained to utilize the pilot information and user weights to optimize the WSR.

Static-GNN2 benchmark

The K users and RIS are modeled using a graph representation of $K + 1$ nodes, indexed by $i \in \mathcal{N}_G = \{1, \dots, K + 1\}$ and the BFN is implemented with an $L_G = 2$ -layer GNN (i.e., the GNN of [52]). The BFN is fed with the pilot information and user weights and is then directly trained to optimize the WSR. The implementation details of the network components forming the static-GNN2 scheme are provided in part B of Table 4.1 and explained in Appendix B.

Static-GNN3 benchmark

The static-GNN2 scheme is extended by adding an additional GNN layer (i.e., $L_G = 3$). Moreover, the sub-networks associated with the input layer as well as the sub-networks of each GNN layer are also extended by adding an additional hidden layer of 512 units. The implementation details are provided in part C of Table 4.1.

Extensions of the static-FCNN scheme above were considered by increasing the number of hidden layers and/or increasing the number of units per hidden layer. However, none of these extensions resulted in significant spectral efficiency performance improvement. Similarly, static-GNN extensions did not yield significant improvements.

4.5.2 Proposed Schemes

The performance of the proposed HNB architecture of Fig. 4.2 is evaluated under two BFN designs:

Proposed Scheme 1- HNB-FCNN

In this HNB scheme, to ensure a fair comparison, the static-FCNN scheme and its inputs are adopted as the BFN. The learning parameters of BFN are partitioned into a static component and a dynamic component. The static component is directly trained, while the dynamic component of the BFN's learning parameters is generated by the hypernetwork. Specifically, the hypernetwork is trained to dynamically generate the learning parameters associated with the last two hidden layers of the static-FCNN scheme. The hypernetwork is implemented using a network of size $H \times 300 \times 200 \times \rho_h$, where ρ_h is specified in Table 4.1. The value of H depends on the input to the hypernetwork: (i) if only the user weights are used then $H = K$ and (ii) if both the user weights and user locations are used then $H = 4K$.

Proposed Scheme 2- HNB-GNN2

To ensure a fair comparison, the BFN is the static-GNN2 described above along with its inputs. The learning parameters of the BFN are divided into a static component and a dynamic component. The hypernetwork is trained to generate the learning parameters associated with the combination networks $\mathbf{q}_{c,I}^{(l)}$ and $\mathbf{q}_{c,U}^{(l)}$ for $l = 1, 2$ (see Table 4.1). The hypernetwork is implemented as a network of size $H \times 300 \times 200 \times (2\chi_{c,1} + 2\chi_{c,2})$, where $\chi_{c,1}$ and $\chi_{c,2}$ are given in Table 4.1.

We remark that the proposed HNB and alternate learning schemes directly exploit noisy pilots to optimize the beamforming configuration of the RIS system. Conversely, the *BCD-perfect-CSI* benchmark employs an iterative algorithm to optimize the non-convex beamforming problem and assumes perfect CSI knowledge. Further, the beamforming learning schemes above can exploit user location information along with user pilots to reduce the number of pilots needed to design the beamforming configuration of the RIS system.

4.6 Numerical Experiments

Numerical experiments are employed to evaluate the efficacy of the proposed HNB framework in maximizing the WSR. The spectral efficiency achieved by the HNB schemes and benchmarks are evaluated under two user weighting strategies: one that emphasizes significant variation in user priorities (*independent user* weights) and another that provides proportional-fairness (*PF* user weights). Further, by adapting the BFN to user locations (if available), numerical experiments demonstrate that the number of pilots needed to configure the RIS system can be significantly reduced compared to the static learning schemes where locations are only fed into the BFN as inputs.

4.6.1 System and Channel Models

Fig. 4.3 outlines the considered downlink TDD system where a base station (BS) serves multiple users and is assisted by an RIS. The RIS is a uniform rectangular array (URA) of $N = 10 \times 10 = 100$ elements and the BS is a URA of $M = 4 \times 2 = 8$ antennas. The BS and the RIS are placed at $(x_{BS}, y_{BS}, z_{BS}) = (0, 250, 25)$ m and $(x_{RIS}, y_{RIS}, z_{RIS}) = (250, 0, 15)$ m, respectively. For $k \in \mathcal{K} = \{1, \dots, 4\}$, the user location $\mathbf{p}_k = (x_k, y_k, z_k)$ is generated such that $z_k = 1.5$ m and (x_k, y_k) coordinates are uniformly generated inside an x-y disc of center $(x_c, y_c) = (250, 30)$ m and radius $R = 25$ m.

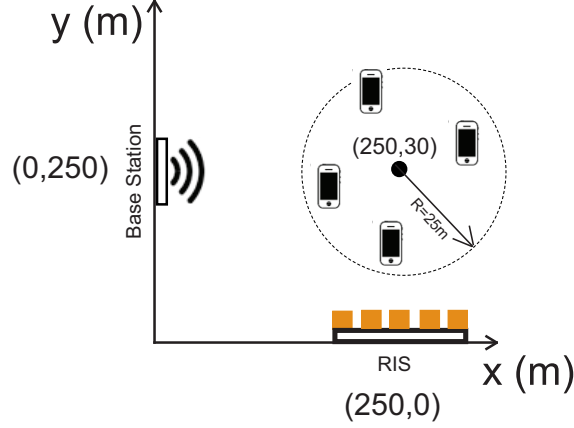


Figure 4.3: Top-view for the considered multi-user RIS system.

The considered downlink system operates at a frequency $f = 1.5$ GHz with a bandwidth B of 20 MHz. The maximum downlink transmission power is $P_{max}^{DL} = 25$ dBm and the downlink noise power is $\sigma_{DL}^2 = B \times N_0 \times F_{DL}$ where the noise figure in the downlink is $F_{DL} = 9$ dB and $N_0 = -170$ dBm/Hz. Over the training period of I slots, uplink pilots are used to acquire pilot information needed for optimizing the beamforming configuration of the considered RIS system. For $k \in \mathcal{K} = \{1, \dots, K\}$ and $i \in \{1, \dots, I\}$, the uplink pilot powers are $p_k^i = 23$ dBm and the uplink noise power is $\sigma_{UL}^2 = B \times N_0 \times F_{UL}$, where the noise figure in uplink is $F_{UL} = 5$ dB. Channel reciprocity in this TDD RIS system is assumed, as established experimentally in [85]. Subsequently, the superscript $t \in \{UL, DL\}$, indicating whether the uplink or downlink channel is considered, is omitted for simplicity.

The channel between the BS and user k undergoes Rayleigh fading and is given by $\mathbf{h}_{d,k} = \sqrt{PL_N(d_{B,k})} \mathbf{h}_{d,k}^{NLOS}$, where the components of $\mathbf{h}_{d,k}^{NLOS}$ are generated i.i.d. from $\mathcal{CN}(0, 1)$. Here, $d_{B,k}$ is the 3D distance between the BS and user k , and the non-line-of-sight (NLOS) path loss is $PL_N(d_{3D}) = 13.54 + 39.08 \log_{10}(d_{3D}) + 20 \log_{10}(f)$ dB [1].

The BS-RIS channel $\mathbf{G}$ and RIS-user k channel $\mathbf{h}_{r,k}$ are modeled as Rician fading. Given that $\mathbf{G}_{LoS} \in \mathbb{C}^{M \times N}$ and $\mathbf{h}_{r,k}^{LoS} \in \mathbb{C}^{N \times 1}$ are the LoS components of the BS-RIS and RIS-user k channels, and $\mathbf{G}_{NLoS} \in \mathbb{C}^{M \times N}$ and $\mathbf{h}_{r,k}^{NLoS} \in \mathbb{C}^{N \times 1}$ are the NLoS components of the BS-RIS and RIS-user k channels, then $\mathbf{G}$ and $\mathbf{h}_{r,k}$ are given by

$$\mathbf{G} = \sqrt{PL_L(d_{B,I})} \left(\sqrt{\frac{\kappa}{1+\kappa}} \mathbf{G}_{LoS} + \sqrt{\frac{1}{1+\kappa}} \mathbf{G}_{NLoS} \right),$$

$$\mathbf{h}_{r,k} = \sqrt{PL_L(d_{I,k})} \left(\sqrt{\frac{\kappa_k}{1+\kappa_k}} \mathbf{h}_{r,k}^{LoS} + \sqrt{\frac{1}{1+\kappa_k}} \mathbf{h}_{r,k}^{NLoS} \right).$$

Here, $d_{B,I}$ and $d_{I,k}$ indicate the 3D distance between the BS and RIS and the distance between the RIS and user k , respectively. The Rician factors associated with the BS-RIS and RIS- k^{th} user channel are given by κ and $\{\kappa_k\}_{k=1}^K$, respectively². Moreover, the LOS path loss in dB is $PL_L(d_{3D}) = 28 + 22 \log_{10}(d_{3D}) + 20 \log_{10}(f)$ dB [1]. For $d_{3D}^{m,n}$ indicating the 3D distance between the m^{th} BS antenna and the n^{th} RIS element and λ_{DL} indicating the wavelength associated with downlink transmission, the components of $\mathbf{G}_{LoS}$ are $[\mathbf{G}_{LoS}]_{m,n} = \exp(2\pi j d_{3D}^{m,n} / \lambda_{DL})$. The components of $\mathbf{G}_{NLoS} \in \mathbb{C}^{M \times N}$ and $\mathbf{h}_{r,k}^{NLoS} \in \mathbb{C}^{N \times 1}$ are generated i.i.d. from $\mathcal{CN}(0, 1)$.

4.6.2 Channel Training and CSI Acquisition

The channel training period, comprised of $I = KT$ slots, is divided into T sub-frames of K symbols each. For each sub-frame $\tau \in \mathcal{T} = \{1, \dots, T\}$, the RIS reflection coefficients can be designed to facilitate CSI acquisition. For numerical experiments, adopting the training RIS reflection coefficients in [52], the RIS is set to a randomly generated training configuration Θ_τ^{UL} such that $\{\Theta_\tau^{UL}\}_{\tau=1}^T$ are independent. In sub-frame $\tau \in \mathcal{T}$, each user k transmits an orthogonal pilot sequence $\mathbf{x}_k^{UL} \in \mathbb{C}^{K \times 1}$ and correspondingly over K symbols, the BS receives the pilot signals $\tilde{\mathbf{Y}}_\tau = [\mathbf{y}_{(\tau-1)K+1}^{UL}, \dots, \mathbf{y}_{\tau K}^{UL}]$. Here, $\mathbf{y}_i^{UL}$ is as given in (4.2). Correlating the sub-frame pilot information $\tilde{\mathbf{Y}}_\tau$ with the k^{th} pilot sequence $\mathbf{x}_k^{UL}$ yields the pilot information associated with user k as $\hat{\mathbf{y}}_{\tau,k}^{UL} = \frac{1}{K} \tilde{\mathbf{Y}}_\tau \mathbf{x}_k^{UL}$. The matrix $\hat{\mathbf{Y}}_k = [\hat{\mathbf{y}}_{1,k}^{UL}, \dots, \hat{\mathbf{y}}_{T,k}^{UL}]$ then represents the pilot information contributed by user k across the T sub-frames and is used to train and test the beamforming learning schemes.

4.6.3 Optimizing WSR under significant variation in user priorities

In this section, we demonstrate the efficiency of the two proposed HNB schemes (HNB-GNN2 and HNB-FCNN) compared to the alternate schemes (e.g., static-FCNN, static-GNN2, and static-GNN3) when there is significant variation in user weights. In particular, the downlink weighted sum-rate optimization problem is first considered for an *independent* user weighting strategy. Under this strategy, the user weights are independently and

²Unless otherwise specified, the Rician factors are all equal and chosen to be either 4, 6, 8, or 10 dB.

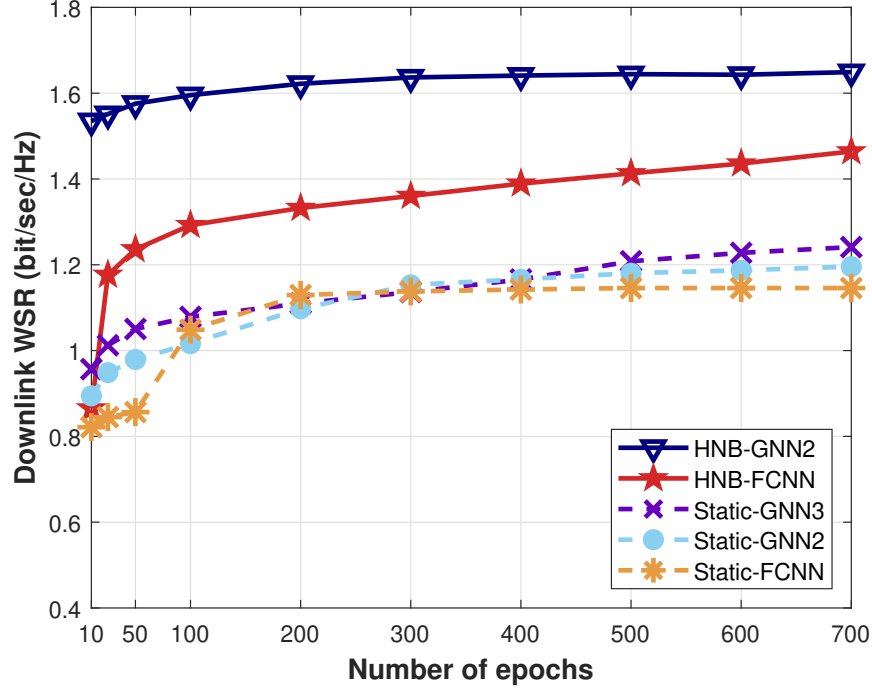


Figure 4.4: Average downlink WSR on testing data after different number of training epochs, for independent user weights, $T = 20$, and Rician factor of 10 dB.

uniformly generated on $[0,1]$, then normalized (i.e., $\sum_{k \in \mathcal{K}} \Gamma_k = 1$). In Section 4.6.4, a proportional-fair weighting scheme will be considered.

Training and Testing of proposed HNB and Alternate Schemes

The *HNB-GNN2* and *HNB-FCNN* hypernetwork-based beamforming schemes, proposed in Section 4.5, were implemented using Tensorflow and the hypernetwork training of [41]. Adam optimizer was used to train the *HNB-GNN2* and *HNB-FCNN* to optimize the hypernetwork-based beamforming learning problem (4.9). Each training epoch is comprised of 250 iterations and the initial learning rate was set to 1×10^{-4} with a decay rate of 0.99 for each 800 iterations. For each iteration, the learning parameters of the schemes were updated using 400 training samples. The performance of the trained models was then assessed by averaging the achievable normalized WSR over 1000 new independently generated test samples.

Impact of BFN Training and Architecture

For $T = 20$ pilots per user and without location information, Fig. 4.4 shows the testing WSR performance of the two proposed hypernetwork-based beamforming learning schemes (HNB-GNN2 and HNB-FCNN) and three alternate static schemes (static-FCNN, static-GNN2, and static-GNN3) as a function of the number of training epochs. The beamforming schemes were trained for a maximum of 700 epochs. Further training yielded no significant improvement in the WSR performance.

Fig. 4.4 demonstrates that the proposed HNB-GNN2 and HNB-FCNN schemes significantly outperform the static schemes, including the static-GNN3 scheme that features an additional GNN layer and deeper sub-networks compared to static-GNN2. For instance, the proposed fully-trained (after 700 epochs) HNB-GNN2 scheme improves on the fully-trained static-FCNN, static-GNN2 and static-GNN3 schemes by 44%, 38% and 33%, respectively. Similarly, after 700 epochs, the proposed HNB-FCNN scheme outperforms the static-FCNN, static-GNN2 and static-GNN3 schemes by 28%, 23% and 18%, respectively. It is to note that for both static and hypernetwork-based schemes, a GNN-based beamforming network outperforms the corresponding FCNN-based network.

From Fig. 4.4, it is also seen that even with limited training, the proposed HNB-GNN2 and HNB-FCN schemes outperform fully-trained static beamforming schemes. For example, in terms of testing WSR, the proposed HNB-GNN2 and HNB-FCNN schemes outperform the 3 fully-trained static learning schemes after only 10 and 50 training epochs, respectively. This demonstrates the efficacy of the proposed HNB schemes in terms of optimizing the WSR in different scenarios.

It is to note that, although not shown, a massive MIMO scenario with $M = 8 \times 8 = 64$ base station antennas and $N = 200$ RIS elements was considered and in this case the proposed HNB scheme also provided significant improvements over the corresponding static learning benchmarks.

Impact of user pilot and location information availability

Fig. 4.5 shows the impact of pilot and location information availability on the testing WSR achieved by the proposed HNB-GNN2 scheme and alternate static-GNN3 scheme. The performance of these is shown for different number of pilots per user, T , both with and without user location information being fed to the BFN (i.e., as part of $\boldsymbol{\eta}$ described in Section 4.4.1). For HNB schemes, the location information, if available, is also fed along with the user weights as the hypernetwork input $\Xi = [\mathbf{I}, \{\mathbf{p}_k\}_{k=1}^K]$. Hence, in this case

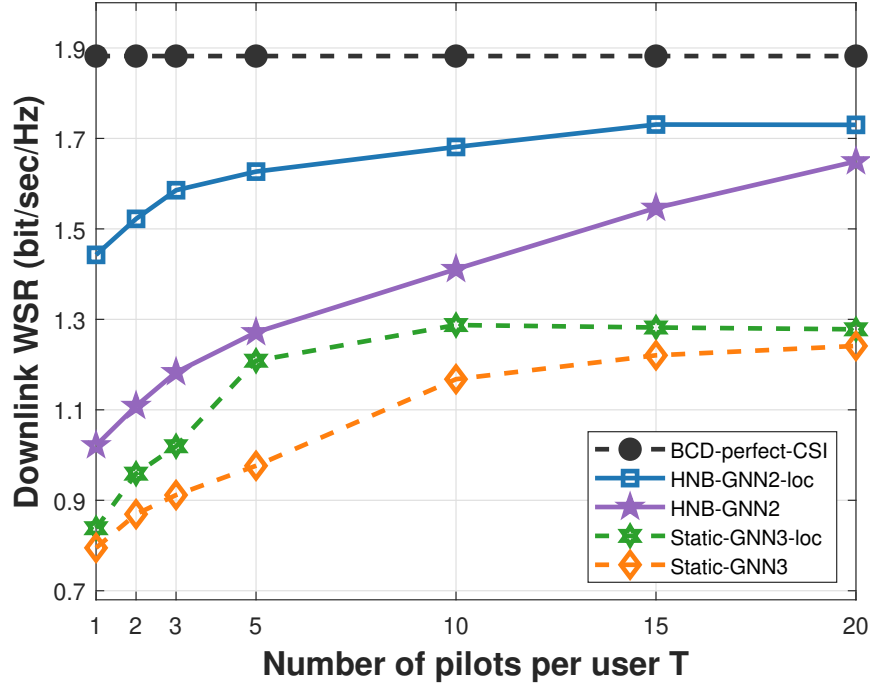


Figure 4.5: Average downlink WSR for different number of pilots per user T , under independent user weights and Rician factors of 10 dB, with user location (e.g., HNB-GNN2-loc) and without user location (e.g., HNB-GNN2).

the learning parameters of the BFN are tuned to both the location information and user weights.

Fig. 4.5 demonstrates that, for each considered T and irrespective of location information availability, the proposed HNB-GNN2 scheme significantly bridges the performance gap to the optimistic BCD-perfect-CSI benchmark compared to the two static-GNN3 schemes. For instance, the static-GNN3 scheme (without user location) with $T = 20$ results in a performance gap of 34% compared to the BCD-perfect-CSI benchmark. This is compared to the proposed HNB-GNN2 scheme (also without user location) that, using only $T = 20$ pilots per user, closes the relative performance gap to the optimistic BCD-perfect-CSI benchmark to 13%. Similarly, with $T = 20$ and with location information available, the proposed HNB-GNN2-loc scheme reduces the gap to the BCD-perfect-CSI benchmark to 8% compared to a gap of 33% associated with the corresponding static-GNN3-loc scheme.

Fig. 4.5 also shows that the proposed HNB-GNN2-loc scheme with location information

continues to operate close to the BCD-perfect-CSI benchmark when the number of user pilots is decreased and hence provides flexibility in trading off spectral efficiency performance with pilot overhead. In fact, compared to higher values of $T = 30$ or $T = 40$ (not shown in the figure), reducing the number of pilots to $T = 15$ under HNB-GNN2-loc increases the gap to the BCD-perfect-CSI benchmark by less than 3%. As shown in Fig. 4.5, further reducing the number of pilots from $T = 15$ to $T = 5$ under the proposed HNB-GNN2-loc leads to a moderate increase in the relative performance gap to the BCD-perfect-CSI benchmark, from 8% to 14%. Conversely, when the HNB-GNN2 scheme employs user pilots only, the relative performance gap is increased from 13% to 33%. Nevertheless, as illustrated in Fig. 4.5, the proposed HNB scheme with user location outperforms the static-GNN3-loc scheme even when a higher number of pilots is employed by the static scheme. Hence, by dynamically leveraging pilots (and possibly user location information), the HNB schemes require significantly fewer pilots per user to outperform static learning schemes and operate close to the BCD-perfect-CSI benchmark.

Impact of Rician Factor

Fig. 4.6 shows the testing WSR performance achieved by the HNB-GNN2-loc scheme and static-GNN3-loc scheme for different Rician factors associated with the RIS links and for $T = 5$ and $T = 10$ pilots per user. For the numerical results in Fig. 4.6, in addition to the user weights and user locations, one pilot per user was also fed to the hypernetwork input as this improved the spectral efficiency performance associated with the HNB schemes.

As captured by the trend in Fig. 4.6, both HNB-GNN2-loc and static-GNN3-loc schemes exploit higher Rician factors to reduce the gap to the BCD-perfect-CSI scheme. For example, the static-GNN3-loc scheme with $T = 10$, increasing the Rician factor from 4 dB to 10 dB results in a decrease of the relative performance gap (to BCD-perfect-CSI) from 41% to 32%. Similarly, the relative performance gap associated with the HNB-GNN2-loc scheme decreases from 17% to 11% as the Rician factor increases from 4 dB to 10 dB. This follows as higher Rician factors correspond to stronger LOS patterns captured by the beamforming schemes during training.

