Imperial College London
Department of Electrical and Electronic Engineering

# Optimization and Reinforcement Learning for Emerging Multi-antenna Wireless Networks

Zhenyuan Feng
冯震远

Supervised by Prof. Bruno Clerckx

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

# Declaration of Originality

I declare that this thesis represents my original research. Main parts of it have appeared in relevant journals. Any external sources utilized have been duly acknowledged.

# Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

*To my wife and parents*

# Abstract

Multi-antenna wireless networks such as wireless power networks (WPN) and wireless communication networks (WCN) have attracted great attention in academia and industry. This thesis is concerned with the optimization and reinforcement learning for emerging multi-antenna wireless networks. Various scenarios and algorithms are elaborated and evaluated.

First, we design a novel WPN based on the joint optimization of beamforming and waveform with intelligent reflecting surfaces (IRS). Generalized multi-user and low-complexity single-user algorithms are demonstrated based on alternating optimization (AO) framework, successive convex approximation (SCA), and semidefinite relaxation (SDR) to maximize the weighted sum output direct current (DC). The energy harvester nonlinearity is explored and two IRS deployment schemes, namely frequency selective IRS (FS-IRS) and frequency flat IRS (FF-IRS), are modeled and analyzed. IRS can provide a promising passive beamforming gain on output DC power over conventional WPT designs and significantly influence the waveform design.

Second, we consider a massive multi-input multi-output (MIMO) WCN under imperfect channel state information at the transmitter (CSIT). In order to tackle the multiuser interference and CSIT imperfections, we develop a multi-agent codebook-based deep reinforcement learning (DRL) framework where the transmit and receive beamforming are jointly designed to maximize the average information rate of all users. Leveraging this deep Q-network (DQN)-based framework, interference management is explored and three multi-agent DQN-based schemes, namely the distributed-learning-distributed-processing scheme, partial-distributed-learning-distributed-processing, and central-learning-distributed-processing scheme, are proposed and analyzed. This work demonstrates the inherent effectiveness and robustness of the proposed designs in the presence of imperfect CSIT.

Finally, we extend the scope of research from codebook-based beamformer prediction to continuous action space. A Multi-agent Deep Deterministic Policy Gradient (DDPG) based strategy is proposed to reduce the effect caused by inaccurate codebook design.

Simulation results demonstrate that the DDPG-based strategy is a very powerful and promising technique for multi-user massive MIMO beamforming.

# Acknowledgements

First, I want to extend my heartfelt thanks to Prof. Bruno Clerckx, my supervisor, for his unwavering responsive emails, timely guidance, and consistent encouragement throughout my PhD journey.

Second, I am also truly thankful to Yunnuo Xu, Jiawei Xu, Hongyu Li, Yumeng Zhang, Yuanwen Liu, Sibo Zhang, Xi Yao, Yang Zhao, Matteo Nerini, Dr. Longfei Yin, Dr. Onur Dizdar, Prof. Yijie Mao, Prof. Shanpu Shen, and other colleagues and friends in Communications and Signal Processing group at the Department of Electrical and Electronic Engineering for the valueless company and wonderful card games and Karaoke time. Also, I would like to thank Simon, Misha, and Alessio, Peimin for their selfless mentorship in Meta.

Finally, special thanks to my girl (now my proud wife) Stella (He Zhao) for every unforgettable second and moment we had. Being able to create a better life for our family has always been my motivation throughout my whole PhD. Also, thanks to my parents for their continuous support and endless love and faith in every decision I made.

I'll grow up through self-discovery and search for meaning.