Fig. 4.6 also demonstrates that the proposed HNB scheme outperforms the static schemes for the considered Rician factors. For example, for $T = 5$, the static-GNN3-loc scheme results in relative performance gaps of 47% and 36% for Rician factors of 4 dB and 10 dB, respectively. This is in contrast to the relative performance gaps of 20% and 14% associated with HNB-GNN2-loc schemes for Rician factors of 4 dB and 10 dB, respectively. Subsequently, the proposed HNB schemes achieve substantial spectral effi-

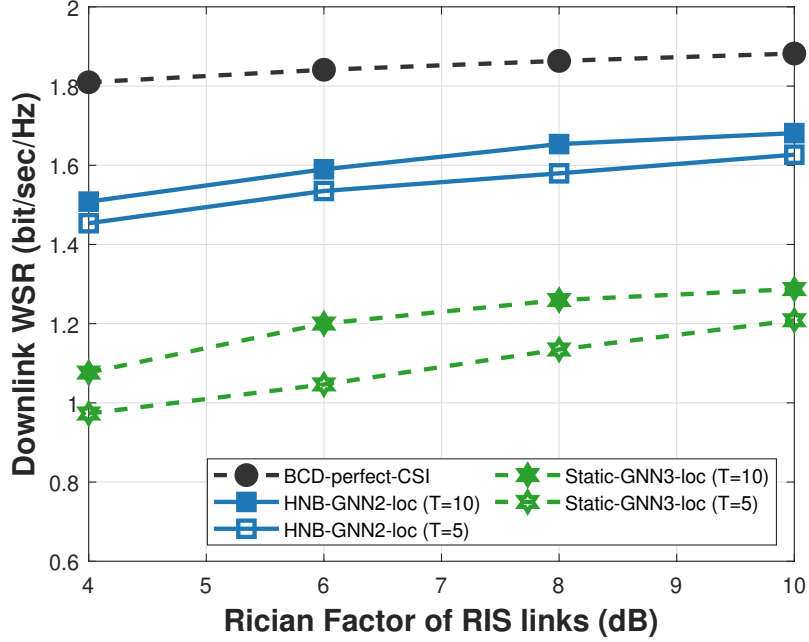


Figure 4.6: Average downlink WSR on testing data for different Rician Factors and for different number of pilots with location information, under independent user weights.

ciency performance gains over static learning schemes for different BFN architecture, user pilots/user location availability scenarios and/or Rician factors.

Robustness to changes in UL/DL noise powers and Rician Factors

In Fig. 4.7, numerical experiments are conducted to demonstrate the robustness of the HNB scheme in terms of generalizing to data generated using different characteristics or parameters. Particularly, Fig. 4.7a (Fig. 4.7b, respectively) shows the performance of the *HNB-GNN2-loc* scheme, where the HNB is individually retrained for each UL noise power (DL noise power, respectively) level and tested under the same conditions. Additionally, the *HNB-GNN2-loc-default-training* scheme represents a single HNB model trained on a default UL noise power of -92 dBm (default DL noise power of -88 dBm, respectively) but tested across different noise power conditions. For comparison, the optimistic *BCD-with-perfect-CSI* benchmark is shown in black. As illustrated, the *HNB-GNN2-loc* scheme continues to perform close to the optimistic BCD-perfect-CSI benchmark across various UL noise powers (DL noise powers, respectively). Additionally, the *HNB-GNN2-loc-default-training* scheme consistently performs well across across different noise power conditions.

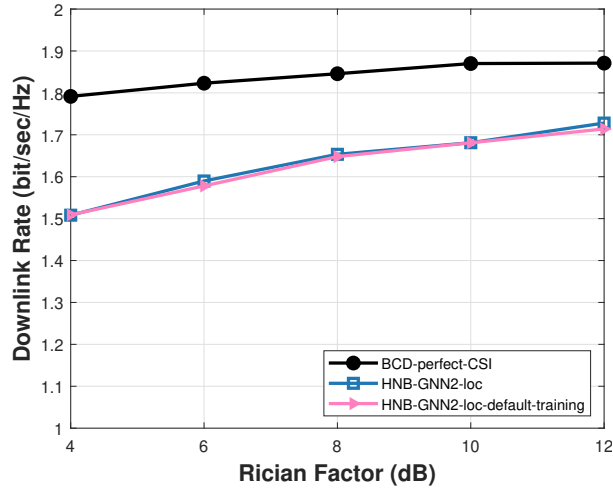
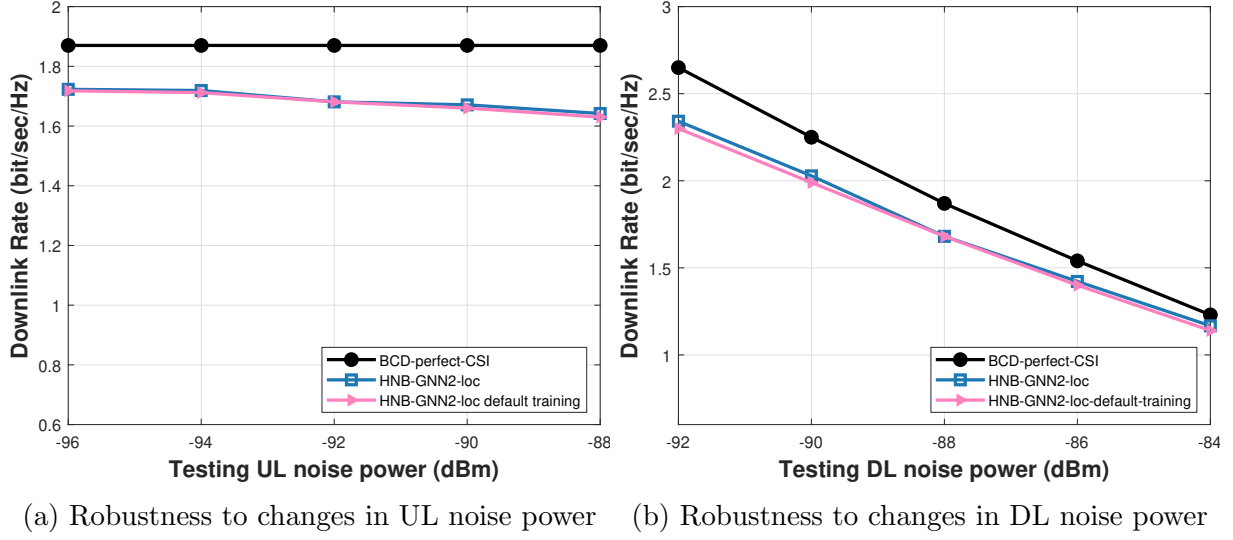


Figure 4.7: Average downlink WSR on testing data for different UL noise powers, DL noise power and Rician factors. Here, $T = 10$ with location information and independent user weights are used.

This demonstrates that the HNB scheme is robust to the case where the testing UL noise (DL noise, respectively) power setting is different from the one used during training. Furthermore, similar results were obtained to demonstrate the robustness of the proposed HNB schemes to different Rician factors as shown in Fig. 4.7c.

4.6.4 Optimizing WSR for Proportional-Fairness

In this section, the proposed HNB schemes are used to assign resources while maintaining fairness based on proportional-fairness. Specifically, we consider data transmissions over S time-slots indexed as $s \in \mathcal{S} = \{1, \dots, S\}$. For each user $k \in \mathcal{K}$, the history of previous data transmissions is reflected by an exponentially weighted moving average factor $\Lambda_k^s = (1 - \lambda)\Lambda_k^{s-1} + \lambda R_k^{s-1}$, where R_k^s is the data rate associated with user k in slot s and λ is a forgetting factor.

Following [19, 4], the proportional-fair (PF) maximization problem in time-slot s can be approximated by maximizing the WSR objective given by $\sum_{k=1}^K \Gamma_k^s R_k^s$, where the user weights $\Gamma_k^s = \nu_s / \Lambda_k^s$, and $\nu_s > 0$ is chosen such that $\sum_{k \in \mathcal{K}} \Gamma_k^s = 1$. Moreover, the number of time-slots is set to $S = 300$ and $\lambda = 0.01$. For each realization, the channel is generated according to the models in Section 4.6.1 with the large scale parameters being fixed across different slots and small scale fading independently generated for each slot.

Training and Testing of proposed HNB and Alternate Schemes

The beamforming learning schemes of Section 4.5 were trained using TensorFlow with Adam optimizer and the hypernetwork training of [41]. The training of the beamforming learning schemes was performed over 100, 50, 10 and 5 time-slots. It was found that learning the beamforming schemes over 10 time-slots was sufficient to capture the distribution of PF user weights while significantly reducing training complexity. For each time-slot, the beamforming learning schemes were trained for 500 epochs of 100 iterations. The initial learning rate was set to 1×10^{-4} with a decay rate of 0.997 for each 2000 iterations. For each iteration, the learning parameters of the beamforming schemes were updated using 400 training samples. The performance of the trained models and BCD-perfect-CSI benchmark was then assessed by averaging the achievable normalized WSR over 1000 new independently generated test samples evaluated at $s = 300$.

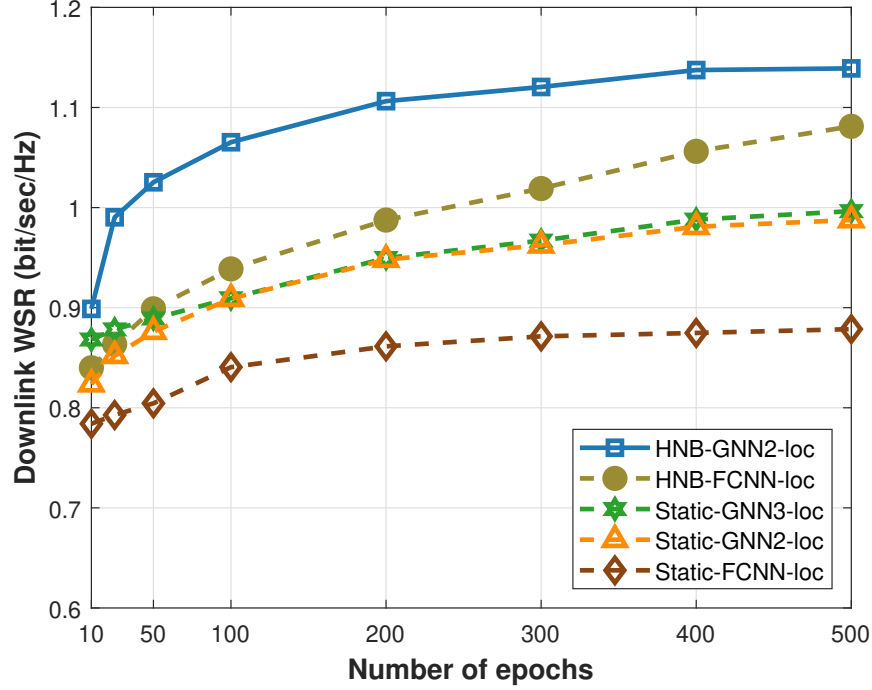


Figure 4.8: Average downlink WSR on testing data after different number of training epochs with $T = 5$ and location information, obtained under a Rician factor of 10 dB, PF user weights and for $s = 300$.

Impact of BFN Training and Architecture

Fig. 4.8 shows the testing WSR achieved by different beamforming learning schemes, trained with $T = 5$ pilots per user and location information, against the number of training epochs. From the figure, the fully-trained HNB-GNN2-loc scheme outperforms the static-GNN3-loc, static-GNN2-loc and static-FCNN-loc schemes by 14%, 15%, and 31%, respectively. Similarly, the fully-trained HNB-FCNN-loc scheme outperforms the static-GNN3-loc, static-GNN2-loc and static-FCNN-loc schemes by 8%, 9% and 25%. Subsequently, the proposed HNB schemes achieve substantial performance improvements over the static schemes, for both GNN and FCNN architectures.

Fig. 4.8 also shows that, similar to the results in Fig. 4.4, the HNB schemes trained with only a few epochs outperform the fully-trained static schemes. Particularly, the HNB-GNN2-loc schemes requires only 25 training epochs to achieve performance improvements over the corresponding fully-trained static schemes trained with 500 training epochs.

Impact of user pilot and location information availability

Fig. 4.9 shows the testing WSR achieved by the proposed HNB-GNN2 scheme and the alternate static-GNN3 scheme, for different number of pilots and depending on whether user location information is available. For the scenarios considered in Fig. 4.9, the HNB-GNN2-loc schemes bring significant improvements over the corresponding static schemes. For example, under $T = 3$ and using location information, the HNB-GNN2-loc scheme reduces the BCD-perfect-CSI relative performance gap to 18% compared to the static-GNN3-loc scheme with a 31% gap. Similarly for $T = 5$ and location information, compared to static-GNN3-loc scheme, the HNB-GNN2-loc scheme reduces the BCD-perfect-CSI relative performance gap from 27% to 16%.

Similar to the results in Section 4.6.3, the proposed HNB-GNN2-loc scheme effectively exploits pilots information along with location information to reduce the required number of pilots. For instance, compared to the HNB-GNN2 scheme, the HNB-GNN2-loc scheme halves the number of required pilots to achieve a WSR performance of 1.2 bit/sec/Hz. Additionally, with HNB-GNN2-loc, reducing the number of pilots available from $T = 20$ to $T = 3$ results in increasing the relative performance gap to the BCD-perfect-CSI benchmark from 8% to 18%. Conversely, for static-GNN3-loc, the performance gap increases from 12% to 31%. Consequently, compared to static-GNN3-loc scheme, HNB-GNN2-loc provides more flexibility in trading off spectral efficiency performance with pilot overhead. Moreover, Fig. 4.9 shows that the static-GNN3-loc scheme with $T = 20$ pilots per user could operate close to the proposed HNB-GNN2 schemes. However, even under this special case, the proposed HNB-GNN2-loc scheme requires only $T = 10$ pilots per user to approach the best performance achieved by the static-GNN3-loc scheme when fed with $T = 20$ pilots per user.

4.6.5 Complexity Analysis and Comparisons

The complexity of different beamforming learning scheme is now described in terms of the number of operations required to generate optimized beamforming configurations. Using the information in Table 4.1 and Section 4.5, the complexities of the static-GNN3 benchmark scheme (which achieves the best performance among the beamforming learning benchmarks) and the two proposed HNB schemes are as follows:

- *Static-GNN3 benchmark*: Each GNN layer consists of $K + 1$ nodes (i.e., K user nodes and 1 RIS node), with each node containing networks that process information from the K neighboring nodes as well as information from the previous layer. Consequently, the

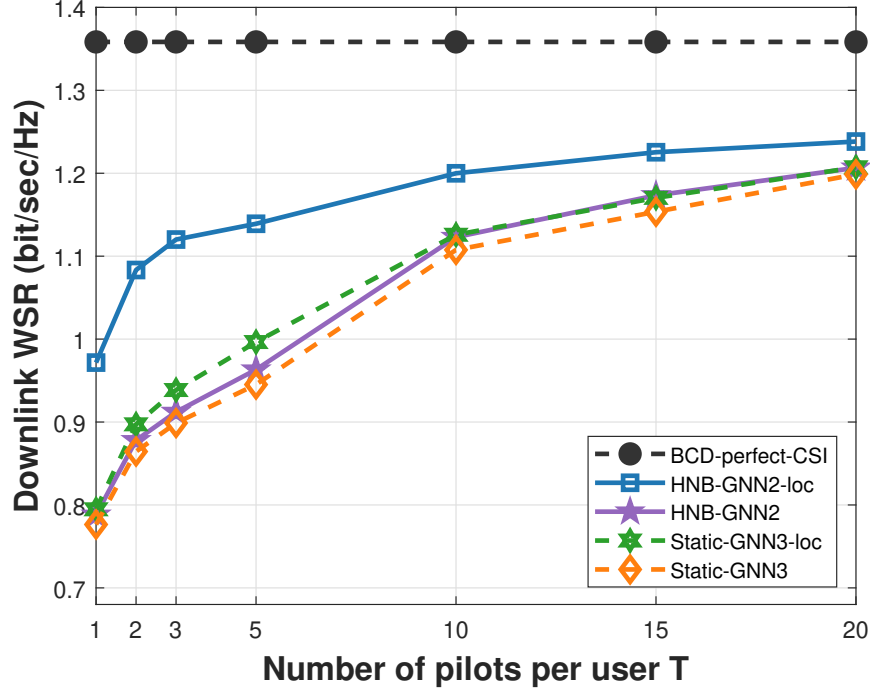


Figure 4.9: Average downlink WSR for different number of pilots per user T , under PF user weights and for $s = 300$, with user location (e.g., HNB-GNN2-loc) and without user location (e.g., HNB-GNN2).

complexity of a single GNN layer is $\mathcal{O}((K+1)^2(\beta_{a,1} + \beta_{a,2} + \beta_{a,3} + \beta_{c,1} + \beta_{c,2}))$. Combining the complexity of $L_G = 3$ GNN layers with that of input and output processing, the complexity of the static-GNN3 benchmark is $\mathcal{O}(K\beta_{i,1} + \beta_{i,2} + 3(K+1)^2(\beta_{a,1} + \beta_{a,2} + \beta_{a,3} + \beta_{c,1} + \beta_{c,2}) + K\beta_{o,1} + \beta_{o,2})$.

- *Proposed Scheme 1- HNB-FCNN*: The complexity is $\mathcal{O}(\rho_i + \rho_h + \rho_o + \zeta_i + \zeta_h + \zeta_o)$. Here, $\zeta_i = 300H$, $\zeta_h = 300 \cdot 200$, $\zeta_o = 200\rho_h$. Moreover, ρ_i , ρ_h , ρ_o are given in Table 4.1. Depending on the input to the hypernetwork, $H = K$ when only user weights are used or $H = 4K$ if both the user weights and user locations are used.

- *Proposed Scheme 2- HNB-GNN2*: Let $\alpha_i = 300H$, $\alpha_h = 300 \cdot 200$ and $\alpha_o = 200(2\chi_{c,1} + 2\chi_{c,2})$. With the information provided in Table 4.1, the complexity is $\mathcal{O}(K\chi_{i,1} + \chi_{i,2} + 2(K+1)^2(\chi_{a,1} + \chi_{a,2} + \chi_{a,3} + \chi_{c,1} + \chi_{c,2}) + K\chi_{o,1} + \chi_{o,2} + \alpha_i + \alpha_h + \alpha_o)$.

We emphasize that, in practice, it was found that an HNB scheme does not significantly increase training/testing time compared to the corresponding static benchmark with the

same BFN. This is because the majority of additional HNB complexity comes from a single operation at the output layer of the hypernetwork and is highly parallelizable. For illustration, we provide relative performance comparisons between the HNB schemes and static learning benchmarks in terms of the training and testing time. Specifically, for the scenario considered in Fig. 4.4 and using an NVIDIA RTX 3090 GPU with a TensorFlow implementation, the training times for the proposed HNB-GNN2 and HNB-FCNN schemes were 25.95 hours and 25.89 hours, respectively. Meanwhile, the training times for the static-GNN3, static-GNN2, and static-FCNN were 24.39 hours, 22.57 hours, and 21.92 hours, respectively.

In terms of runtime, testing the performance on a single realization required 1.4 ms, 2.1 ms and 4.6 ms for the static-FCNN, static-GNN2, and static-GNN3 benchmarks, respectively. Conversely, the proposed HNB-FCNN and HNB-GNN2 required 1.7 ms and 2.8 ms, respectively. This demonstrates that, compared to the corresponding static learning benchmarks using the same BFN, an HNB schemes do not significantly increase runtime.

4.7 Conclusions

This chapter has investigated the problem of learning beamforming schemes that leverage user pilots to maximize the weighted sum-rate in multi-user RIS systems while accommodating different service priorities and accounting for fairness via user weights. Subsequently, an adaptive beamforming configuration framework was developed to exploit noisy pilots and user weights to generate the base station active beamforming vectors and RIS passive beamforming reflection coefficients that optimize the downlink weighted sum-rate. The formulated adaptive beamforming configuration problem is analyzed and, under mild regularity conditions, shown to attain a maximum. To dynamically approximate solutions to the formulated beamforming configuration problem, an HNB framework is proposed.

Under the proposed HNB framework, a BFN is designed to leverage pilot information and user weights to generate the beamforming configuration of the RIS system. Rather than directly learning a single BFN, a hypernetwork was trained to dynamically generate BFN learning parameters based on an input conditioning vector. When the conditioning vector is the user weights, a single trained HNB is able to tune the learning parameters of the BFN to the user weights without the need for retraining the BFN. Additionally, by leveraging the hypernetwork's conditioning vector to tune the BFN to user location information (if available), the HNB scheme can significantly reduce the number of pilots needed to configure the RIS system.

Numerical results have corroborated the efficacy of the proposed HNB schemes in achieving significant performance improvements over static learning schemes that directly train a single BFN to work for all possible user weights. For different user weighting strategies, the proposed HNB schemes were shown to achieve significant performance improvements over corresponding fully-trained static schemes. Furthermore, it was illustrated that the proposed HNB schemes can exploit a small number of user pilots to operate close to an optimistic BCD with perfect CSI benchmark.

Chapter 5

Practical Configuration of a Discrete RIS with Phase-dependent Amplitude Response

5.1 Introduction

To relax hardware constraints and enable RIS implementation with diverse technologies, in this chapter we consider the case where an RIS is configured with discrete phase shifts. Moreover, depending on the RIS implementation and as illustrated by the transmission line model of [3], the amplitude response associated with an RIS element can in general be non-linearly dependent on the phase shift introduced by the element. Hence, different phase shifts may result in different reflection amplitudes.

To address the discrete RIS problem, one possible approach is to relax the discrete formulation into a continuous one, optimize the relaxed problem, and subsequently discretize the obtained continuous phase shifts. With sufficiently fine resolution, a discrete RIS configuration obtained in this manner can potentially approximate the performance of its continuous counterpart. This is supported by analytical results in [131, 124], which demonstrate that discretizing continuous phase shifts with fine resolution (i.e., 4-bit) achieves near-continuous performance under certain conditions.

The objective of this chapter is to develop an RIS configuration framework that effectively supports implementations with lower resolution, specifically using binary (1-bit) or 2-bit phase shifts. In this context, extensive numerical experiments demonstrate that directly addressing the discrete problem can yield substantial gains compared to continuous

relaxation followed by discretization approach. Conversely, for different scenarios considered and even in the presence of significant amplitude selectivity, numerical results confirm that 3- or 4-bit discretization of continuous phase shifts allows the performance to closely approach that of the continuous case.

In this chapter, the discrete RIS configuration problem with phase-dependent amplitude response is formulated as a constrained virtual-channel selection problem. It is demonstrated that the configuration problem, when restricted to subsets of the virtual channels where the maximum phase-variation is $\theta_{\max} < \pi$, can be bounded between monotone upper bound (UB) and lower bound (LB) matroid-constrained problems. This in turn motivates the use of a greedy configuration framework, as greedy algorithms are well known for their computational efficiency and performance guarantees in matroid settings [29].

We illustrate discrete RIS configuration problems whose optimal solution satisfy this maximum phase-variation property. Subsequently, under the maximum phase-variation property, performance guarantees are obtained using a low-complexity greedy configuration scheme. Motivated by this, a practical low-complexity configuration framework is proposed to optimize general RIS problems with discrete phase shifts and where phase shifts result in phase-dependent amplitude response.

In [3], a block coordinate descent (BCD) algorithm was proposed to configure an RIS with continuous phase shifts but with phase-dependent amplitude response. On the other hand, configuring the discrete phase shifts of an RIS under a perfect-reflection RIS configuration model was addressed in works such as [89, 88]. Moreover, in [88], the special case of configuring an N -element RIS with binary phase shifts of different amplitudes was considered in a single antenna setting. However, this technique does not work for (general) complex-valued discrete phase shifts configuration problems. In contrast to the aforementioned works, the goal of this chapter is to develop practical schemes that configure an RIS with (i) $L \geq 2$ discrete phase shift levels and (ii) where the reflection coefficients have non-linear phase-dependent amplitude response.

5.1.1 Contributions and Key Findings:

The contributions and key findings of this chapter are summarized as follows:

- We consider configuring an RIS with phase shifts that (i) are discrete and (ii) exhibit non-linear phase-dependent amplitude response. Subsequently, the discrete RIS configuration problem with phase-dependent amplitude response is formulated as a constrained virtual-channel selection problem.

- The formulated configuration problem, considered over a subset of the virtual channels where the maximum phase-variation is $\theta_{\max} < \pi$, is shown to be bounded between monotone upper bound (UB) and lower bound (LB) matroid-constrained problems. Moreover, when the optimal RIS configuration lies within such subsets, it is shown that optimizing either the UB or LB yields a configuration that guarantees a worst-case optimality gap of $|2 \log_2 \cos(\theta_{\max}/2)|$ bit/sec/Hz, uniformly over all system parameters. Discrete RIS configuration problems whose optimal solution lies within such maximum phase-variation subsets are demonstrated.
- Motivated by the results above, a low-complexity configuration framework is proposed to optimize general RIS problems where the discrete phase shifts are specified by k bits and exhibit an arbitrary phase-dependent amplitude response. Moreover, the complexity of the proposed scheme is analyzed and compared to the considered benchmarks.
- Numerical experiments demonstrate the efficacy of the proposed RIS configuration scheme. For instance, for single antenna case and under binary phase shifts with significant amplitude-response variations, the proposed framework performs within 4% of the optimal configuration. Furthermore, the proposed scheme requires only 3 control bits per RIS element to achieve more than 90% of the performance associated with a benchmark scheme that uses continuous phase shifts and a numerical solver.
- Additionally, the proposed scheme is shown to either significantly outperform the commonly adopted closest point projection (CPP) benchmark (that discretizes the continuous phase shifts from a block-coordinate descent algorithm), or at least achieve similar performance. Furthermore, even in the cases where the proposed scheme and the CPP benchmark provide similar performance, it may be preferable to adopt the proposed scheme due to its practical gains in terms of time complexity and simplicity of the operations involved.

5.2 Discrete RIS Configuration: Problem Formulation and Analysis

We first consider a system where an N -element RIS is configured to assist signal transmissions between a single-antenna transmitter (TX) and a single-antenna receiver (RX).¹ We

¹Depending on whether the TX is the base station or the user, this setting can represent either a downlink or an uplink scenario.

note that a multi-antenna extension of this scenario will be considered in Section 5.4.

An RIS element $n \in \mathcal{N} = \{1, \dots, N\}$ introduces a reflection coefficient $\alpha(\varphi_n)e^{-j\varphi_n}$ where $\varphi_n \in [0, 2\pi)$ is the phase shift introduced by the n^{th} element and $\alpha(\varphi_n) > 0$ is the corresponding amplitude. Under a discrete model where each RIS element is specified by k bits, the RIS phase shifts are constrained to $L = 2^k$ possible discrete values $\{\phi_1, \dots, \phi_L\}$. Thus, the n^{th} element of the RIS is configured with reflection coefficient $\Gamma_n \in \{\gamma_1, \dots, \gamma_L\}$ where $\gamma_\ell \triangleq \alpha(\phi_\ell)e^{-j\phi_\ell}$.

Under a block fading model, we denote the direct TX-RX channel as $h_{\text{TR}} \in \mathbb{C}$. Moreover, the TX-RIS channel is indicated by $\mathbf{f} \in \mathbb{C}^{N \times 1}$ and the RIS-RX channel is denoted by $\mathbf{g} \in \mathbb{C}^{1 \times N}$. Here, the channels are assumed to be known. Indeed, acquiring the CSI information has been the focus of many research works including [134, 10] as well as the works reviewed in Section 2.3.

With an RIS configuration $\mathbf{\Gamma} = [\Gamma_1, \dots, \Gamma_N]$, the effective TX-RX channel is given by $h_{\text{TR}} + \sum_{n=1}^N g_n f_n \Gamma_n$. Subsequently, the RIS configuration problem associated with maximizing the rate is

$$\begin{aligned} & \max_{\mathbf{\Gamma}=[\Gamma_1, \dots, \Gamma_N]} \log_2 \left(1 + \frac{P|h_{\text{TR}} + \sum_{n=1}^N g_n f_n \Gamma_n|^2}{\sigma^2} \right) \\ & \text{subject to} \quad \Gamma_n \in \{\gamma_1, \dots, \gamma_L\}. \end{aligned} \tag{P1}$$

Here, P is the power of the signal transmitted by the TX and the additive white Gaussian noise (AWGN) at the RX is distributed as $\mathcal{CN}(0, \sigma^2)$.

5.2.1 Virtual-antenna Selection Formulation

Rather than relaxing the discrete optimization problem in (P1), we investigate the problem to obtain worst-case performance guarantees under a low-complexity configuration scheme. Towards this goal, the RIS configuration problem in (P1) is cast as a virtual-channel selection problem.

We note that, for each RIS element $n \in \mathcal{N}$, there are L possible choices of the reflection coefficients $\Gamma_n \in \{\gamma_1, \dots, \gamma_L\}$ where $\gamma_1 = \alpha(\phi_1)e^{-j\phi_1}, \dots, \gamma_L = \alpha(\phi_L)e^{-j\phi_L}$. This results in L possible effective channels

$$\tilde{h}_{n,\ell} = g_n f_n \gamma_\ell, \quad n \in \mathcal{N}, \ell \in \{1, \dots, L\}, \tag{5.1}$$

of $g_n f_n \gamma_\ell$ in (P1) and these are called virtual-channels as only *one* will be selected.

Algorithm 2 A greedy algorithm for configuring an RIS

Input: $\Sigma = (\mathcal{G}, \mathcal{I}_{\mathcal{G}})$, set function $F : 2^{\mathcal{G}} \rightarrow \mathbb{R}^+$
 $\mathcal{C}_{\mathcal{I}} \leftarrow \emptyset$
while $\exists e \in \mathcal{G}$ such that $e \notin \mathcal{C}_{\mathcal{I}}$ and $\mathcal{C}_{\mathcal{I}} \cup \{e\} \in \mathcal{I}_{\mathcal{G}}$ **do**
 $e \leftarrow \arg \max_{\substack{e \in \mathcal{G}, e \notin \mathcal{C}_{\mathcal{I}} \\ \mathcal{C}_{\mathcal{I}} \cup \{e\} \in \mathcal{I}_{\mathcal{G}}}} \{F(\mathcal{C}_{\mathcal{I}} \cup \{e\}) - F(\mathcal{C}_{\mathcal{I}})\}$
 $\mathcal{C}_{\mathcal{I}} \leftarrow \mathcal{C}_{\mathcal{I}} \cup \{e\}$
end while
Output: $\mathcal{C}_{\mathcal{I}}$

Let $\mathcal{V}_0 = \{h_{\text{TR}}\}$, and for $n \in \{1, \dots, N\}$, let $\mathcal{V}_n = \{\tilde{h}_{n,1}, \dots, \tilde{h}_{n,L}\}$ indicate the subset of virtual-channels associated with the n^{th} RIS element. The set of all virtual-channels is given by $\mathcal{V} \triangleq \bigcup_{n=0}^N \mathcal{V}_n$ and $|\mathcal{V}| = NL+1$. Then, the RIS configuration problem corresponds to selecting a subset $\mathcal{C} = \{h_0, \dots, h_N\}$, where $h_n \in \mathcal{V}_n$ for $n = 0, \dots, N$. Subsequently, over a family of sets $\mathcal{F}_{\mathcal{V}} \triangleq \{\mathcal{C} \subseteq \mathcal{V} : |\mathcal{C} \cap \mathcal{V}_n| = 1, \forall n = 0, \dots, N\}$, maximizing (P1) is equivalent to

$$\max_{\mathcal{C} \in \mathcal{F}_{\mathcal{V}}} R(\mathcal{C}), \quad (\text{P2})$$

where $R(\mathcal{C}) \triangleq \log_2 (1 + P |\sum_{h \in \mathcal{C}} h|^2 / \sigma^2)$ is a set function.

Under the formulated constrained RIS configuration problem (P2), we propose to select the RIS reflection coefficients using the low complexity greedy algorithm shown in Algorithm 2. For given input ground set $\mathcal{G}$ and a family of subsets $\mathcal{I}_{\mathcal{G}}$ of $\mathcal{G}$, Algorithm 2 greedily optimizes a given set function $F : 2^{\mathcal{G}} \rightarrow \mathbb{R}^+$, over the family of subsets $\mathcal{I}_{\mathcal{G}}$.

5.2.2 Analysis and a Performance Guarantee

To gain insights into the efficacy of Algorithm 2 in optimizing the RIS configuration, we first introduce the notion of a slice of virtual-channels. A $(\theta_c, \theta_{\max})$ slice is a subset $\mathcal{S}(\theta_c, \theta_{\max}) = \mathcal{V} \cap \mathcal{U}(\theta_c, \theta_{\max})$, where $\mathcal{U}(\theta_c, \theta_{\max}) = \{z \in \mathbb{C} : z = re^{j\theta} \text{ where } r \geq 0, |\theta - \theta_c| \leq \theta_{\max}/2\}$. Restricting (P2) to a slice $\mathcal{S}(\theta_c, \theta_{\max})$ with $\theta_{\max} < \pi$ may be reasonable in many cases since coherent combining requires selecting virtual-channels with similar phases.

Using a simplified example, Fig. 5.1 illustrates the concept that the selected virtual channels should lie within a slice. Indeed, Proposition 5.1 below shows that under mild assumptions, except for possibly the effective channel associated with the direct path, the optimal effective channels all lie in some slice $\mathcal{S}(\theta_c, \theta_{\max})$ with $\theta_{\max} < \pi$.

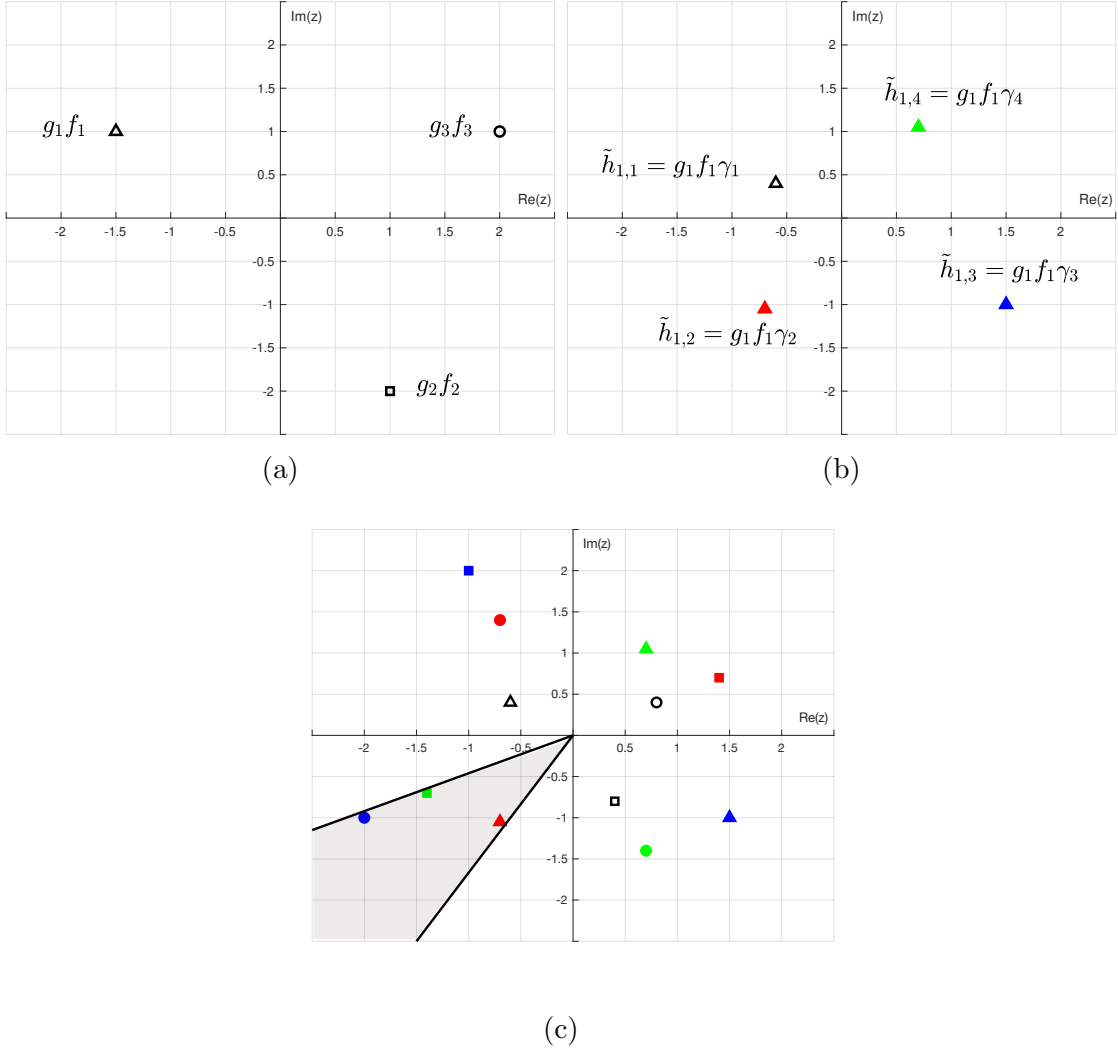


Figure 5.1: This example illustrates a scenario involving 3 RIS elements and no direct TX – RX path, where in (a) each element’s channel is represented by a triangle, square, or circle. Each RIS element introduces one of four possible phase shifts, identified by different colors: black, green, blue, or red. In (b), the possible virtual channels associated with the first RIS element are shown. In (c), the 12 possible virtual channels associated with the 3 elements are illustrated. To coherently align the reflected channels, the selected effective channels should lie within a slice of the complex plane as shown in (c).

Proposition 5.1. *Consider the RIS problem (P2) with $L \geq 4$ and phase shifts $\phi_\ell = 2\pi(\ell - 1)/L$. Then, for an optimal configuration $\mathcal{C}^{\text{OPT}} \triangleq \{h_0^{\text{OPT}}, h_1^{\text{OPT}}, \dots, h_N^{\text{OPT}}\}$, we have that $\{h_1^{\text{OPT}}, \dots, h_N^{\text{OPT}}\} \subseteq \mathcal{S}(\theta_c, \theta_{\max})$ for some θ_c and $\theta_{\max} < \pi$.*

The proof is given in Appendix A.4.

We conclude that for an RIS system that satisfies the assumptions of Proposition 5.1 and has a blocked TX-RX direct path (i.e., $h_{\text{TR}} = 0$), then the optimal RIS configuration $\mathcal{C}^{\text{OPT}}$ of (P2) lies in some slice $\mathcal{S}(\theta_c, \theta_{\max})$ with $\theta_{\max} < \pi$.

Over a slice $\mathcal{S}(\theta_c, \theta_{\max})$, let $\mathcal{F}_S \triangleq \{\mathcal{C} \subseteq \mathcal{S}(\theta_c, \theta_{\max}) : |\mathcal{C} \cap \mathcal{V}_n| = 1, \forall n = 0, \dots, N\}$ and $\mathcal{I}_S \triangleq \{\mathcal{C} \subseteq \mathcal{S}(\theta_c, \theta_{\max}) : |\mathcal{C} \cap \mathcal{V}_n| \leq 1, \forall n = 0, \dots, N\}$. Proposition 5.2 below demonstrates that, when the RIS configuration problem (P2) is considered over a slice $\mathcal{S}(\theta_c, \theta_{\max})$ with $\theta_{\max} < \pi$, the rate $R(\mathcal{C})$ of (P2) can be closely bounded between an upper bound

$$f_u(\mathcal{C}) \triangleq \log_2 \left(1 + P \left(\sum_{h \in \mathcal{C}} |h|^2 / \sigma^2 \right) \right) \quad (5.2)$$

and a lower bound

$$f_l(\mathcal{C}) \triangleq \log_2 \left(1 + P \left(\sum_{h \in \mathcal{C}} \cos(\theta_{\max}/2) |h|^2 / \sigma^2 \right) \right). \quad (5.3)$$

Moreover, if $\mathcal{S}(\theta_c, \theta_{\max})$ contains a feasible solution to (P2), the optimization of both $\max_{\mathcal{C} \in \mathcal{F}_S} f_u(\mathcal{C})$ and $\max_{\mathcal{C} \in \mathcal{F}_S} f_l(\mathcal{C})$ is equivalent to maximizing a modular function subject to a matroid constraint.

Proposition 5.2. *Let a slice $\mathcal{S}(\theta_c, \theta_{\max})$ have $\theta_{\max} < \pi$. Then*

1. *Over any subset $\mathcal{C} \subseteq \mathcal{S}(\theta_c, \theta_{\max})$, the set function $R(\mathcal{C})$ of (P2), can be bounded as*

$$f_l(\mathcal{C}) \leq R(\mathcal{C}) \leq f_u(\mathcal{C}). \quad (5.4)$$

Here, $f_u(\mathcal{C})$ and $f_l(\mathcal{C})$ are monotone, i.e., for any $A \subseteq B \subseteq \mathcal{S}(\theta_c, \theta_{\max})$, $f_u(A) \leq f_u(B)$. Additionally, if $\theta_{\max} \leq \pi/2$, then $R(\mathcal{C})$ is also monotone.

2. *If $\mathcal{S}(\theta_c, \theta_{\max})$ contains a feasible solution of (P2), i.e., $|\mathcal{F}_S| > 0$, then there exists $\tilde{\mathcal{C}}_S \in \mathcal{F}_S$ such that*

$$\begin{aligned} \max_{\mathcal{C} \in \mathcal{I}_S} f_l(\mathcal{C}) &= f_l(\tilde{\mathcal{C}}_S) \leq R(\tilde{\mathcal{C}}_S) \leq \max_{\mathcal{C} \in \mathcal{F}_S} R(\mathcal{C}) \\ &\leq \max_{\mathcal{C} \in \mathcal{I}_S} f_u(\mathcal{C}) = f_u(\tilde{\mathcal{C}}_S), \end{aligned} \quad (5.5)$$

and $\tilde{\mathcal{C}}_S$ can be obtained by applying Algorithm 2 to either $\max_{\mathcal{C} \in \mathcal{I}_S} f_l(\mathcal{C})$ or $\max_{\mathcal{C} \in \mathcal{I}_S} f_u(\mathcal{C})$. Moreover, $\mathcal{I}_S$ are the independent sets of a matroid $(S, \mathcal{I}_S)$.

The proof is given in Appendix A.5.

Under the appropriate conditions, the next corollary provides a worst-case guarantee on the performance of Algorithm 2 with respect to an optimal RIS configuration.

Corollary 5.1. *If an optimal RIS configuration is contained within a slice $\mathcal{S}(\theta_c, \theta_{\max})$ with $\theta_{\max} < \pi$, then Algorithm 2 applied with $(\mathcal{S}(\theta_c, \theta_{\max}), \mathcal{I}_S)$ and $f_u(\mathcal{C})$ produces an RIS configuration $\tilde{\mathcal{C}}_S$ that is within $|2 \log_2 \cos(\theta_{\max}/2)|$ bit/sec/Hz of the optimal objective of (P2).*

Proof. The slice $\mathcal{S}(\theta_c, \theta_{\max})$ contains a feasible RIS configuration and hence by (5.5) of Proposition 5.2, the RIS configuration $\tilde{\mathcal{C}}_S$ has an optimality gap of

$$|\max_{\mathcal{C} \in \mathcal{F}_S} R(\mathcal{C}) - R(\tilde{\mathcal{C}}_S)| \leq |f_u(\tilde{\mathcal{C}}_S) - f_l(\tilde{\mathcal{C}}_S)| \leq |2 \log_2 \cos(\theta_{\max}/2)|.$$

□

Example of a Configuration Problem

Now, consider an RIS system that satisfies the assumptions of Proposition 5.1 (i.e., $L \geq 4$ and phase shifts $\phi_\ell = 2\pi(\ell - 1)/L$), has equal gains $\alpha(\phi_1) = \dots = \alpha(\phi_L) > 0$ and $h_{\text{TR}} \neq 0$. Then for an optimal RIS configuration $\mathcal{C}^{\text{OPT}} = \{h_0^{\text{OPT}}, h_1^{\text{OPT}}, \dots, h_N^{\text{OPT}}\}$, it must hold that

$$\left| \arg \left(\bar{h}_n^{\text{OPT}} \sum_{\substack{m=0 \\ m \neq n}}^N h_m^{\text{OPT}} \right) \right| \leq \pi/L, \quad n \in \{0, \dots, N\}, \quad (5.6)$$

where $\bar{z}$ is the complex conjugate of z . By extending the method of [88, Lemma 1], we first show that (5.6) holds for $n = 0$. Suppose that

$$\left| \arg \left(\bar{h}_0^{\text{OPT}} \sum_{m=1}^N h_m^{\text{OPT}} \right) \right| > \pi/L.$$

Then there is an $l' \in \mathbb{Z}$ such that $|\arg(\bar{h}_0^{\text{OPT}} \sum_{m=1}^N e^{j2\pi l'/L} h_m^{\text{OPT}})| \leq \pi/L$ and therefore

$$\left| h_0^{\text{OPT}} + \sum_{m=1}^N h_m^{\text{OPT}} \right| < \left| h_0^{\text{OPT}} + \sum_{m=1}^N e^{j2\pi l'/L} h_m^{\text{OPT}} \right|,$$

contradicting the optimality of $\mathcal{C}^{\text{OPT}}$.

Next, for $n \neq 0$, suppose that for some $n' \in \{1, \dots, N\}$,

$$\left| \arg \left(\bar{h}_{n'}^{\text{OPT}} \sum_{\substack{m \neq n' \\ m=0}}^N h_m^{\text{OPT}} \right) \right| > \pi/L.$$

Then there exists an $l'' \in \mathbb{Z}$ such that $|\arg(e^{-j2\pi l''/L} \bar{h}_{n'}^{\text{OPT}} \sum_{m \neq n'} h_m^{\text{OPT}})| \leq \pi/L$, and

$$\left| \sum_{\substack{m \neq n' \\ m=0}}^N h_m^{\text{OPT}} + h_{n'}^{\text{OPT}} \right| < \left| \sum_{\substack{m \neq n' \\ m=0}}^N h_m^{\text{OPT}} + e^{j2\pi l''/L} h_{n'}^{\text{OPT}} \right|,$$

which contradicts the optimality of $\mathcal{C}^{\text{OPT}}$.

Subsequently, (5.6) implies that, for $n \in \{0, \dots, N\}$, $|\arg(\bar{h}_n^{\text{OPT}} \sum_{m=0}^N h_m^{\text{OPT}})| \leq \pi/L$. Therefore, for $n, m \in \{0, \dots, N\}$, $|\arg(\bar{h}_n^{\text{OPT}} h_m^{\text{OPT}})| \leq 2\pi/L$. Consequently, since $h_0^{\text{OPT}} = h_{\text{TR}}$ and if $h_{\text{TR}} \neq 0$, then the slice $\mathcal{S}(\arg h_{\text{TR}}, 4\pi/L)$ contains an optimal RIS configuration. Moreover, if $h_{\text{TR}} = 0$, the slice $\mathcal{S}(0, 4\pi/L)$ contains an optimal RIS configuration. Thus, in both cases and for $L = 8$, Corollary 5.1 implies that the adopted greedy algorithm with the above identified slices as input is guaranteed to find a solution that is at most $|2 \log_2 \cos(\pi/4)| = 1$ bit/sec/Hz away from the optimal solution regardless of the number of RIS elements N , and channel realization.

5.3 Proposed RIS Configuration Scheme and Numerical Results

5.3.1 RIS Model and Proposed Scheme

We now consider the $L = 2^k$ discrete phase shifts $\phi_\ell = 2\pi(\ell - 1)/L$, $\ell = 1, \dots, L$ and adopt the non-linear amplitude-phase shift model developed in [3] where

$$\alpha(\phi) = \alpha_{\min} + (1 - \alpha_{\min}) \left(\frac{1}{2} (\sin(\phi - \vartheta) + 1) \right)^\zeta. \quad (5.7)$$

Here $\alpha_{\min} \geq 0$ is the minimum amplitude associated with the RIS element and $\vartheta, \zeta \geq 0$ are shape parameters [3].

With this setting, the $|\mathcal{C} \cap \mathcal{V}_n| = 1$ constraint in $\mathcal{F}_{\mathcal{V}}$ of (P2) can be replaced by $|\mathcal{C} \cap \mathcal{V}_n| \leq 1$. This follows as for each $n \in \{1, \dots, N\}$, since $L \geq 2$ there always exists an element $\hat{h}_{n,l} \in \mathcal{V}_n$ that can be selected by a greedy algorithm to improve the $R(\mathcal{C})$ objective in (P2) (or at least keep it unchanged). Thus, without loss in performance, the RIS configuration is expressed as:

$$\max_{\mathcal{C} \in \mathcal{I}_{\mathcal{V}}} R(\mathcal{C}), \quad (\text{P3})$$

where $\mathcal{I}_{\mathcal{V}} \triangleq \{\mathcal{C} \subseteq \mathcal{V} : |\mathcal{C} \cap \mathcal{V}_n| \leq 1, \forall n = 0, \dots, N\}$. We note that $\mathcal{I}_{\mathcal{V}}$ corresponds to the collection of independent sets of a partition matroid formed over the ground set $\mathcal{V}$. Subsequently, we adopt the matroid-constrained virtual-channel selection formulation in (P3) and Algorithm 2 (with $\mathcal{G} \leftarrow \mathcal{V}$, $\mathcal{I}_{\mathcal{G}} \leftarrow \mathcal{I}_{\mathcal{V}}$, $F(\cdot) \leftarrow R(\cdot)$ and always choose the direct path as the first virtual-channel) to optimize the general discrete RIS problem.

5.3.2 Simulation Settings and Benchmarks

We consider an RIS-assisted system operating at $f_c = 1.5$ GHz, a bandwidth of 20 MHz, and a transmission power $P = 25$ dBm. The BS is located at $(0, 250, 25)$ m. The RIS is modeled as a uniform rectangular array of N elements and is located at $(250, 0, 15)$ m. The location of the user is uniformly generated in a disc of center $(250, 20, 1.5)$ m and a radius of 5m. We take $N_0 = -170$ dBm/Hz with a noise figure of 7 dB. For numerical simulations, the user acts as the TX and the BS acts as the RX.

The direct TX-RX channel undergoes Rayleigh fading and is given by

$$h_{\text{TR}} = \sqrt{PL_{\text{NL}}(d_{\text{TR}})} h_{\text{TR,NLoS}}, \quad (5.8)$$

where $h_{\text{TR,NLoS}}$ is sampled from $\mathcal{CN}(0, 1)$. Moreover, for a TX-RX distance d_{TR} , the pathloss is [1]

$$PL_{\text{NL}}(d_{\text{TR}}) = 13.54 + 39.08 \times \log_{10}(d_{\text{TR}}) + 20 \times \log_{10}(f_c).$$

The TX-RIS channel $\mathbf{f} \in \mathbb{C}^{N \times 1}$ undergoes Rician fading. With the line-of-sight (LoS) component $\mathbf{f}_{\text{LoS}}$, non-line-of-sight (NLoS) component $\mathbf{f}_{\text{NLoS}}$, and d_{TI} indicating the distance between the TX and RIS,

$$\mathbf{f} = \sqrt{PL_{\text{L}}(d_{\text{TI}})} (\sqrt{\kappa_1/(1 + \kappa_1)} \mathbf{f}_{\text{LoS}} + \sqrt{1/(1 + \kappa_1)} \mathbf{f}_{\text{NLoS}}). \quad (5.9)$$

Here, the n^{th} component of $\mathbf{f}_{\text{LoS}}$ is given by $\mathbf{f}_{\text{LoS}}^n = \exp(2\pi j d_{\text{T,n}}/\lambda_c)$, where $d_{\text{T,n}}$ is the 3D distance between the TX and the n^{th} RIS element and λ_c is the wavelength associated with f_c . Moreover, the Rician factor is $\kappa_1 = 10$ dB, the LoS pathloss is given by [1]

$$PL_{\text{L}}(d) = 28 + 22 \times \log_{10}(d) + 20 \times \log_{10}(f_c)$$

and the components of $\mathbf{f}_{NLoS}$ are i.i.d sampled from $\mathcal{CN}(0, 1)$. Similarly, the RIS-RX channel $\mathbf{g} \in \mathbb{C}^{1 \times N}$ is modeled with Rician fading and with a Rician factor $\kappa_2 = 10$ dB,

$$\mathbf{g} = \sqrt{PL_L(d_{IR})}(\sqrt{\kappa_2/(1 + \kappa_2)}\mathbf{g}_{LoS} + \sqrt{1/(1 + \kappa_2)}\mathbf{g}_{NLoS}), \quad (5.10)$$

where d_{IR} indicates the distance between the RIS and RX.

To evaluate the performance of the proposed RIS configuration framework, the following benchmarks are considered:

- *Continuous phase shift*: The RIS with phase-dependent amplitude response is configured using Couenne numerical solver ([11]) with a time limit of 1 hour per channel realization.
- *Closest Point Projection (CPP)*: The continuous phase shifts from the BCD Algorithm 1 of [3] are projected componentwise to the closest point in $\{\phi_1, \dots, \phi_L\}$. Such projection is commonly adopted in the literature.
- *Optimal Binary configuration*: In the special case that $L = 2$, the scheme developed in [88] to configure a binary-RIS under phase-dependent amplitude response is adopted. We emphasize that in contrast with [88], the proposed scheme configures an RIS with an arbitrary number of phase-shift levels per RIS element.

5.3.3 Numerical Experiments

The robustness of the proposed framework is illustrated in two scenarios: first with the TX-RX path being blocked and second with the existence of a direct TX-RX path. Moreover, for each scenario, the performance is shown under two different RIS reflection settings, namely moderate phase-dependent amplitude selectivity ($\alpha_{min} = 0.6$, $\zeta = 3$, $\vartheta = 0.5\pi$) and significant phase-dependent amplitude selectivity ($\alpha_{min} = 0.2$, $\zeta = 3$, $\vartheta = 0.5\pi$). The numerical experiments are conducted using 500 independent channel realizations.

For a 100 element RIS controlled with binary phase shifts and under the blocked direct TX-RX path scenario, Fig. 5.2a shows the empirical CDF of the spectral efficiency (SE) associated with the proposed scheme and benchmarks. From Fig. 5.2a, with moderate amplitude selectivity ($\alpha_{min} = 0.6$), the median SE performance associated with the proposed scheme and CPP benchmark, respectively, result in a 3%, and an 11% gap compared with that of the optimal configuration. Moreover, from Fig. 5.2a with significant amplitude selectivity ($\alpha_{min} = 0.2$), the median SE associated with the proposed scheme is 4% away

from the optimal binary-RIS performance. This is in contrast to a median SE gap of 47% associated with the CPP benchmark. Thus, under both reflection models, the proposed scheme operates close to the optimal binary-RIS performance and outperforms the CPP benchmark. Similarly, under the direct TX-RX path scenario and for $N = 100$, Fig. 5.2b shows that the proposed scheme operates close to the optimal binary-RIS performance in both cases and significantly outperforms the CPP benchmark in one case.

For $N = 100$, Figs. 5.2c and 5.2d show the impact of the number of bits k used to configure the RIS on the average spectral efficiency performance associated with the proposed scheme and benchmarks. Figs. 5.2c and 5.2d were obtained under significant amplitude selectivity ($\alpha_{\min} = 0.2$) and moderate amplitude selectivity ($\alpha_{\min} = 0.6$). From the figures, under both $\alpha_{\min} = 0.2$ and $\alpha_{\min} = 0.6$, the proposed scheme requires only $k = 3$ control bits per RIS element to achieve more than 90% of the performance associated with a benchmark that uses continuous phase shifts and a numerical solver. Moreover, as demonstrated by Figs. 5.2c and 5.2d, the proposed scheme significantly outperforms the CPP benchmark in some scenarios while achieving the same performance in all other cases.

The performance of the proposed scheme and benchmarks was also evaluated for different numbers of RIS elements N . Indeed, the trends and conclusions obtained under $N = 100$ continue to hold for $N = 200$ and $N = 400$ scenarios. For instance, under $N = 200$ and blocked TX-RX path, Fig. 5.3a shows that the proposed scheme performs close to the optimal binary configuration. Moreover, under significant amplitude selectivity ($\alpha_{\min} = 0.2$), the proposed scheme significantly outperforms the corresponding CPP benchmark. Additionally, Fig. 5.3c indicates that for $N = 200$ and blocked TX-RX path, the proposed scheme requires $k = 3$ bits per RIS element to approach the performance associated with a benchmark that uses continuous phase shifts and a numerical solver. Similarly, Figs. 5.3b and 5.3d suggest that a similar conclusion can be drawn for $N = 200$ when a direct TX-RX path exists.

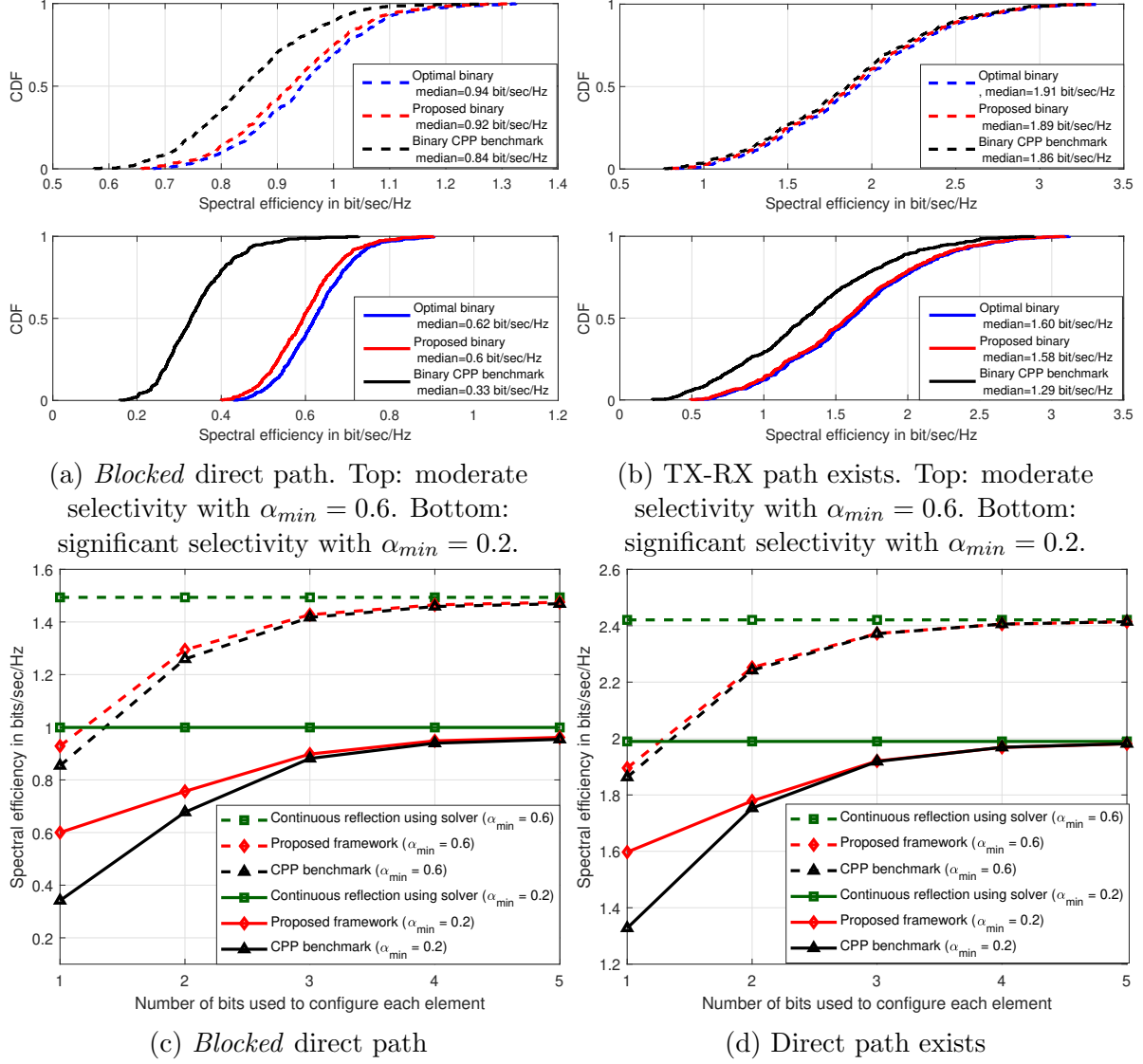


Figure 5.2: Figs. 5.2a and 5.2b show the empirical CDF of the spectral efficiency performance for ($k = 1$) binary RIS. Figs. 5.2c and 5.2d show average spectral efficiency performance vs number of bits k used to control the RIS under $\alpha_{min} = 0.2$ and $\alpha_{min} = 0.6$. Here, $N = 100$.

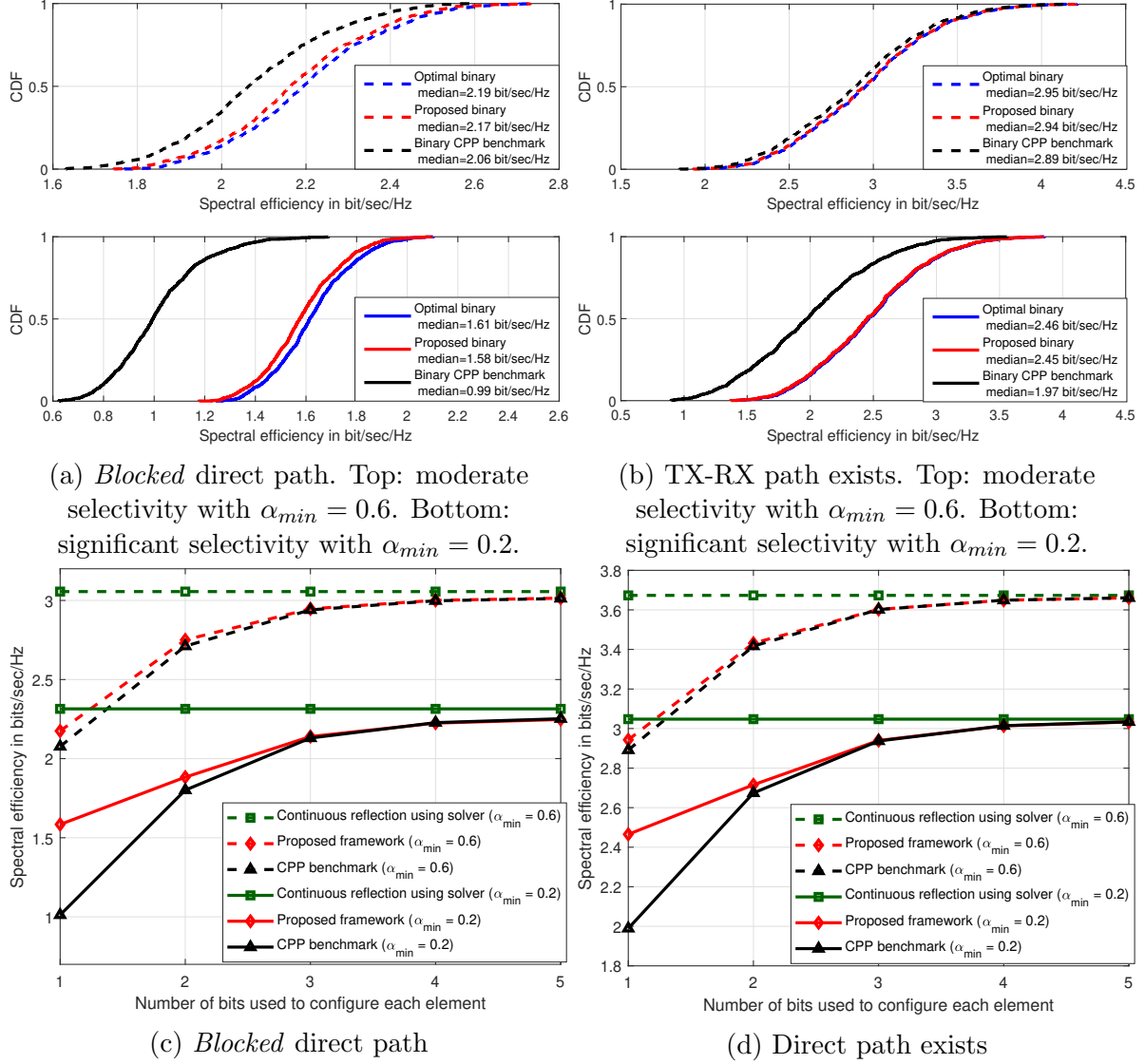


Figure 5.3: Figs. 5.3a and 5.3b show the empirical CDF of the spectral efficiency performance for ($k = 1$) binary RIS. Figs. 5.3c and 5.3d show average spectral efficiency performance vs number of bits k used to control the RIS under $\alpha_{min} = 0.2$ and $\alpha_{min} = 0.6$. Here, $N = 200$.

5.3.4 Complexity Comparisons

The proposed framework of Algorithm 2 loops for N iterations, where in iteration $n = 0, \dots, N-1$, $(N-n)L$ complex additions and the computation of $(N-n)L$ squared magnitudes are needed to compute the magnitude square of the effective channels. Then, $(N-n)L$ comparisons are needed to find the arg max element. Thus, Algorithm 2 requires a total of $\frac{N^2L+NL}{2}$ complex additions, $\frac{N^2L+NL}{2}$ squared magnitudes, and $\frac{N^2L+NL}{2}$ comparisons. Here, L is the number of RIS phase shift levels, N is the number of RIS elements. Since for practical scenarios $L \ll N$, the overall complexity is $\mathcal{O}(N^2)$. Further, Algorithm 2 can be easily parallelized to reduce its time complexity. Particularly, for each iteration $n = 0, \dots, N-1$, the complex additions and squared magnitudes computation needed in the arg max operation can be parallelized in one step. It remains to perform the comparisons needed to find the arg max element and parallelization allows this to be done in $\log(N)$ time. Thus, a fully parallelized version of Algorithm 2 has a time complexity of $\mathcal{O}(N \log(N))$.

The CPP benchmark is based on the continuous phase shifts from a BCD Algorithm. We note the BCD algorithm involves I_O outer iterations where each iteration involves N non-linear optimization problems each with complexity P . Hence, the BCD algorithm has a complexity of $\mathcal{O}(I_O NP)$ where P is the complexity of the non-linear optimization step. The number of required iterations I_O depends on the stopping criteria, and is random in general. For instance, the stopping criteria could be to stop once the difference between successive outer iterations is less than some $\epsilon > 0$. Simulations for $N = 100$ RIS elements and over 1000 channel realizations show that $\epsilon = 10^{-4}$ resulted in a maximum of $I_O^{\max} = 217$ iterations, average $I_O^{AVG} = 42$ iterations, and 90% quantile of $I_O^{90\%} = 89$ iterations. Thus, the number of outer iterations can be of the order of N or greater in this practical case of interest.

It is important to emphasize that in addition to the aforementioned complexity comparison, the low-complexity of the proposed framework stems from the simplicity of the operations involved (addition and squared magnitude) and the fact that it *always* loops for exactly N iteration steps regardless of the channel realization. This is in contrast with the BCD algorithm of the CPP benchmark which involves solving non-linear optimization problems and the number of outer iteration can vary greatly depending on the channel realization as previously outlined. Additionally, rather than requiring the knowledge of an accurate model that continuously captures the non-linear behavior between the amplitude and phase, the proposed framework only requires the knowledge of the amplitude response at the few considered discrete points. Thus, the proposed framework is suitable for practical implementation.

5.4 Multi-antenna receivers

In this section, we consider a multi-antenna extension of the constrained virtual-antenna selection formulation of Section 5.2. We also expand upon the results of Section 5.2 and investigate the performance of a greedy configuration scheme (i.e., Algorithm 2) when employed to configure the RIS in a multi-antenna scenario.

We consider a system where a transmitter (TX) sends a symbol $s \in \mathbb{C}$ with power P to an M -antenna receiver (RX). The signal transmissions are assisted by an N -element RIS with configuration $\mathbf{\Gamma} = [\Gamma_1, \dots, \Gamma_N]$. Similar to Section 5.2, the n^{th} element of the RIS is configured with reflection coefficient $\Gamma_n \in \{\gamma_1, \dots, \gamma_L\}$ where $\gamma_\ell \triangleq \alpha(\phi_\ell)e^{-j\phi_\ell}$.

The direct TX-RX channel is represented as $\mathbf{h}_{\text{TR}} \in \mathbb{C}^{M \times 1}$. Moreover, the TX-RIS channel is denoted as $\mathbf{f} \in \mathbb{C}^{N \times 1}$ and the RIS-RX channel is $\mathbf{G} \in \mathbb{C}^{M \times N}$. Thus, the signal received by the RX is given by

$$\mathbf{y} = (\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma}) \sqrt{P} s + \mathbf{n}, \quad (5.11)$$

where $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2 \mathbb{I}_M)$ is the AWGN at the receiver. With the RX employing a unit-norm receive combining vector $\mathbf{v} \in \mathbb{C}^{M \times 1}$ and $\mathbf{\Gamma} = [\Gamma_1, \dots, \Gamma_N]$, the rate maximization problem can be formulated as

$$\begin{aligned} & \max_{\mathbf{\Gamma}, \mathbf{v}} && \log_2 \left(1 + \frac{P |(\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma})^H \mathbf{v}|^2}{\sigma^2} \right) \\ & \text{subject to} && \Gamma_n \in \{\gamma_1, \dots, \gamma_L\} \\ & && \|\mathbf{v}\|^2 = 1. \end{aligned} \quad (\text{P4})$$

5.4.1 Virtual-antenna Selection Formulation and Analysis

The objective in (P4) is equivalent to $\max_{\mathbf{\Gamma}} \max_{\mathbf{v}} |(\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma})^H \mathbf{v}|^2$. We note that a maximum-ratio combining (MRC) strategy $\mathbf{v}^{\text{MRC}} = \frac{\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma}}{\|\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma}\|}$ optimizes the inner problem $\max_{\mathbf{v}} |(\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma})^H \mathbf{v}|^2$ subject to $\|\mathbf{v}\|^2 = 1$. By substituting $\mathbf{v}^{\text{MRC}}$ into (P4), the RIS configuration problem is equivalent to

$$\begin{aligned} & \max_{\mathbf{\Gamma}} && \|\mathbf{h}_{\text{TR}} + \mathbf{G} \text{diag}(\mathbf{f}) \mathbf{\Gamma}\|^2 \\ & \text{subject to} && \Gamma_n \in \{\gamma_1, \dots, \gamma_L\}. \end{aligned} \quad (\text{P5})$$

Next, we extend the virtual-antenna formulation of Section 5.2 to optimize (P5) for the general $M > 1$ case.

To illustrate the virtual-antenna selection formulation, we note that (P5) is equivalent to

$$\max_{\Gamma_n \in \{\gamma_1, \dots, \gamma_L\}, n \in \mathcal{N}} \left\| \begin{bmatrix} h_{1,\text{TR}} + \sum_{n=1}^N g_{1,n} f_n \Gamma_n \\ \vdots \\ h_{M,\text{TR}} + \sum_{n=1}^N g_{M,n} f_n \Gamma_n \end{bmatrix} \right\| \quad (5.12)$$

where for $m \in \{1, \dots, M\}$, $h_{m,\text{TR}}$ is the m^{th} component of $\mathbf{h}_{\text{TR}}$ and $g_{m,n} = [\mathbf{G}]_{m,n}$. For each RIS element $n \in \mathcal{N}$, there are L possible choices of $\Gamma_n \in \{\gamma_1, \dots, \gamma_L\}$ resulting in L possible virtual-channels

$$\tilde{\mathbf{h}}_{n,\ell} = \begin{bmatrix} \tilde{h}_{1,n,\ell} \\ \vdots \\ \tilde{h}_{M,n,\ell} \end{bmatrix} = \gamma_\ell f_n \begin{bmatrix} g_{1,n} \\ \vdots \\ g_{M,n} \end{bmatrix} \quad n \in \mathcal{N}, \ell \in \{1, \dots, L\}. \quad (5.13)$$

These are called virtual-channels as only *one* will be selected per RIS element.

Let $\mathcal{V}_0 = \{\mathbf{h}_{\text{TR}}\}$ and for $n \in \mathcal{N} = \{1, \dots, N\}$, let $\mathcal{V}_n = \{\tilde{\mathbf{h}}_{n,1}, \dots, \tilde{\mathbf{h}}_{n,L}\}$. Then, the RIS configuration problem corresponds to selecting a subset $\mathcal{C} = \{\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_N\}$, where $\mathbf{h}_n \in \mathcal{V}_n$ for $n = 0, \dots, N$. Subsequently, for $\mathcal{V} \triangleq \bigcup_{n=0}^N \mathcal{V}_n$ and over a family of sets $\mathcal{F}_{\mathcal{V}} \triangleq \{\mathcal{C} \subseteq \mathcal{V} : |\mathcal{C} \cap \mathcal{V}_n| = 1, \forall n = 0, \dots, N\}$, maximizing (P5) is equivalent to

$$\max_{\mathcal{C} \in \mathcal{F}_{\mathcal{V}}} Q(\mathcal{C}), \quad (P6)$$

where $Q(\mathcal{C}) \triangleq \|\sum_{\mathbf{h} \in \mathcal{C}} \mathbf{h}\|$ yields a rate $R(\mathcal{C}) = \log_2(1 + P(Q(\mathcal{C}))^2/\sigma^2)$.

To motivate adopting a greedy algorithm (i.e., Algorithm 2) to optimize (P6), we extend the concept of slices (introduced in Section 5.2.1) to the multi-antenna case considered here. For $\mathbf{u}_c \in \mathbb{C}^M$ with $\|\mathbf{u}_c\| = 1$, a $(\mathbf{u}_c, \theta_{\max})$ slice is a subset $\mathcal{S}(\mathbf{u}_c, \theta_{\max}) = \mathcal{V} \cap \mathcal{U}(\mathbf{u}_c, \theta_{\max})$, where $\mathcal{U}(\mathbf{u}_c, \theta_{\max}) = \{\mathbf{z} \in \mathbb{C}^M : \Re\{\langle \mathbf{z}, \mathbf{u}_c \rangle\} \geq \|\mathbf{z}\| \cos(\theta_{\max}/2)\}$.

The Proposition below shows that, under some mild conditions and except for possibly the direct path, the optimal solution to (P6) lies within a slice.

Proposition 5.3. *Consider (P6) with $L \geq 4$ and phase shifts $\phi_\ell = 2\pi(\ell - 1)/L$. Then, for an optimal configuration $\mathcal{C}^{\text{OPT}} \triangleq \{\mathbf{h}_0^{\text{OPT}}, \mathbf{h}_1^{\text{OPT}}, \dots, \mathbf{h}_N^{\text{OPT}}\}$, it holds that $\{\mathbf{h}_1^{\text{OPT}}, \dots, \mathbf{h}_N^{\text{OPT}}\} \subseteq \mathcal{S}(\mathbf{u}_c, \theta_{\max})$ for some $\mathbf{u}_c$ and $\theta_{\max} < \pi$.*

The proof is given in Appendix A.6.

Proposition 5.4 below shows that, over a slice $\mathcal{S}(\mathbf{u}_c, \theta_{\max})$ with $\theta_{\max} < \pi$, the objective function $Q(\mathcal{C})$ of (P6) can be bounded between an upper bound

$$g_u(\mathcal{C}) \triangleq \sum_{\mathbf{h} \in \mathcal{C}} \|\mathbf{h}\|, \quad (5.14)$$

and a lower bound

$$g_l(\mathcal{C}) \triangleq \sum_{\mathbf{h} \in \mathcal{C}} \|\mathbf{h}\| \cos(\theta_{\max}/2). \quad (5.15)$$

Proposition 5.4. *Consider a slice $\mathcal{S}(\mathbf{u}_c, \theta_{\max})$ with $\theta_{\max} < \pi$, then for any subset $\mathcal{C} \subseteq \mathcal{S}(\mathbf{u}_c, \theta_{\max})$,*

$$g_l(\mathcal{C}) \leq Q(\mathcal{C}) \leq g_u(\mathcal{C}). \quad (5.16)$$

Here, $g_u(\mathcal{C})$ and $g_l(\mathcal{C})$ are monotone and modular functions. Further, if $\theta_{\max} \leq \pi/2$, then $Q(\mathcal{C})$ is monotone.

The proof is given in Appendix A.7.

For a slice $\mathcal{S}(\mathbf{u}_c, \theta_{\max})$, define $\mathcal{F}_S \triangleq \{\mathcal{C} \subseteq \mathcal{S}(\mathbf{u}_c, \theta_{\max}) : |\mathcal{C} \cap \mathcal{V}_n| = 1, \forall n = 0, \dots, N\}$, and $\mathcal{I}_S \triangleq \{\mathcal{C} \subseteq \mathcal{S}(\mathbf{u}_c, \theta_{\max}) : |\mathcal{C} \cap \mathcal{V}_n| \leq 1, \forall n = 0, \dots, N\}$.

By the monotonicity of $g_u(\mathcal{C})$, a maximizer of $\max_{\mathcal{C} \in \mathcal{I}_S} g_u(\mathcal{C})$ also maximizes $\max_{\mathcal{C} \in \mathcal{F}_S} g_u(\mathcal{C})$. Since $\max_{\mathcal{C} \in \mathcal{I}_S} g_u(\mathcal{C})$ involves maximizing a modular function under a matroid constraint, it can be optimized by the greedy Algorithm 2 [24]. Moreover, for $\theta_{\max} < \pi$, a maximizer of $\max_{\mathcal{C} \in \mathcal{I}_S} g_u(\mathcal{C})$ also maximizes $\max_{\mathcal{C} \in \mathcal{I}_S} g_l(\mathcal{C})$. Subsequently, by proceeding similar to Corollary 5.1, Proposition 5.4 implies that if an optimal RIS configuration is contained within a slice $\mathcal{S}(\mathbf{u}_c, \theta_{\max})$ with $\theta_{\max} < \pi$, then Algorithm 2 applied with $\mathcal{S}(\mathbf{u}_c, \theta_{\max})$, $\mathcal{I}_S$, and $g_u(\mathcal{C})$ produces an RIS configuration $\tilde{\mathcal{C}}_S$ that is optimal for both $\max_{\mathcal{C} \in \mathcal{F}_S} g_u(\mathcal{C})$ and $\max_{\mathcal{C} \in \mathcal{F}_S} g_l(\mathcal{C})$, and thereafter $R(\tilde{\mathcal{C}}_S)$ is within $|2 \log_2 \cos(\theta_{\max}/2)|$ bit/sec/Hz of the optimal objective of (P4).

5.4.2 Simulation Settings and Benchmarks

In this section, we adopt the constrained virtual-channel selection formulation of this section and investigate the performance of the corresponding Algorithm 2 when employed to optimize the general multi-antenna receiver problem (P4).

For $L = 2^k$ discrete phase shifts $\phi_\ell = 2\pi(\ell - 1)/L, \ell = 1, \dots, L$, the amplitude-phase shift model (5.7) presented in Section 5.3.1 is adopted. Moreover, the simulation parameters of Section 5.3.2 are adopted and extended for the case where the receiver has M antennas. In particular, $\mathbf{h}_{\text{TR}} \in \mathbb{C}^{M \times 1}$ and $\mathbf{G} \in \mathbb{C}^{M \times N}$ now extend $h_{\text{TR}} \in \mathbb{C}$ of (5.8) and $\mathbf{g} \in \mathbb{C}^{1 \times N}$ of (5.10), respectively. Subsequently, numerical experiments are conducted under various scenarios to demonstrate the efficacy of the proposed configuration scheme, with particular attention given to the impact of the number of antennas, M .

Benchmarks:

It is important to note that, unlike the single-antenna receiver case discussed in Section 5.3, the optimal binary configuration for cases where $M > 1$ remains unknown. To the best of our knowledge, this is the first time the discrete RIS configuration problem with amplitude selectivity is being investigated in multi-antenna scenarios (i.e., $M > 1$). Consequently, to assess the performance of the proposed configuration scheme, the following benchmarks are adopted:

- *Continuous phase shift*: the continuous relaxation of the multi-antenna problem (P4) (i.e., RIS configuration problem with continuous phase shifts under phase-dependent amplitude response) is configured using Couenne numerical solver ([11]) with a time limit of 1 hour per channel realization.
- *Closest Point Projection (CPP)*: the continuous phase shifts from the BCD Algorithm 1 of [3] are projected componentwise to the closest point in $\{\phi_1, \dots, \phi_L\}$.

5.4.3 Numerical Results

We first examine a scenario where the RX has $M = 2$ antennas and the RIS is equipped with $N = 100$ elements. Under this setup, Fig. 5.4 illustrates the average spectral efficiency performance associated with the proposed scheme and benchmarks for different number of bits k used to configure the RIS. We consider two cases: (i) a blocked TX-RX scenario (Fig. 5.4a) and (ii) a direct TX-RX scenario (Fig. 5.4b). From the figures, under a moderate amplitude selectivity ($\alpha_{\min} = 0.6$), the proposed scheme performs similarly to the CPP benchmark. However, under significant amplitude selectivity ($\alpha_{\min} = 0.2$), the proposed scheme can provide substantial gains over the CPP benchmark.

For instance, Fig. 5.4a shows that in the blocked TX-RX case, configuring the RIS elements with $k = 1$ bits (i.e. binary configuration) using the proposed scheme results

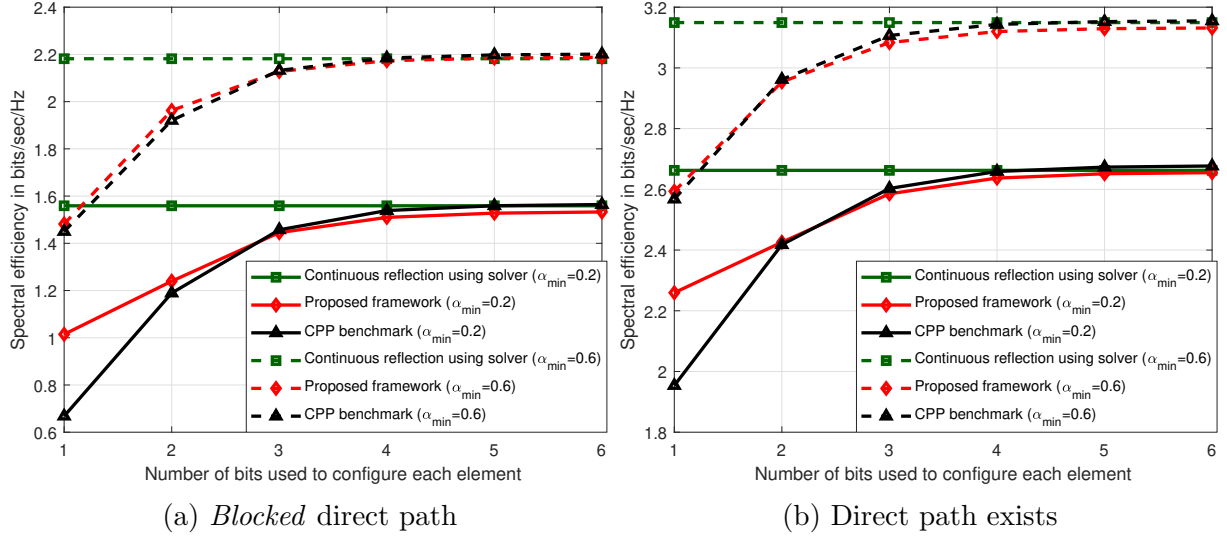


Figure 5.4: The average spectral efficiency performance vs number of bits used to control the RIS under $\alpha_{min} = 0.2$ and $\alpha_{min} = 0.6$. Here $M = 2$ and $N = 100$.

in a 52% spectral efficiency improvement compared to the binary configuration of the CPP benchmark. Moreover, for the direct TX-RX case, Fig. 5.4b shows that the binary configuration obtained by the proposed scheme yields a 16% improvement over the corresponding CPP benchmark performance. Additionally, across all scenarios illustrated in Fig. 5.4, the proposed scheme requires $k = 3$ bits to achieve $> 90\%$ of the performance associated with the continuous phase shift benchmark.

For a receiver with $M = 4$ antennas and an RIS of either $N = 100$ or $N = 200$ elements, Fig. 5.5 illustrates the average spectral efficiency performance of the proposed scheme compared to the benchmarks as a function of the number of bits k used to configure the RIS. Here, the results were obtained under a significant amplitude selectivity ($\alpha_{min} = 0.2$) scenario. As shown in the figures, the proposed scheme consistently matches or surpasses the performance of the CPP benchmark, with significant improvements in certain scenarios. For example, under the blocked TX-RX path and $N = 100$ elements (Fig. 5.5a), the binary configuration generated by the proposed scheme results in a 45% improvement in spectral efficiency over the CPP benchmark. Similarly, for $N = 200$ elements and $k = 1$, the proposed scheme achieves a 32% improvement compared to the CPP benchmark.

For $N = 100$ and significant amplitude selectivity ($\alpha_{min} = 0.2$), Fig. 5.6 illustrates the impact of the number of antennas M on spectral efficiency performance. The figure highlights the performance of the proposed scheme for $k = 1$ (red dotted line) and $k = 2$

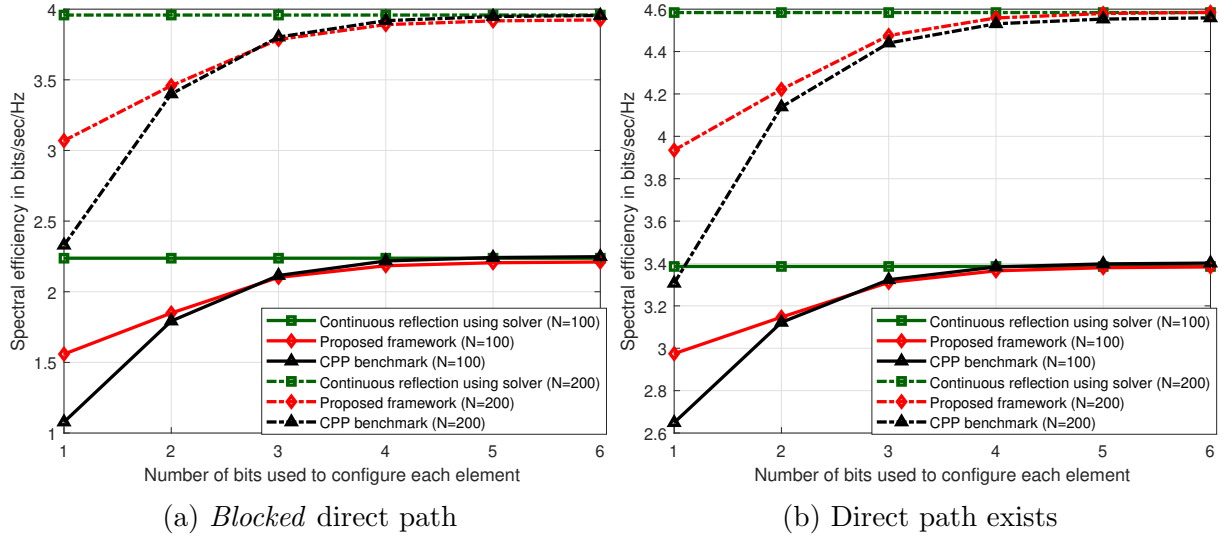


Figure 5.5: The average spectral efficiency performance vs number of bits used to control the RIS for both $N = 100$ and $N = 200$. Here $M = 4$ and $\alpha_{min} = 0.2$.

(red dashed line), compared to the CPP benchmark for $k = 1$ (black dotted line) and $k = 2$ (black dashed line). Additionally, it includes the performance of an RIS configured with continuous phase shifts, shown as a solid green line.

For the $k = 1$ case, regardless of whether the direct TX-RX path is blocked or not, the proposed scheme consistently provides significant improvements over the CPP benchmark, effectively narrowing the gap to the continuous phase shift benchmark across different values of M . Specifically, in the blocked direct path scenario (Fig. 5.6a), the CPP benchmark results in performance gaps of 64%, 58%, 53%, and 47% for $M=1, 2, 4, 8$, respectively, compared to the continuous benchmark. In contrast, the proposed binary ($k = 1$) configuration nearly halves these gaps, yielding 37%, 35%, 30%, and 25% gaps for the same values of M . For $k = 2$, the proposed scheme continues to outperform the CPP benchmark for some cases. For instance, when $M = 1$, the proposed scheme reduces the performance gap to 20%, compared to a 29% gap under the CPP benchmark. Similar trends and conclusions hold in the scenario where the direct TX-RX path exists, as shown in Fig. 5.6b.

We note that, although not shown in Fig. 5.6, similar results were obtained for a massive MIMO scenario with $M = 8 \times 8 = 64$ antennas at the receiver. For instance, under the binary case ($k = 1$) and *blocked* direct path, the proposed scheme nearly halves the gap to the continuous benchmark to 16% compared to the CPP benchmark, which results in a 30% gap relative to the continuous benchmark.

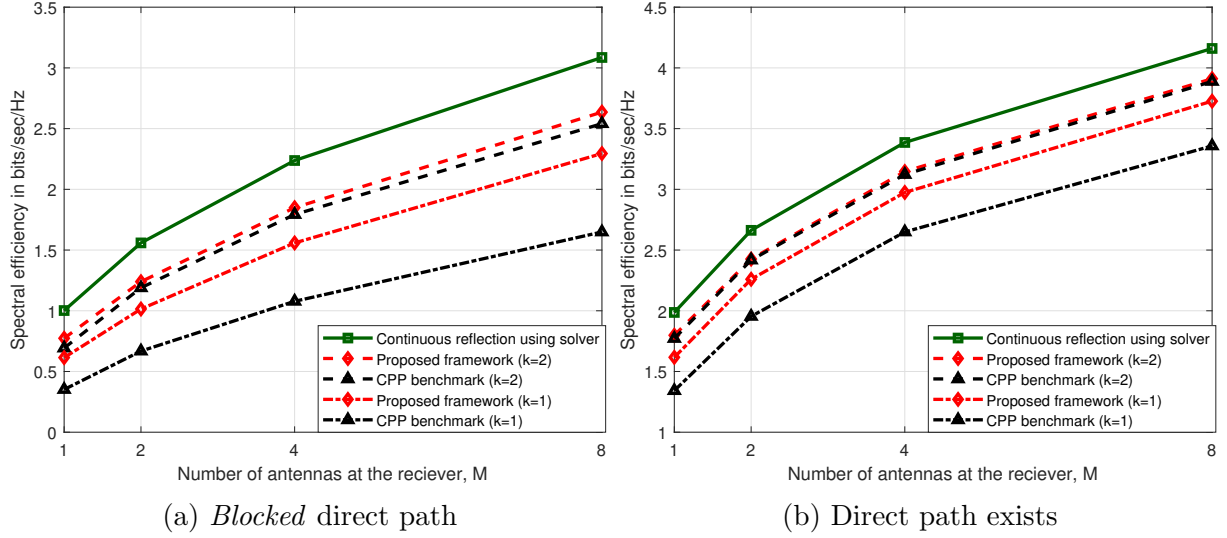


Figure 5.6: The average spectral efficiency performance vs number of antennas at the receiver for both $k = 1$ and $k = 2$ bits used to control the RIS. Here $N = 100$ and $\alpha_{min} = 0.2$.

The numerical results of this section demonstrate that the proposed scheme consistently outperforms the CPP benchmark or achieves comparable performance. Notably, as discussed in Section 5.3.4, the proposed scheme offers significant practical advantages in terms of time complexity and simplicity of the operation needed for optimizing the RIS configuration. Therefore, even in scenarios where the spectral efficiency performance of the proposed scheme matches that of the CPP benchmark, such as under moderate phase-dependent amplitude response or with a higher number of bits per RIS element, it remains a favorable option to adopt the proposed scheme due to its practical gains in terms of time complexity and simplicity of the operations involved.

5.5 Conclusions

The discrete reconfigurable intelligent surfaces (RIS) configuration problem under phase-dependent amplitude response was formulated as a constrained virtual-channel selection problem. This problem, when considered over a subset of virtual-channels with maximum phase-variation $\theta_{max} < \pi$, was shown to be bounded between two monotone matroid-constrained problems. Moreover, when the optimal RIS configuration lies within such

subsets, a greedy configuration framework was shown to be able to guarantee a worst-case optimality gap of $|2\log_2 \cos(\theta_{\max}/2)|$ bit/sec/Hz uniformly over all system parameters. RIS configuration problems whose optimal configuration lies within such maximum phase-variation subsets were demonstrated. Subsequently, a practical low-complexity configuration framework was proposed for general discrete RIS configuration problems. Numerical experiments corroborated the effectiveness of the proposed framework which required only 3 control bits per RIS element to achieve at least 90% of the continuous phase shift performance obtained by a numerical solver. Additionally, in the single-antenna case with binary phase shifts and significant amplitude-response variations, the proposed framework achieves performance within 4% of the optimal configuration. Furthermore, the proposed scheme either significantly outperforms the commonly adopted closest point projection (CPP) benchmark, where continuous phase shifts from a block-coordinate descent algorithm are discretized, or achieves comparable performance. Even in scenarios where both schemes yield similar results, the proposed scheme has practical advantages, including reduced time complexity and simplicity of the operations needed to optimize the RIS configuration.

Chapter 6

Concluding Remarks and Future Work

Current systems focus on optimizing the transmitter and receiver, leaving channel configuration largely beyond the control of a designer. A reconfigurable intelligent surface (RIS) is a 2D array of nearly-passive reflecting elements that can be configured to dynamically shape electromagnetic waves, offering a cost-effective and energy-efficient approach to engineering the channel between a transmitter and receiver. However, the lack of active components at an RIS presents challenges in terms of channel acquisition and optimization of the RIS reflection coefficient. This thesis addresses these challenges by developing efficient resource allocation and RIS configuration schemes.

6.1 Summary of Contributions and Key Findings

In the first part of this research, we propose a joint uplink-downlink RIS design for multi-user frequency-division-duplexing (FDD) and time-division-duplexing (TDD) systems. This joint RIS design is essential for supporting simultaneous uplink (UL) and downlink (DL) transmissions in an FDD system, as designing separate configurations for UL and DL is not physically realizable. Furthermore, in a TDD system, the joint design can reduce signaling overhead, power consumption, and configuration periods compared to individually designing UL and DL RIS configurations. To investigate the trade-off between uplink and downlink rates achieved by the joint design, we formulate a weighted-sum optimization problem and optimize it using a developed block-coordinate descent algorithm.

With the RIS configuration optimized through the proposed framework, numerical simulations are conducted to compute the uplink-downlink trade-off regions in practical scenarios where RIS can be deployed to assist signal transmission in communication systems.

To evaluate the performance of the proposed framework, we introduce several benchmarks. These include an individual design upper bound, which allows different RIS configurations for UL and DL, a fixed-uplink (fixed-downlink) heuristic that uses the RIS configuration optimized for uplink (downlink) to assist downlink (uplink), and slicing benchmarks, where RIS elements are split into two groups dedicated to assisting either UL or DL transmissions exclusively. Numerical investigations of the UL-DL trade-off regions show that the proposed joint design can offer significant performance improvements over the fixed-downlink/fixed-uplink designs. In many scenarios, the UL-DL region achieved by the joint design is shown to strictly contain those associated with the fixed-downlink/fixed-uplink designs and slicing benchmarks. Additionally, the joint design substantially bridges the gap to the individual design upper bound, which cannot be realized in an FDD system.

Furthermore, complexity analysis of the proposed scheme and benchmarks is combined with the computed trade-off regions to highlight scenarios where minor losses in spectral efficiency can be traded for reductions in configuration complexity, configuration periods, and signaling overhead. Specifically, for all FDD scenarios, the proposed joint design consistently outperforms fixed-downlink/fixed-uplink designs, making it indispensable in such systems. For TDD scenarios, the joint design can yield substantial spectral efficiency improvements over fixed-downlink/fixed-uplink designs in scenarios with significant variations in user priorities. Meanwhile, in other TDD scenarios, the joint design and fixed uplink/fixed downlink configurations perform almost as well as the individual design upper bound, offering reduced complexity, overhead, and configuration periods without significant losses in spectral efficiency.

The second part of the research develops a learning-based framework that directly exploits noisy pilots to optimize an RIS system while accommodating different service priorities and fairness through user weights. To achieve this, an adaptive beamforming configuration problem is formulated to generate the RIS system's beamforming configurations that maximize the weighted sum-rate. The problem is analyzed to demonstrate that, under mild regularity conditions, it attains a maximum. To approximate solutions to this problem while accounting for various sets of user weights, one approach is to extend existing learning methods by including the user weights as input alongside the pilot information and training a single neural network to work for all possible sets of user weights. This is referred to as the static learning approach. However, this approach has a fundamental limitation: once the deep learning model is trained, its parameters are fixed and cannot be adapted to reflect changes in user weights or other available information.

To overcome this limitation, we propose a novel hypernetwork-based beamforming (HNB) framework to dynamically approximate solutions to the formulated beamforming problem. In the HNB framework, a beamforming network (BFN) is designed to leverage available information, including noisy pilots, to generate optimized beamforming configurations. Instead of directly training the BFN, another network, referred to as the hypernetwork, is trained to dynamically generate a subset of the BFN’s learning parameters based on an input conditioning vector. By using user weights as the conditioning vector, the HNB scheme is capable of adapting the BFN’s learning parameters to the user weights. This allows the HNB framework, during deployment, to generate a different BFN for each input condition (e.g., user weights) without requiring retraining for new conditions.

Extensive simulation results across a wide range of practical scenarios demonstrate that tuning the BFN to the user weights enables the proposed HNB framework, even with limited training, to significantly outperform static learning schemes. Furthermore, the proposed HNB framework directly leverages noisy pilots to produce beamforming configurations that achieve performance comparable to an optimistic BCD algorithm with perfect CSI as a benchmark.

When location information is available, the hypernetwork’s conditioning vector can be used to tune the BFN learning parameters to the user location information (in addition to user weights). Numerical investigations show that adapting the BFN to both user weights and locations allows the HNB framework to significantly outperform static learning schemes and considerably reduces the number of pilots needed to provide performance close to the BCD with perfect CSI benchmark. Simulation results show that the improvements introduced by the HNB framework over existing methods remain consistent across various scenarios and under different user weighting strategies, number of user pilots, noise considerations, Rician factors, and user location availability.

In the third part of this research, we focus on developing RIS configuration schemes under relaxed hardware constraints to facilitate RIS implementation across diverse technologies. Specifically, we consider the scenario where RIS elements have discrete phase shifts, and their amplitude response is non-linearly dependent on the introduced phase shift. This discrete RIS configuration problem is formulated as a constrained virtual-channel selection problem.

By considering the problem over a subset of virtual channels with a maximum phase variation of $\theta_{\max} < \pi$, it is shown that the discrete RIS configuration problem is bounded between monotone upper and lower matroid-constrained problems. When the optimal RIS configuration lies within these subsets, optimizing either the upper or lower bound ensures a worst-case optimality gap of $|2 \log_2 \cos(\theta_{\max}/2)|$ bit/sec/Hz, regardless of the system

parameters. We present discrete RIS configuration problems whose optimal solution satisfy this maximum phase-variation property.

Subsequently, a low-complexity configuration framework is proposed to optimize general RIS problems where the discrete phase shifts are specified by k bits and exhibit an arbitrary phase-dependent amplitude response. Numerical evaluations validate the efficacy of the proposed framework. For instance, under binary phase shifts with significant amplitude-response dependency, the proposed framework achieves performance within 4% of the optimal configuration for a single-antenna case. Moreover, with only three control bits per RIS element, the scheme delivers over 90% of the performance of a benchmark scheme using continuous phase shifts and numerical solvers.

Furthermore, the proposed scheme is shown to either significantly outperform the commonly adopted closest point projection (CPP) benchmark (which discretizes the continuous phase shifts from a block-coordinate descent algorithm) or achieve similar performance. The computational complexity of the proposed configuration scheme is also analyzed and shown to be advantageous compared to existing benchmarks. Thus, even in scenarios where both approaches provide similar results, the proposed scheme offers notable practical gains in terms of reduced time complexity and operational simplicity, making it a compelling choice for efficient and scalable RIS configuration.

6.2 Future Research Directions

This thesis focuses on developing efficient resource allocation and beamforming configuration schemes to integrate a RIS into future communication systems, where nearly-passive reflecting elements optimize the channel between transmitters and receivers. This work can be extended by exploring different RIS architectures, use cases, and deployment strategies.

6.2.1 Extended RIS Architectures

Hybrid Active-Passive RIS

The nearly-passive RIS implementation, widely studied in the literature, effectively mitigates noise amplification and self-interference. On the other hand, an active RIS implementation introduces active components that amplify reflected signals (along with noise), increasing power consumption, noise power, and hardware complexity. A promising alternative is a hybrid active-passive RIS design, which combines a few active elements with passive ones. For instance, active elements could enhance channel estimation. While hybrid RIS implementations have been explored in works such as [101], research in this area remains in its early stages. Incorporating active components presents challenges, including noise modeling and mitigation, resource allocation for active and passive elements, and other technical considerations that remain largely unaddressed in the literature.

Simultaneously Transmitting and Reflecting Surface

Another extended RIS architecture is explored in works such as [74] and [38], where each RIS element is designed with both refracting (sometimes referred to as transmitting) and reflecting capabilities. This RIS model is known as simultaneously transmitting and reflecting (STAR) RIS. By controlling the relative degree of refraction and reflection, the RIS can serve users on both sides. In [74, 38], the reflection and refraction coefficients of each RIS element were optimized to enhance signal transmission. The STAR-RIS implementation was shown to extend IRS coverage and improve the power efficiency of IRS-assisted systems compared to reflection-only RIS. However, it introduces several technical challenges, including the efficient acquisition of CSI information and the development of low-complexity configuration schemes under practical reflection models, which remain active areas of research.

6.2.2 RIS for Localization and Mapping Tasks

Although this thesis focuses on wireless communications, RIS reflection patterns can also be optimized to enhance localization and mapping accuracy. An RIS can be particularly useful for localization tasks, acting as a tunable anchor to control NLOS propagation [103]. A promising future direction involves designing RIS reflection coefficients to support joint sensing and communication tasks, leveraging the additional degrees of freedom provided by the RIS to address trade-offs between these tasks [55]. Additionally, determining the optimal placement of one or multiple RISs to assist localization remains an important challenge. Furthermore, the role of extended RIS architectures (e.g., hybrid RIS or STAR RIS) in localization tasks has not received sufficient attention in the literature.

6.2.3 RIS deployment strategy

The focus so far has been on centralized RIS deployment, where a single RIS assists signal transmissions. However, deploying multiple RISs collaboratively in a distributed manner could offer significant benefits for both communication and localization tasks. One key application where distributed RIS deployment could provide substantial improvements is multi-target localization and sensing [103]. A distributed RIS deployment introduces several technical challenges, including the development of effective resource allocation schemes and synchronization protocols. Additionally, in such a deployment, many users may operate within the near-field regime of their associated RIS, which must be carefully considered. Addressing these challenges makes distributed RIS deployment a promising direction for future research.

References

- [1] 3GPP. Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report TS 38.901, 3GPP, 2020. Version 16.1.0.
- [2] 3GPP. SA1 6G Study. <https://www.3gpp.org/news-events/3gpp-news/sa1-6g>, 2024.
- [3] Samith Abeywickrama, Rui Zhang, Qingqing Wu, and Chau Yuen. Intelligent reflecting surface: Practical phase shift model and beamforming optimization. *IEEE Trans. on Commun.*, 68(9):5849–5863, 2020.
- [4] Mahmoud Saad Abouamer and Patrick Mitran. Joint uplink-downlink resource allocation for multiuser IRS-assisted systems. *IEEE Trans. on Wireless Commun.*, 21(12):10918–10933, 2022.
- [5] Mahmoud Saad Abouamer and Patrick Mitran. Flexible resource allocation in IRS-assisted systems using hypernetworks. In *IEEE Wireless Commun. and Netw. Conf.*, pages 1–6, 2023.
- [6] Mahmoud Saad Abouamer and Patrick Mitran. Practical configuration of a discrete ris with phase-dependent amplitude response. *IEEE Communications Letters*, 28(11):2678–2682, 2024.
- [7] Andrea Abrardo, Davide Dardari, and Marco Di Renzo. Intelligent reflecting surfaces: Sum-rate optimization based on statistical position information. *IEEE Trans. Commun.*, 69(10):7121–7136, 2021.
- [8] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2007.
- [9] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis : a hitchhiker’s guide*. Springer, Berlin, 3rd ed. edition, 2007.

- [10] Jiancheng An, Chao Xu, Lu Gan, and Lajos Hanzo. Low-complexity channel estimation and passive beamforming for RIS-assisted mimo systems relying on discrete phase shifts. *IEEE Trans. on Commun.*, 70(2):1245–1260, 2022.
- [11] Pietro Belotti, Jon Lee, Leo Liberti, François Margot, and Andreas Wächter. Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software*, 24(4-5):597–634, 2009.
- [12] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.
- [13] E. Björnson, M. Bengtsson, and B. Ottersten. Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]. *IEEE Signal Process. Magazine*, 31(4):142–148, 2014.
- [14] Emil Björnson, Henk Wymeersch, Bho Matthiesen, Petar Popovski, Luca Sanguinetti, and Elisabeth de Carvalho. Reconfigurable intelligent surfaces: A signal processing perspective with wireless applications. *IEEE Signal Processing Magazine*, 39(2):135–158, 2022.
- [15] Emil Björnson, Özgecan Özdoğan, and Erik G. Larsson. Reconfigurable intelligent surfaces: Three myths and two critical questions. *IEEE Commun. Magazine*, 58(12):90–96, Dec. 2020.
- [16] Emil Björnson, Özgecan Özdoğan, and Erik G. Larsson. Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying? *IEEE Wireless Communications Letters*, 9(2):244–248, 2020.
- [17] L. D. Brown and R. Purves. Measurable Selections of Extrema. *The Annals of Statistics*, 1(5):902 – 912, 1973.
- [18] Yashuai Cao, Tiejun Lv, Wei Ni, and Zhipeng Lin. Sum-rate maximization for multi-reconfigurable intelligent surface-assisted device-to-device communications. *IEEE Trans. Commun.*, 69(11):7283–7296, Nov. 2021.
- [19] Etienne F. Chaponniere, Peter J. Black, Jack M. Holtzman, and David N.C. Tse. Transmitter directed code division multiple access system using path diversity to equitably maximize throughput, 2002.
- [20] Guangji Chen and Qingqing Wu. Fundamental limits of intelligent reflecting surface aided multiuser broadcast channel. *IEEE Transactions on Communications*, 71(10):5904–5919, 2023.

- [21] Jie Chen, Ying-Chang Liang, Yiyang Pei, and Huayan Guo. Intelligent reflecting surface: A programmable wireless environment for physical layer security. *IEEE Access*, 7:82599–82612, 2019.
- [22] Yingyang Chen, Miaowen Wen, Ertugrul Basar, Yik-Chung Wu, Li Wang, and Weiping Liu. Exploiting reconfigurable intelligent surfaces in edge caching: Joint hybrid beamforming and content placement optimization. *IEEE Trans. Wireless Commun.*, 20(12):7799–7812, Dec. 2021.
- [23] Sundeep Prabhakar Chepuri, Nir Shlezinger, Fan Liu, George C. Alexandropoulos, Stefano Buzzi, and Yonina C. Eldar. Integrated sensing and communications with reconfigurable intelligent surfaces: From signal modeling to processing. *IEEE Signal Processing Magazine*, 40(6):41–62, 2023.
- [24] Michele Conforti and Gérard Cornuéjols. Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the Rado-Edmonds theorem. *Disc. Appl. Math.*, 7:251–274, 1984.
- [25] Linglong Dai, Bichai Wang, Min Wang, Xue Yang, Jingbo Tan, Shuangkaisheng Bi, Shenheng Xu, Fan Yang, Zhi Chen, Marco Di Renzo, Chan-Byoung Chae, and Lajos Hanzo. Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results. *IEEE Access*, 8:45913–45923, March 2020.
- [26] Tomislav Debogovic and Julien Perruisseau-Carrier. MEMS-reconfigurable metamaterials and antenna applications. *International J. of Antennas and Propag.*, 2014:1–8, Apr. 2014.
- [27] M. Di Renzo, K. Ntontin, J. Song, F. H. Danufane, X. Qian, F. Lazarakis, J. De Rosny, D. T. Phan-Huy, O. Simeone, R. Zhang, M. Debbah, G. Lerosey, M. Fink, S. Tretyakov, and S. Shamai. Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison. *IEEE Open J. of the Commun. Soc.*, 1:798–807, 2020.
- [28] Marco Di Renzo, Alessio Zappone, Merouane Debbah, Mohamed-Slim Alouini, Chau Yuen, Julien de Rosny, and Sergei Tretyakov. Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead. *IEEE Journal on Selected Areas in Communications*, 38(11):2450–2525, 2020.
- [29] Jack Edmonds. Matroids and the greedy algorithm. *Mathematical programming*, 1:127–136, 1971.

- [30] Yasoub Eghbali, Sajad Faramarzi, Shiva Kazemi Taskou, Mohammad Robat Mili, Mehdi Rasti, and Ekram Hossain. Beamforming for STAR-RIS-Aided Integrated Sensing and Communication Using Meta DRL. *IEEE Wireless Communications Letters*, 13(4):919–923, 2024.
- [31] Ahmet M. Elbir, Anastasios Papazafeiropoulos, Pandelis Kourtessis, and Symeon Chatzinotas. Deep channel learning for large intelligent surfaces aided mm-Wave massive MIMO systems. *IEEE Wireless Communications Letters*, 9(9):1447–1451, 2020.
- [32] Roman Fara, Philippe Ratajczak, Dinh-Thuy Phan-Huy, Abdelwaheb Ourir, Marco Di Renzo, and Julien de Rosny. A prototype of reconfigurable intelligent surface with continuous control of the reflection phase. *IEEE Wireless Communications*, 29(1):70–77, 2022.
- [33] Armin Farhadi, Mohsen Moomivand, Shiva Kazemi Taskou, Mohammad Robat Mili, Mehdi Rasti, and Ekram Hossain. A Meta-DDPG Algorithm for Energy and Spectral Efficiency Optimization in STAR-RIS-Aided SWIPT. *IEEE Wireless Communications Letters*, 13(5):1473–1477, 2024.
- [34] Keming Feng, Qisheng Wang, Xiao Li, and Chao-Kai Wen. Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems. *IEEE Wireless Communications Letters*, 9(5):745–749, 2020.
- [35] Jiabao Gao, Caijun Zhong, Xiaoming Chen, Hai Lin, and Zhaoyang Zhang. Unsupervised learning for passive beamforming. *IEEE Communications Letters*, 24(5):1052–1056, 2020.
- [36] Ying Gao, Qingqing Wu, Wen Chen, Yang Liu, Ming Li, and Daniel Benevides da Costa. Irs-aided overloaded multi-antenna systems: Joint user grouping and resource allocation. *IEEE Transactions on Wireless Communications*, 23(8):8297–8313, 2024.
- [37] Xinrong Guan, Qingqing Wu, and Rui Zhang. Anchor-assisted intelligent reflecting surface channel estimation for multiuser communications. In *IEEE Global Communications Conference*, pages 1–6, 2020.
- [38] Dulaj Gunasinghe, Dhanushka Kudathanthirige, and Gayan Amarasuriya Aruma Baduge. The achievable rate performance of STAR-RIS aided Massive MIMO Systems. *IEEE Transactions on Communications*, 72(6):3681–3700, 2024.

- [39] H. Guo, Y. Liang, J. Chen, and E. G. Larsson. Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks. *IEEE Trans. Wireless Commun.*, 19(5):3064–3076, May 2020.
- [40] Huayan Guo, Ying-Chang Liang, Jie Chen, and Erik G. Larsson. Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks. *IEEE Transactions on Wireless Communications*, 19(5):3064–3076, 2020.
- [41] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.
- [42] Huimei Han, Jun Zhao, Dusit Niyato, Marco Di Renzo, and Quoc-Viet Pham. Intelligent reflecting surface aided network: Power control for physical-layer broadcasting. In *IEEE International Conference on Communications (ICC)*, pages 1–7, 2020.
- [43] Shaowei He, Huimin Yang, Yunhui Jiang, Wenjun Deng, and Weiming Zhu. Recent advances in MEMS metasurfaces and their applications on tunable lens. *Micromachines*, 10(8), July 2019.
- [44] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [45] Tianwei Hou, Yuanwei Liu, Zhengyu Song, Xin Sun, Yue Chen, and Lajos Hanzo. Reconfigurable intelligent surface aided NOMA networks. *IEEE Journal on Selected Areas in Communications*, 38(11):2575–2588, 2020.
- [46] Xiaoyan Hu, Christos Masouros, and Kai-Kit Wong. Reconfigurable intelligent surface aided mobile edge computing: From optimization-based to location-only learning-based solutions. *IEEE Transactions on Communications*, 69(6):3709–3725, 2021.
- [47] Chongwen Huang, Ronghong Mo, and Chau Yuen. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE Journal on Selected Areas in Communications*, 38(8):1839–1850, 2020.
- [48] Chongwen Huang, Alessio Zappone, George C. Alexandropoulos, M  rouane Debbah, and Chau Yuen. Reconfigurable intelligent surfaces for energy efficiency in wireless communication. *IEEE Transactions on Wireless Communications*, 18(8):4157–4170, 2019.

- [49] Zixuan Huang, Beixiong Zheng, and Rui Zhang. Transforming fading channel from fast to slow: Irs-assisted high-mobility communication. In *ICC 2021 - IEEE International Conference on Communications*, pages 1–6, 2021.
- [50] Tobias Lindstrøm Jensen and Elisabeth De Carvalho. An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5000–5004, 2020.
- [51] Tao Jiang, Hei Victor Cheng, and Wei Yu. Learning to beamform for intelligent reflecting surface with implicit channel estimate. In *IEEE Global Communications Conference*, pages 1–6, 2020.
- [52] Tao Jiang, Hei Victor Cheng, and Wei Yu. Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation. *IEEE Journal on Selected Areas in Communications*, 39(7):1931–1945, 2021.
- [53] Wei Jiang, Bin Han, Mohammad Asif Habibi, and Hans Dieter Schotten. The road towards 6g: A comprehensive survey. *IEEE Open Journal of the Communications Society*, 2:334–366, 2021.
- [54] Steven M. Kay. *Fundamentals of statistical signal processing*. Prentice-Hall signal processing series. Prentice-Hall PTR, Englewood Cliffs, NJ, 1993 - 1998.
- [55] Kamran Keykhosravi, Benoît Denis, George C. Alexandropoulos, Zhongxia Simon He, Antonio Albanese, Vincenzo Sciancalepore, and Henk Wymeersch. Leveraging RIS-Enabled smart signal propagation for solving infeasible localization problems: Scenarios, key research directions, and open challenges. *IEEE Vehicular Technology Magazine*, 18(2):20–28, 2023.
- [56] Wahab Khawaja, Ozgur Ozdemir, Yavuz Yapici, Fatih Erden, and Ismail Guvenc. Coverage enhancement for NLOS mmWave links using passive reflectors. *IEEE Open Journal of the Communications Society*, 1:263–281, 2020.
- [57] M. Kountouris and D. Gesbert. Memory-based opportunistic multi-user beamforming. In *Proc. Intern. Symp. on Information Theory*, pages 1426–1430, Sept. 2005.
- [58] Ivan Ku, Cheng-Xiang Wang, and John Thompson. Spectral-energy efficiency trade-off in relay-aided cellular networks. *IEEE Transactions on Wireless Communications*, 12(10):4970–4982, 2013.

- [59] Neel Kanth Kundu and Matthew R. McKay. Channel estimation for reconfigurable intelligent surface aided MISO communications: From LMMSE to deep learning solutions. *IEEE Open Journal of the Communications Society*, 2:471–487, 2021.
- [60] Hongyu Li, Wenhao Cai, Yang Liu, Ming Li, Qian Liu, and Qingqing Wu. Intelligent reflecting surface enhanced wideband MIMO-OFDM communications: From practical model to reflection optimization. *IEEE Transactions on Communications*, 69(7):4807–4820, 2021.
- [61] Hongyu Li, Rang Liu, Ming Liy, Qian Liu, and Xuanheng Li. IRS-Enhanced Wideband MU-MISO-OFDM Communication Systems. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2020.
- [62] Jiachen Li, Boning Gao, Zhexuan Yu, Cen Li, Wankai Tang, Le Liang, Xiao Li, Shi Jin, Qiang Cheng, and Tie Jun Cui. Coverage enhancement of 5g commercial network based on reconfigurable intelligent surface. In *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, pages 1–5, 2022.
- [63] Xiaohui Li, Hong Wang, Yunpei Chen, and Shuran Sheng. Joint resource allocation and reflecting precoding design for ris-assisted isac systems. *IEEE Wireless Communications Letters*, 13(4):1193–1197, 2024.
- [64] Minghua Lin. Remarks on Kreĭn’s Inequality. *The Mathematical intelligencer*, 34(1):3–4, 2012.
- [65] Shaoe Lin, Beixiong Zheng, George C. Alexandropoulos, Miaowen Wen, Fangjiong Chen, and Shahid sMumtaz. Adaptive transmission for reconfigurable intelligent surface-assisted OFDM wireless communications. *IEEE Journal on Selected Areas in Communications*, 38(11):2653–2665, 2020.
- [66] Jingxian Liu, Ke Xiong, Yang Lu, Derrick Wing Kwan Ng, Zhangdui Zhong, and Zhu Han. Energy efficiency in secure IRS-aided SWIPT. *IEEE Wireless Communications Letters*, 9(11):1884–1888, 2020.
- [67] Yusha Liu and Osvaldo Simeone. Learning how to transfer from uplink to downlink via hyper-recurrent neural network for FDD massive MIMO. *IEEE Trans. on Wireless Commun.*, 21(10):7975–7989, 2022.
- [68] Wei Lu and Marco Di Renzo. Performance evaluation of relay-aided cellular networks by using stochastic geometry. In *IEEE International Workshop on Computer Aided*

Modeling and Design of Communication Links and Networks (CAMAD), pages 265–269, 2014.

- [69] Xiao Lu, Ekram Hossain, Taniya Shafique, Shaohan Feng, Hai Jiang, and Dusit Niyato. Intelligent reflecting surface enabled covert communications in wireless networks. *IEEE Network*, 34(5):148–155, 2020.
- [70] Lu Lv, Qingqing Wu, Zan Li, Zhiguo Ding, Naofal Al-Dhahir, and Jian Chen. Covert communication in intelligent reflecting surface-assisted NOMa systems: Design, analysis, and optimization. *IEEE Transactions on Wireless Communications*, 21(3):1735–1750, 2022.
- [71] Wenyan Ma, Lipeng Zhu, and Rui Zhang. Passive beamforming for 3-d coverage in irs-assisted communications. *IEEE Wireless Communications Letters*, 11(8):1763–1767, 2022.
- [72] Kaitao Meng, Qingqing Wu, Wen Chen, Enrico Paolini, and Elisabetta Matricardi. Intelligent surface empowered sensing and communication: A novel mutual assistance design. *IEEE Communications Letters*, 27(8):2212–2216, 2023.
- [73] Deepak Mishra and Håkan Johansson. Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4659–4663, 2019.
- [74] Xidong Mu, Yuanwei Liu, Li Guo, Jiaru Lin, and Robert Schober. Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications. *IEEE Transactions on Wireless Communications*, 21(5):3083–3098, 2022.
- [75] Umair Ahmad Mughal, Yazeed Alkhrijah, Ahmad Almadhor, and Chau Yuen. Deep learning for secure uav-assisted ris communication networks. *IEEE Internet of Things Magazine*, 7(2):38–44, 2024.
- [76] Q. U. A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M. S. Alouini. Asymptotic Max-Min SINR analysis of reconfigurable intelligent surface assisted MISO systems. *IEEE Trans. Wireless Commun.*, 19(12):7748–7764, 2020.
- [77] Qurrat-Ul-Ain Nadeem, Hibatallah Alwazani, Abba Kammoun, Anas Chaaban, Mérouane Debbah, and Mohamed-Slim Alouini. Intelligent reflecting surface-assisted multi-user MISO communication: Channel estimation and beamforming design. *IEEE Open Journal of the Communications Society*, 1:661–680, 2020.

- [78] Qurrat-Ul-Ain Nadeem, Abla Kammoun, Anas Chaaban, Mérouane Debbah, and Mohamed-Slim Alouini. Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems. *IEEE Transactions on Wireless Communications*, 19(12):7748–7764, 2020.
- [79] Aviv Navon, Aviv Shamsian, Gal Chechik, and Ethan Fetaya. Learning the pareto front with hypernetworks. *arXiv preprint arXiv:2010.04104*, 2020.
- [80] Mahyar Nemati, Jihong Park, and Jinho Choi. RIS-assisted coverage enhancement in millimeter-wave cellular networks. *IEEE Access*, 8:188171–188185, 2020.
- [81] Konstantinos Ntontin, Marco Di Renzo, and Fotis Lazarakis. On the rate and energy efficiency comparison of reconfigurable intelligent surfaces with relays. In *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2020.
- [82] Cunhua Pan, Hong Ren, Kezhi Wang, Jonas Florentin Kolb, Maged El Kashlan, Ming Chen, Marco Di Renzo, Yang Hao, Jiangzhou Wang, A. Lee Swindlehurst, Xiaohu You, and Lajos Hanzo. Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions. *IEEE Communications Magazine*, 59(6):14–20, 2021.
- [83] Cunhua Pan, Hong Ren, Kezhi Wang, Wei Xu, Maged El Kashlan, Arumugam Nallanathan, and Lajos Hanzo. Multicell MIMO communications relying on intelligent reflecting surfaces. *IEEE Transactions on Wireless Communications*, 19(8):5218–5233, 2020.
- [84] Cunhua Pan, Gui Zhou, Kangda Zhi, Sheng Hong, Tuo Wu, Yijin Pan, Hong Ren, Marco Di Renzo, A. Lee Swindlehurst, Rui Zhang, and Angela Yingjun Zhang. An overview of signal processing techniques for RIS/IRS-aided wireless systems. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–35, 2022.
- [85] Xilong Pei, Haifan Yin, Li Tan, Lin Cao, Zhanpeng Li, Kai Wang, Kun Zhang, and Emil Björnson. RIS-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials. *IEEE Transactions on Communications*, 69(12):8627–8640, 2021.
- [86] Nemanja Stefan Perović, Marco Di Renzo, and Mark F. Flanagan. Channel capacity optimization using reconfigurable intelligent surfaces in indoor mmWave environments. In *IEEE International Conference on Communications (ICC)*, pages 1–7, 2020.

- [87] Nemanja Stefan Perović, Le-Nam Tran, Marco Di Renzo, and Mark F. Flanagan. On the achievable sum-rate of the RIS-aided MIMO broadcast channel. In *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 571–575, 2021.
- [88] Narayan Prasad, Yavuz Yapici, Tao Luo, Junyi Li, and Peter Gaal. Optimal pattern determination in reconfigurable intelligent surface aided communications. In *IEEE Int. Symp. Pers. Indoor Mob. Radio Commun.*, pages 1–7, 2023.
- [89] Shuyi Ren, Kaiming Shen, Xin Li, Xin Chen, and Zhi-Quan Luo. A linear time algorithm for the optimal discrete IRS beamforming. *IEEE Wireless Commun. Letters*, 12(3):496–500, 2023.
- [90] Marco Di Renzo, Merouane Debbah, Dinh-Thuy Phan-Huy, Alessio Zappone, Mohamed-Slim Alouini, Chau Yuen, Vincenzo Sciancalepore, George C. Alexandropoulos, Jakob Hoydis, Haris Gacanin, and et al. Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come, May 2019.
- [91] H. L. Royden. *Real analysis*. Prentice Hall, Boston, 4th ed. edition, 2010.
- [92] Jian Sang, Yifei Yuan, Wankai Tang, Ya Li, Xiao Li, Shi Jin, Qiang Cheng, and Tie Jun Cui. Coverage enhancement by deploying ris in 5g commercial mobile networks: Field trials. *IEEE Wireless Communications*, 31(1):172–180, 2024.
- [93] Martin. Schubert. *Interference Calculus: A General Framework for Interference Management and Network Utility Optimization*. Springer, 2012.
- [94] K. Shen and W. Yu. Fractional programming for communication systems—Part I: Power control and beamforming. *IEEE Trans. Signal Process.*, 66(10):2616–2630, 2018.
- [95] K. Shen and W. Yu. Fractional programming for communication systems—Part II: Uplink scheduling via matching. *IEEE Trans. Signal Process.*, 66(10):2631–2644, 2018.
- [96] Wenhan Shen, Zhijin Qin, and Arumugam Nallanathan. Deep learning for super-resolution channel estimation in reconfigurable intelligent surface aided systems. *IEEE Transactions on Communications*, 71(3):1491–1503, 2023.

- [97] Yifei Shen, Jun Zhang, S.H. Song, and Khaled B. Letaief. Graph neural networks for wireless communications: From theory to practice. *IEEE Trans. on Wireless Commun.*, pages 1–1, 2022.
- [98] Qingjiang Shi, Meisam Razaviyayn, Zhi-Quan Luo, and Chen He. An iteratively Weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel. *IEEE Trans. Signal Process.*, 59(9):4331–4340, 2011.
- [99] Eyad Shtaiwi, Hongliang Zhang, Ahmed Abdelhadi, A. Lee Swindlehurst, Zhu Han, and H. Vincent Poor. Sum-rate maximization for ris-assisted integrated sensing and communication systems with manifold optimization. *IEEE Transactions on Communications*, 71(8):4909–4923, 2023.
- [100] A. Lee Swindlehurst, Gui Zhou, Rang Liu, Cunhua Pan, and Ming Li. Channel estimation with reconfigurable intelligent surfaces—a general framework. *Proceedings of the IEEE*, pages 1–27, 2022.
- [101] A. Taha, M. Alrabeiah, and A. Alkhateeb. Enabling large intelligent surfaces with compressive sensing and deep learning. *IEEE Access*, 9:44304–44321, 2021.
- [102] Wankai Tang, Xiangyu Chen, Ming Zheng Chen, Jun Yan Dai, Yu Han, Shi Jin, Qiang Cheng, Geoffrey Ye Li, and Tie Jun Cui. On channel reciprocity in reconfigurable intelligent surface assisted wireless networks. *IEEE Wireless Communications*, 28(6):94–101, 2021.
- [103] Anum Umer, Ivo Mürsepp, Muhammad Mahtab Alam, and Henk Wymeersch. Reconfigurable intelligent surfaces in 6G radio localization: A survey of recent developments, opportunities, and challenges. *IEEE Communications Surveys & Tutorials*, pages 1–1, 2025.
- [104] Cheng-Xiang Wang, Xiaohu You, Xiqi Gao, Xiuming Zhu, Zixin Li, Chuan Zhang, Haiming Wang, Yongming Huang, Yunfei Chen, Harald Haas, John S. Thompson, Erik G. Larsson, Marco Di Renzo, Wen Tong, Peiying Zhu, Xuemin Shen, H. Vincent Poor, and Lajos Hanzo. On the road to 6g: Visions, requirements, key technologies, and testbeds. *IEEE Communications Surveys & Tutorials*, 25(2):905–974, 2023.
- [105] Peilan Wang, Jun Fang, Huiping Duan, and Hongbin Li. Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems. *IEEE Signal Processing Letters*, 27:905–909, 2020.

- [106] Zhaorui Wang, Liang Liu, and Shuguang Cui. Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis. *IEEE Transactions on Wireless Communications*, 19(10):6607–6620, 2020.
- [107] Zhaorui Wang, Liang Liu, and Shuguang Cui. Intelligent reflecting surface assisted massive MIMO communications. In *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2020.
- [108] Muhua Wei, Jun Wu, Huili Liu, Lin Lin, Zheng Zhang, and Mayuan Sun. Envisioning the potential of 6g use cases—a dual perspective of innovation and business. *IEEE Access*, 13:9831–9843, 2025.
- [109] Qingqing Wu and Rui Zhang. Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming. *IEEE Transactions on Wireless Communications*, 18(11):5394–5409, 2019.
- [110] Qingqing Wu and Rui Zhang. Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts. *IEEE Transactions on Communications*, 68(3):1838–1851, 2020.
- [111] Qingqing Wu, Shuowen Zhang, Beixiong Zheng, Changsheng You, and Rui Zhang. Intelligent reflecting surface-aided wireless communications: A tutorial. *IEEE Transactions on Communications*, 69(5):3313–3351, 2021.
- [112] Qingqing Wu, Beixiong Zheng, Changsheng You, Lipeng Zhu, Kaiming Shen, Xiaodan Shao, Weidong Mei, Boya Di, Hongliang Zhang, Ertugrul Basar, Lingyang Song, Marco Di Renzo, Zhi-Quan Luo, and Rui Zhang. Intelligent surfaces empowered wireless network: Recent advances and the road to 6g. *Proceedings of the IEEE*, pages 1–40, 2024.
- [113] Yue Xiu, Jun Zhao, Wei Sun, Marco Di Renzo, Guan Gui, Zhongpei Zhang, and Ning Wei. Reconfigurable intelligent surfaces aided mmWave NOMA: Joint power allocation, phase shifts, and hybrid beamforming optimization. *IEEE Trans. Wireless Commun.*, 20(12):8393–8409, Dec. 2021.
- [114] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober. Resource allocation for IRS-assisted Full-Duplex cognitive radio systems. *IEEE Trans. on Commun.*, 68(12):7376–7394, 2020.
- [115] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

- [116] Hyunjun Yang, Sunghyun Kim, Hogleom Kim, Seungwoo Bang, Yongwan Kim, Seongkwan Kim, Kyujin Park, Doyle Kwon, and Jungsuek Oh. Beyond limitations of 5g with ris: Field trial in a commercial network, recent advances, and future directions. *IEEE Communications Magazine*, 62(10):132–138, 2024.
- [117] Kai Yang, Yuanming Shi, Yong Zhou, Zhanpeng Yang, Liqun Fu, and Wei Chen. Federated machine learning for intelligent iot via reconfigurable intelligent surface. *IEEE Network*, 34(5):16–22, 2020.
- [118] Yifei Yang, Beixiong Zheng, Shuowen Zhang, and Rui Zhang. Intelligent reflecting surface meets OFDM: Protocol design and rate maximization. *IEEE Transactions on Communications*, 68(7):4522–4535, 2020.
- [119] Changsheng You, Zhenyu Kang, Yong Zeng, and Rui Zhang. Enabling smart reflection in integrated air-ground wireless network: Irs meets uav. *IEEE Wireless Communications*, 28(6):138–144, 2021.
- [120] Li You, Jiayuan Xiong, Yufei Huang, Derrick Wing Kwan Ng, Cunhua Pan, Wenjin Wang, and Xiqi Gao. Reconfigurable intelligent surfaces-assisted multiuser MIMO uplink transmission with partial CSI. *IEEE Transactions on Wireless Communications*, 20(9):5613–5627, 2021.
- [121] Xianghao Yu, Dongfang Xu, and Robert Schober. MISO wireless communication systems via intelligent reflecting surfaces. In *Proc. IEEE Intern. Conf. on Commun. in China*, pages 735–740, Aug. 2019.
- [122] Alessio Zappone, Marco Di Renzo, Farshad Shams, Xuewen Qian, and Mérouane Debbah. Overhead-aware design of reconfigurable intelligent surfaces in smart radio environments. *IEEE Transactions on Wireless Communications*, 20(1):126–141, 2021.
- [123] Hongliang Zhang, Boya Di, Lingyang Song, and Zhu Han. Reconfigurable intelligent surfaces assisted communications with limited phase shifts: How many phase shifts are enough? *IEEE Transactions on Vehicular Technology*, 69(4):4498–4502, 2020.
- [124] Hongliang Zhang, Boya Di, Lingyang Song, and Zhu Han. Reconfigurable intelligent surfaces assisted communications with limited phase shifts: How many phase shifts are enough? *IEEE Transactions on Vehicular Technology*, 69(4):4498–4502, 2020.

- [125] Jing Zhang, Chao-Kai Wen, and Shi Jin. Adaptive MIMO detector based on hyper-network: Design, simulation, and experimental test. *IEEE J. on Sel. Areas Commun.*, 40(1):65–81, 2022.
- [126] Lei Zhang, Yu Wang, Weige Tao, Ziyang Jia, Tiecheng Song, and Cunhua Pan. Intelligent reflecting surface aided MIMO cognitive radio systems. *IEEE Trans. Veh. Technol.*, 69(10):11445–11457, Oct. 2020.
- [127] Shuhang Zhang, Hongliang Zhang, Boya Di, Yunhua Tan, Marco Di Renzo, Zhu Han, H. Vincent Poor, and Lingyang Song. Intelligent omni-surfaces: Ubiquitous wireless transmission by reflective-refractive metasurfaces. *IEEE Transactions on Wireless Communications*, 21(1):219–233, 2022.
- [128] Shuowen Zhang and Rui Zhang. Capacity characterization for intelligent reflecting surface aided MIMO communication. *IEEE Journal on Selected Areas in Communications*, 38(8):1823–1838, 2020.
- [129] Zhongze Zhang, Tao Jiang, and Wei Yu. Learning based user scheduling in reconfigurable intelligent surface assisted multiuser downlink. *IEEE J. of Sel. Topics in Signal Process.*, 16(5):1026–1039, 2022.
- [130] Zhongze Zhang, Tao Jiang, and Wei Yu. Localization with reconfigurable intelligent surface: An active sensing approach. *IEEE Transactions on Wireless Communications*, 23(7):7698–7711, 2024.
- [131] Donglai Zhao, Gang Wang, Jinlong Wang, and Chenxu Wang. Optimal passive beamforming for reconfigurable intelligent surface-assisted communications with discrete phase shifts. *IEEE Wireless Communications Letters*, 13(2):427–431, 2024.
- [132] J. Zhao, M. Chen, M. Chen, Z. Yang, Y. Wang, B. Cao, and M. Shikh-Bahaei. Energy efficient Full-Duplex communication systems with reconfigurable intelligent surface. In *IEEE Veh. Technol. Conf.*, pages 1–5, 2020.
- [133] Ming-Min Zhao, An Liu, Yubo Wan, and Rui Zhang. Two-timescale beamforming optimization for intelligent reflecting surface aided multiuser communication with QoS constraints. *IEEE Transactions on Wireless Communications*, 20(9):6179–6194, 2021.
- [134] Beixiong Zheng, Changsheng You, Weidong Mei, and Rui Zhang. A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface

- aided wireless communications. *IEEE Commun. Surveys & Tut.*, 24(2):1035–1071, 2022.
- [135] Beixiong Zheng, Changsheng You, and Rui Zhang. Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design. *IEEE Transactions on Wireless Communications*, 20(7):4513–4526, 2021.
 - [136] Beixiong Zheng and Rui Zhang. Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization. *IEEE Wireless Communications Letters*, 9(4):518–522, 2020.
 - [137] Kangda Zhi, Cunhua Pan, Hong Ren, and Kezhi Wang. Ergodic rate analysis of reconfigurable intelligent surface-aided massive MIMO systems with ZF detectors. *IEEE Communications Letters*, pages 1–1, 2021.
 - [138] Kai Zhong, Jinfeng Hu, Huiyong Li, Ren Wang, Dongxu An, Gangyong Zhu, Kah Chan Teh, CunhuaSenior Pan, and Yonina C. Eldar. RIS-aided beamforming design for MIMO systems via unified manifold optimization. *IEEE Transactions on Vehicular Technology*, pages 1–11, 2024.
 - [139] Gui Zhou, Cunhua Pan, Hong Ren, Petar Popovski, and A. Lee Swindlehurst. Channel estimation for RIS-aided multiuser millimeter-wave systems. *IEEE Transactions on Signal Processing*, 70:1478–1492, 2022.
 - [140] Qi Zhu, Hongyu Li, Rang Liu, Ming Li, and Qian Liu. Hybrid beamforming and passive reflection design for RIS-assisted mmWave MIMO systems. In *IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, 2021.
 - [141] Özgecan Özdoğan, Emil Björnson, and Erik G. Larsson. Using intelligent reflecting surfaces for rank improvement in MIMO communications. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9160–9164, 2020.

APPENDICES

Appendix A

Proofs

A.1 Lemma 1

The following lemma illustrates a practical sufficient condition for the $\mathbb{E}[U_k] < \infty$ assumption of Proposition 4.1.

Lemma 1. *If the components of the $\mathbf{h}_{r,k}^{DL}$, $\mathbf{h}_{d,k}^{DL}$, and $\mathbf{G}^{DL}$ channels satisfy $\mathbb{E}[|h_{r,k,n}^{DL}|] < \infty$, $\mathbb{E}[|h_{d,k,m}^{DL}|] < \infty$, and $\mathbb{E}[|g_{m,n}^{DL}|] < \infty$, then $\mathbb{E}[U_k] < \infty$.*

Proof. First,

$$\begin{aligned} U_k &= \log_2 \left(1 + \frac{P_{max}^{DL}}{\sigma_{DL}^2} (||\mathbf{h}_{d,k}^{DL}|| + \sum_{n=1}^N |h_{r,k,n}^{DL}| ||\mathbf{g}_n^{DL}||)^2 \right) \\ &\stackrel{(a)}{\leq} \frac{4\sqrt{P_{max}^{DL}}}{\sigma_{DL}} (||\mathbf{h}_{d,k}^{DL}|| + \sum_{n=1}^N |h_{r,k,n}^{DL}| ||\mathbf{g}_n^{DL}||) \\ &\leq \frac{4\sqrt{P_{max}^{DL}}}{\sigma_{DL}} \left(\sum_{m=1}^M |h_{d,k,m}^{DL}| + \sum_{n=1}^N \sum_{m=1}^M |h_{r,k,n}^{DL}| |g_{m,n}^{DL}| \right), \end{aligned}$$

where (a) follows as $\log_2(1 + x^2) \leq 4|x|$. Then, since $\mathbb{E}[|h_{r,k,n}^{DL}|] < \infty$, $\mathbb{E}[|h_{d,k,m}^{DL}|] < \infty$, and $\mathbb{E}[|g_{m,n}^{DL}|] < \infty$ and by independence of $h_{r,k,n}^{DL}$ and $g_{m,n}^{DL}$ channel components, we conclude that $\mathbb{E}[U_k] < \infty$. $\square$

A sufficient condition for the $\mathbb{E}[U_k] < \infty$ above is that the channels have finite second moments. Thus, Lemma 1 above is satisfied by Rician, Rayleigh and Nakagami fading models.

A.2 Lemma 2

To prove Proposition 4.1, we also establish a lemma about continuity of the objective in (4.5) (as a function of $\mathbf{W}$ and Θ) and show that the conditional problem in (4.5) attains a maximum for almost every γ and $\mathbf{y}$.

Lemma 2. *If $\sum_{k=1}^K \mathbb{E}[U_k] < \infty$, then the objective of (4.5), $\mathbb{E}[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \Theta) | \mathbf{Z} = (\mathbf{y}, \gamma)]$, is a continuous function of $\mathbf{W}$ and Θ for almost all γ and $\mathbf{y}$ and the conditional WSR problem in (4.5) attains a maximum for almost every γ and $\mathbf{y}$.*

Proof. We first note that $\mathbb{E}[\sum_{k=1}^K U_k] < \infty$ implies that, for a set $A \in \sigma(\mathbf{Z})$ of measure 1, $\mathbb{E}[\sum_{k=1}^K U_k | \mathbf{Z} = \mathbf{z}] < \infty$. The bound in (4.4) then implies that, over such set A ,

$$\mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \Theta) | \mathbf{Z} = \mathbf{z}\right] \leq \mathbb{E}\left[\sum_{k=1}^K U_k | \mathbf{Z} = \mathbf{z}\right],$$

where U_k 's do *not* depend on the choice of $\mathbf{W}$ and Θ . Suppose that $\mathbf{W}_l \rightarrow \mathbf{W}$ and $\Theta_l \rightarrow \Theta$, then by Lebesgue's dominated convergence theorem [9, Theorem 11.21], over A ,

$$\lim_{l \rightarrow \infty} \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}_l, \Theta_l) | \mathbf{Z} = \mathbf{z}\right] = \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \Theta) | \mathbf{Z} = \mathbf{z}\right].$$

To show that problem in (4.5) attains a maximum for almost every γ and $\mathbf{y}$, we note that for every γ and $\mathbf{y}$, the search space for $\mathbf{W}$ and Θ of Problem (4.5) is the intersection of (4.5b) and (4.5c) and thus compact. $\square$

A.3 Proof of Proposition 4.1

1. To prove part (a) of Proposition 4.1, we construct a solution of the problem in (4.3) using a family of solutions that optimize the conditional problem in (4.5). Let $\mathbf{S} \triangleq (\mathbf{W}, \Theta)$ and let $\mathbf{1}_A$ be indicator function of a set A . We first note that by lemma 2, there exists a measure 1 set $A \in \sigma(\mathbf{Z})$ such that the function $f(\mathbf{S}, \mathbf{Z}) \triangleq \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \Theta) | \mathbf{Z} = \mathbf{z}\right] \mathbf{1}_A$ is a continuous function of $\mathbf{S} = (\mathbf{W}, \Theta)$ for each $\mathbf{z}$. Moreover, for each $\mathbf{s} \in \mathbf{S}$, $f(\mathbf{s}, \cdot)$ is measurable. It follows that $f : \mathbf{S} \times \mathbf{Z} \rightarrow \mathbb{R}$ is jointly measurable [9, Lemma 4.51]. By [17, Corollary 1], there exists a measurable mapping

$\tilde{\mathcal{F}}(\mathbf{Z})$ such that $\mathbb{E}[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}] \mathbf{1}_A = \max_{\boldsymbol{\Theta}, \mathbf{W}} \mathbb{E}[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \boldsymbol{\Theta}) | \mathbf{Z} = \mathbf{z}] \mathbf{1}_A$. Finally, we show that $\tilde{\mathcal{F}}(\mathbf{Z})$ can solve (4.3) as over the domain $\mathcal{Z}$ of $\mathbf{Z}$,

$$\begin{aligned}
& \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \boldsymbol{\Theta})\right] \\
&= \int_{\mathcal{Z}} \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \boldsymbol{\Theta}) | \mathbf{Z} = \mathbf{z}\right] \mu(d\mathbf{z}) \\
&= \int_{\mathcal{Z}} \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathbf{W}, \boldsymbol{\Theta}) | \mathbf{Z} = \mathbf{z}\right] \mathbf{1}_A \mu(d\mathbf{z}) \\
&\leq \int_{\mathcal{Z}} \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}\right] \mathbf{1}_A \mu(d\mathbf{z}) \\
&= \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z}))\right]
\end{aligned}$$

□

2. To contradict the proposition, suppose that there exists a subset $C \in \sigma(\mathbf{Z})$ of positive measure, such that for $\mathbf{z} \in C$

$$\begin{aligned}
& \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathcal{F}^*(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}\right] \\
&< \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}\right].
\end{aligned} \tag{A.1}$$

Then the objective of the adaptive beamforming configuration problem in (4.3) can be expressed as [91]:

$$\begin{aligned}
& \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathcal{F}^*(\mathbf{Z}))\right], \\
&= \int_{\mathcal{Z}} \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathcal{F}^*(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}\right] \mu(d\mathbf{z}), \\
&\stackrel{(a)}{<} \int_{\mathcal{Z} \setminus C} \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\mathcal{F}^*(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}\right] \mu(d\mathbf{z}), \\
&+ \int_C \mathbb{E}\left[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z})) | \mathbf{Z} = \mathbf{z}\right] \mu(d\mathbf{z}),
\end{aligned} \tag{A.2}$$

where the inequality (a) holds if C has positive measure. However, this contradicts the

optimality of $\mathcal{F}^*(\mathbf{Z})$ for the problem in (4.3). Therefore C in (A.1) has measure 0, and

$$\begin{aligned} \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\tilde{\mathcal{F}}(\mathbf{Z})) \middle| \mathbf{Z} = \mathbf{z} \right] \\ = \mathbb{E} \left[\sum_{k=1}^K \Gamma_k R_k(\mathcal{F}^*(\mathbf{Z})) \middle| \mathbf{Z} = \mathbf{z} \right] \end{aligned}$$

holds for almost all $\mathbf{z}$.

A.4 Proof of Proposition 5.1

For $\mathbf{\Gamma}$ with $\Gamma_n \in \{\alpha(\phi_\ell)e^{-j\phi_\ell}\}_{\ell=1}^L$, (P1) is then equivalent to solving

$$\max_{\theta} \max_{\mathbf{\Gamma}} \Re\{e^{-j\theta}(h_{\text{TR}} + \sum_{n=1}^N g_n f_n \Gamma_n)\}.$$

For any given θ (including an optimal θ), the inner maximization is equivalent to

$$\Re\{e^{-j\theta}h_{\text{TR}}\} + \max_{\mathbf{\Gamma}} \sum_{n=1}^N |g_n f_n \Gamma_n| \cos(\arg g_n f_n \Gamma_n - \theta). \quad (\text{A.3})$$

Say an optimal solution of (A.3) yields effective channels $h_n^{\text{OPT}} \triangleq \Gamma_n^{\text{OPT}} g_n f_n$. If

$$\{h_1^{\text{OPT}}, \dots, h_N^{\text{OPT}}\} \not\subseteq \mathcal{S}(\theta, \theta_{\max})$$

for all $\theta_{\max} < \pi$, then for at least one element $a \in \{1, \dots, N\}$, $\cos(\arg h_a^{\text{OPT}} - \theta)$ is negative or zero. Since the phase shifts are in increments of $2\pi/L$ with $L \geq 4$, then there exists at least one effective channel $h_a^{\text{ALT}} \in \mathcal{V}_a$ such that $\cos(\arg h_a^{\text{ALT}} - \theta)$ is positive, thus contradicting the optimality of h_a^{OPT} .

A.5 Proof of Proposition 5.2

(a) For any $\mathcal{C} \subseteq \mathcal{S}(\theta_c, \theta_{\max})$ with $\theta_{\max} < \pi$,

$$\begin{aligned} \sum_{h \in \mathcal{C}} \cos(\theta_{\max}/2) |h| &\leq \sum_{h \in \mathcal{C}} \Re\{h e^{-j\theta_c}\} \\ &\leq \left| \sum_{h \in \mathcal{C}} h e^{-j\theta_c} \right| \leq \sum_{h \in \mathcal{C}} |h|. \end{aligned} \quad (\text{A.4})$$

Then (5.4) can be obtained by applying $r(x) = \log_2(1 + \frac{P}{\sigma^2}x^2)$ to the appropriate steps of (A.4). Moreover, $f_u(\mathcal{C})$ and $f_l(\mathcal{C})$ are monotone as they are both increasing functions of the sum $\sum_{h \in \mathcal{C}} |h|$. To show the monotonicity of $R(\mathcal{C})$ when restricted to $\mathcal{C} \subseteq \mathcal{S}(\theta_c, \pi/2)$, we note that for any $s_a, s_b \in \mathcal{S}(\theta_c, \pi/2)$, $|s_a + s_b|^2 = |s_a|^2 + |s_b|^2 + 2|s_a||s_b|\cos(\arg(s_a\bar{s}_b)) \geq \max(|s_a|^2, |s_b|^2)$.

(b) $\mathcal{I}_{\mathcal{S}}$ is the collection of independent sets of a partition matroid formed over the ground set $\mathcal{S} = \mathcal{S}(\theta_c, \theta_{\max})$ and hence $(\mathcal{S}, \mathcal{I}_{\mathcal{S}})$ is a matroid. Subsequently, Algorithm 2 applied to the upper and lower bound problems is equivalent to using Algorithm 2 to maximize $\sum_{h \in \mathcal{C}} |h|$ over $\mathcal{C} \in \mathcal{I}_{\mathcal{S}}$, which is a matroid-constrained modular function (i.e. submodular with curvature $c = 0$). Hence, Algorithm 2 yields a solution $\tilde{\mathcal{C}}_{\mathcal{S}}$ that is optimal for both problems [24]. By the monotonicity of $f_i(\mathcal{C})$ for $i \in \{u, l\}$ and $\mathcal{F}_{\mathcal{S}} \neq \emptyset$, it follows that $\tilde{\mathcal{C}}_{\mathcal{S}} \cap \mathcal{V}_n \neq \emptyset$ and hence $\tilde{\mathcal{C}}_{\mathcal{S}} \in \mathcal{F}_{\mathcal{S}}$. Then $\max_{\mathcal{C} \in \mathcal{I}_{\mathcal{S}}} f_l(\mathcal{C}) = f_l(\tilde{\mathcal{C}}_{\mathcal{S}}) \leq R(\tilde{\mathcal{C}}_{\mathcal{S}}) \leq \max_{\mathcal{C} \in \mathcal{F}_{\mathcal{S}}} R(\mathcal{C}) \leq \max_{\mathcal{C} \in \mathcal{I}_{\mathcal{S}}} f_u(\mathcal{C}) = f_u(\tilde{\mathcal{C}}_{\mathcal{S}})$, proving (5.5), where the first inequality follows by (5.4) and the last inequality follows because $\mathcal{F}_{\mathcal{S}} \subseteq \mathcal{I}_{\mathcal{S}}$ and $R(\mathcal{C}) \leq f_u(\mathcal{C})$.

A.6 Proof of Proposition 5.3

Let $\mathbf{w}_0 = \mathbf{h}_{\text{TR}}$ and $\mathbf{w}_n = [f_n g_{1,n}, \dots, f_n g_{M,n}]^T$. Then, for $\mathbf{u} = [u_1, \dots, u_M]^T \in \mathbb{C}^{M \times 1}$ such that $\|\mathbf{u}\| = 1$, (5.12) (correspondingly (P6)) is equivalent to

$$\max_{\mathbf{u}: \|\mathbf{u}\|=1} \max_{\mathbf{\Gamma}} \Re\{\langle \mathbf{w}_0 + \sum_{n=1}^N \Gamma_n \mathbf{w}_n, \mathbf{u} \rangle\}. \quad (\text{A.5})$$

For any $\mathbf{u}$, the inner maximization is

$$\max_{\mathbf{\Gamma}} \sum_{n=1}^N \Re\{\langle \Gamma_n \mathbf{w}_n, \mathbf{u} \rangle\} \quad (\text{A.6})$$

Let an optimal solution of (A.6) be $\Gamma_1^{\text{OPT}}, \dots, \Gamma_N^{\text{OPT}}$. Suppose $\{\Gamma_1^{\text{OPT}} \mathbf{w}_1, \dots, \Gamma_N^{\text{OPT}} \mathbf{w}_N\} \not\subseteq \mathcal{S}(\mathbf{u}, \theta_{\max})$ for all $\theta_{\max} < \pi$. Then, there exists at least one element $n' \in \{1, \dots, N\}$, such that $\Re\{\langle \Gamma_{n'}^{\text{OPT}} \mathbf{w}_{n'}, \mathbf{u} \rangle\} = \Re\{\bar{\Gamma}_{n'}^{\text{OPT}} \langle \mathbf{w}_{n'}, \mathbf{u} \rangle\}$ is negative or zero. However, since $\Gamma_{n'} \in \{\gamma_1, \dots, \gamma_L\}$ can be selected from $L \geq 4$ possible phase shifts in increments of $2\pi/L$, there exists at least one $\Gamma_{n'}^{\text{ALT}}$ such that $\Re\{\langle \Gamma_{n'}^{\text{ALT}} \mathbf{w}_{n'}, \mathbf{u} \rangle\} = \Re\{\bar{\Gamma}_{n'}^{\text{ALT}} \langle \mathbf{w}_{n'}, \mathbf{u} \rangle\}$ is positive which contradicts the optimality $\Gamma_{n'}^{\text{OPT}}$.

A.7 Proof of Proposition 5.4

Let $\mathbf{u} = [u_1, \dots, u_M]^T \in \mathbb{C}^{M \times 1}$ such that $\|\mathbf{u}\| = 1$, then an upper bound on $Q(\mathcal{C})$ can be obtained as follows:

$$\begin{aligned}
 Q(\mathcal{C}) &= \left\| \sum_{\mathbf{h} \in \mathcal{C}} \mathbf{h} \right\| = \max_{\|\mathbf{u}\|=1} \Re\left\{ \left\langle \sum_{\mathbf{h} \in \mathcal{C}} \mathbf{h}, \mathbf{u} \right\rangle \right\} \\
 &= \max_{\|\mathbf{u}\|=1} \sum_{\mathbf{h} \in \mathcal{C}} \Re\{ \langle \mathbf{h}, \mathbf{u} \rangle \} \\
 &\leq \sum_{\mathbf{h} \in \mathcal{C}} \|\mathbf{h}\| \\
 &\triangleq g_u(\mathcal{C}).
 \end{aligned} \tag{A.7}$$

For the lower bound, we note that

$$\begin{aligned}
 Q(\mathcal{C}) &= \max_{\|\mathbf{u}\|=1} \Re\left\{ \left\langle \sum_{\mathbf{h} \in \mathcal{C}} \mathbf{h}, \mathbf{u} \right\rangle \right\} \geq \Re\left\{ \left\langle \sum_{\mathbf{h} \in \mathcal{C}} \mathbf{h}, \mathbf{u}_c \right\rangle \right\} \\
 &= \sum_{\mathbf{h} \in \mathcal{C}} \Re\{ \langle \mathbf{h}, \mathbf{u}_c \rangle \} \\
 &\stackrel{(a)}{\geq} \sum_{\mathbf{h} \in \mathcal{C}} \|\mathbf{h}\| \cos(\theta_{\max}/2) \\
 &\triangleq g_l(\mathcal{C}),
 \end{aligned} \tag{A.8}$$

where (a) follows since $\mathbf{h} \in \mathcal{C} \subseteq \mathcal{S}(\mathbf{u}_c, \theta_{\max})$. Both the monotonicity and modularity of $g_u(\mathcal{C})$ (correspondingly, $g_l(\mathcal{C})$) follow since each of the terms in $g_u(\mathcal{C})$ (correspondingly, $g_l(\mathcal{C})$) is a sum of non-negative terms proportional to $\|\mathbf{h}\|$.

Finally, we show the monotonicity of $Q(\mathcal{C})$ for $\mathcal{C} \subseteq \mathcal{S}(\mathbf{u}_c, \pi/2)$. Consider $\mathbf{h}_a, \mathbf{h}_b \in \mathcal{S}(\mathbf{u}_c, \pi/2)$, it follows that $\Phi_{\mathbf{h}_a, \mathbf{u}_c} = \cos^{-1}\left(\frac{\Re\{\langle \mathbf{h}_a, \mathbf{u}_c \rangle\}}{\|\mathbf{h}_a\| \|\mathbf{u}_c\|}\right) \leq \pi/4$ and $\Phi_{\mathbf{h}_b, \mathbf{u}_c} = \cos^{-1}\left(\frac{\Re\{\langle \mathbf{h}_b, \mathbf{u}_c \rangle\}}{\|\mathbf{h}_b\| \|\mathbf{u}_c\|}\right) \leq \pi/4$. By Kreĭn's Inequality [64], $\cos^{-1}\left(\frac{\Re\{\langle \mathbf{h}_a, \mathbf{h}_b \rangle\}}{\|\mathbf{h}_a\| \|\mathbf{h}_b\|}\right) = \Phi_{\mathbf{h}_a, \mathbf{h}_b} \leq \Phi_{\mathbf{h}_a, \mathbf{u}_c} + \Phi_{\mathbf{h}_b, \mathbf{u}_c} \leq \pi/2$ and therefore $\Re\{\langle \mathbf{h}_a, \mathbf{h}_b \rangle\} \geq 0$. Since $\|\mathbf{h}_a + \mathbf{h}_b\|^2 = \|\mathbf{h}_a\|^2 + \|\mathbf{h}_b\|^2 + 2\Re\{\langle \mathbf{h}_a, \mathbf{h}_b \rangle\} \geq \max(\|\mathbf{h}_a\|^2, \|\mathbf{h}_b\|^2)$, it follows that $Q(\mathcal{C})$ is monotone in this case.

Appendix B

Static-GNN2 implementation

Let $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_K$ be the matrices that collect pilot information for users $1, \dots, K$, respectively (See Section 4.6.2 for more information). The pilot information matrices $\{\hat{\mathbf{Y}}_k\}_{k=1}^K$ along with the additional information vector $\boldsymbol{\eta}$ (comprised of the user weights vectors $\boldsymbol{\Gamma}$, and if applicable, the user location information $\{\mathbf{p}_k\}_{k \in \mathcal{K}}$) are fed to the BFN. For user $i \in \mathcal{K}$, the node representation is initialized as $\mathbf{d}_i^{(0)} = \mathbf{q}_{\text{in}}^{\text{user}}([\text{vec}(\Re\{\hat{\mathbf{Y}}_k\})^T, \text{vec}(\Im\{\hat{\mathbf{Y}}_k\})^T, \boldsymbol{\eta}^T]^T)$. The network $\mathbf{q}_{\text{in}}^{\text{user}}$ is described in part B of Table 4.1. For the RIS node ($i = K + 1$), the initial node representation is $\mathbf{d}_{K+1}^{(0)} = \mathbf{q}_{\text{in}}^{\text{RIS}}(\text{MEAN}(\{\mathbf{d}_i^{(0)}\}_{i=1}^K))$, where $\text{MEAN}(\cdot)$ denotes the element-wise mean function.

For each GNN layer l , the RIS node includes an RIS representation aggregation network $\mathbf{q}_{a,I}^{(l)}$ and a user representation aggregation network $\mathbf{q}_{a,I,U}^{(l)}$. The RIS node combination network is $\mathbf{q}_{c,I}^{(l)}$, and the pooling function is the element-wise mean function. The K user nodes share the same aggregation and combination networks. Specifically, for all user nodes, the user representation aggregation network is $\mathbf{q}_{a,U,U}^{(l)}$, and the RIS representation aggregation network reuses $\mathbf{q}_{a,I}^{(l)}$ from the RIS node. The combination network associated with the user nodes is $\mathbf{q}_{c,U}^{(l)}$. For user nodes, the pooling function is the element-wise max function. Finally, $\mathbf{q}_o^{(I)}$ and $\mathbf{q}_o^{(U)}$ process the node representations at the final GNN layer to generate the (real and imaginary components of) the RIS configuration and K beamforming vectors, respectively.