*Zhenyuan Feng*

*London, July 2024*

# Abbreviations

**4G** fourth generation

**5G** fifth generation

**6G** sixth generation

**AO** alternating optimization

**BC** broadcast channel

**BS** base station

**CDF** cumulative distribution function

**CDRL** central-learning-distributed-processing DRL-based scheme

**CE** conversion efficiency

**CSCG** circularly symmetric complex Gaussian

**CSI** channel state information

**CSIT** channel state information at the transmitter

**CW** continuous caveform

**DC** direct current

**DDPG** Deep Deterministic Policy Gradient

**DDRL** distributed-learning-distributed-processing DRL-based scheme

**DL** deep learning

**DNN** deep neural network

**DoF** degree of freedom

**DQN** deep Q-network

**DRL** deep reinforcement learning

**EPTE** end-to-end power transfer efficiency

**ER** energy receiver

**ET** energy transmitter

**EVD** eigenvalue decomposition

**FDD** frequency division duplexing

**FIFO** first-input-first-output

**FF-IRS** frequency-flat IRS

**FS-IRS** frequency-selective IRS

**GBS** Greedy Beam Selection

**GP** Geometric Program

**IoT** Internet of Things

**IRS** intelligent reflecting surfaces

**KKT** Karush-Kuhn-Tucker

**LPF** low pass filter

**MAC** media access control

**MDP** Markov decision process

**MIMO** Multiple-Input-Multiple-Output

**MISO** multi-input-single-output

**ML** machine learning

**MU** multi-user

**OFDM** orthogonal frequency division multiplexing

**PAPR** peak-to-average-ratio

**PDRL** partial-distributed-learning-distributed-processing DRL-based scheme

**POW** power-optimized waveforms

**RE** reflection element

**REs** reflection elements

**RF** ratio frequency

**RFID** Radio Frequency Identification

**RIS** reconfigurable intelligent surfaces

**RL** reinforcement learning

**RMS** root mean square

**RSMA** Rate-Splitting Multiple Access

**SAH** sample-and-hold

**SCA** successive convex approximation

**SDMA** Space Division Multiple Access

**SDP** Semi-definite Program

**SDR** semidefinite relaxation

**SISO** single-input-single-output

**SU** single-use

**TDD** time division duplexing

**WCN** wireless communication networks

**WPN** wireless power networks

**WPT** wireless power transfer

**UAV** unmanned aerial vehicles

**ZF-CI** zero-forcing channel-inversion

**ZF-CI PCSI** ZF-CI method with perfect CSI

# Contents

**Declaration of Originality** 1

**Copyright Declaration** 3

**Abstract** 7

**Acknowledgements** 9

**Abbreviations** 11

**1 Introduction** 26

1.1 Motivation and Outline . . . . . 27

1.1.1 Motivation . . . . . 27

1.1.2 Outline . . . . . 30

1.2 List of Contributions . . . . . 31

1.2.1 For Wireless Power Networks . . . . . 31

1.2.2 For Wireless Communication Networks . . . . . 32

1.3 Publications . . . . . 33

1.4 Notation . . . . . 33

**2 Fundamentals and Literature Review** **34**

2.1 Some Fundamentals of WPT . . . . . 34

2.1.1 WPT Waveform Design . . . . . 35

2.1.2 Rectenna . . . . . 37

2.2 Literature Review for WPT . . . . . 40

2.2.1 Multi-antenna Wireless Power Networks . . . . . 40

2.3 Some Fundamentals of Multiuser MIMO . . . . . 44

2.3.1 Multi-user MIMO Systems . . . . . 45

2.3.2 Broadcast Channel . . . . . 45

2.3.3 Signal Model . . . . . 46

2.3.4 Channel Model . . . . . 47

2.3.5 CSIT Acquisition . . . . . 48

2.3.6 CSIT Types . . . . . 48

2.3.7 Performance Metric . . . . . 49

2.3.8 Precoding Techniques . . . . . 50

2.4 Literature Review of Multi-antenna Wireless Communication Networks 52

2.5 Mathematical Tools . . . . . 56

2.5.1 Convex Optimization . . . . . 56

2.5.2 Reinforcement Learning . . . . . 57

**3 Waveform and Beamforming Design for IRS-aided Wireless Power Networks** **59**

3.1 Introduction . . . . . 60

3.2 System Model . . . . . 63

3.2.1 Transmit Signal . . . . . 63

3.2.2 Channel Model and Reflection Pattern . . . . . 64

3.2.3 Receive Signal . . . . . 66

3.3 Joint Design of Waveform and Reflection Elements for Multi-user Scenario 66

3.3.1 Passive Beamforming for Multi-User FF-IRS . . . . . 68

3.3.2 Passive Beamforming for Multi-User FS-IRS . . . . . 72

3.3.3 Waveform Design for Multi-User FF-IRS and Multi-User FS-IRS 75

3.3.4 Alternating Optimization . . . . . 77

3.3.5 Convergence . . . . . 78

3.4 Low Complexity Design for Single-user Scenario . . . . . 79

3.4.1 Passive Beamforming for Single-User FS-IRS . . . . . 80

3.4.2 Frequency Domain Power Allocation for Single-User FS-IRS . 81

3.5 Large Scale Analysis of SU FS-IRS . . . . . 82

3.6 Discussion on Multi-Antenna Scenario . . . . . 83

3.7 Discussion on Discrete-Phase . . . . . 84

3.8 Simulation Result and Analysis . . . . . 84

3.8.1 Single-User . . . 86
3.8.2 Convergence Analysis . . . 90
3.8.3 Bandwidth . . . 91
3.8.4 Multi-User . . . 92
3.8.5 Discrete-Phase . . . 95
3.9 Summary . . . 98

**4 A Novel Multi-agent DQN-based Beamforming Design for Massive MIMO Wireless Communication Networks 100**

4.1 Introduction . . . 101
4.2 System Model . . . 103
4.2.1 Channel Model . . . 105
4.2.2 Problem Formulation . . . 106
4.3 Multi-agent DRL-based Discrete Beamforming Optimization . . . 107
4.3.1 A Brief Overview of DQN . . . 107
4.3.2 The Distributed-learning-distributed-processing DRL-based Algorithm . . . 110
4.3.3 The Low-complexity Centralized-learning-distributed-processing DRL-based Algorithm . . . 118
4.3.4 Bridging the DDRL and CDRL: Partial-distributed-learning-distributed-processing Scheme . . . 119
4.4 Simulation Result and Analysis . . . 121

4.4.1 Simulation Setup . . . . . 121
4.4.2 DDRL vs CDRL . . . . . 123
4.4.3 Multi-antenna and Multi-stream . . . . . 129
4.4.4 Reward, Penalty Analysis and Statistical Test . . . . . 131
4.5 Summary . . . . . 134

**5 A Novel Multi-agent DDPG-based Beamforming Design for Massive MIMO Wireless Communication Networks 137**

5.1 Introduction . . . . . 138
5.2 Multi-agent DRL-based Continuous Beamforming Optimization . . . 139
5.2.1 A Brief Overview of DDPG . . . . . 140
5.2.2 Proposed DRL Algorithm . . . . . 142
5.3 Simulation Result and Analysis . . . . . 150
5.3.1 Single User . . . . . 150
5.3.2 Multi User . . . . . 151
5.3.3 Different Reward Function . . . . . 152
5.4 Summary . . . . . 153

**6 Conlusion 156**

6.1 Summary of Thesis Achievements . . . . . 156
6.2 Future Work . . . . . 157

**A Proof of Propositions** **161**

A.1 Proof of Proposition 1 . . . . . 161

A.2 Proof of Proposition 2 . . . . . 162

**References** **163**

# List of Figures


2.1 Time domain of a typical 3-subcarrier multisine and single sine CW. 36

2.2 Frequency domain of a typical 3-subcarrier multisine and single sine CW. The multisine signal appears as pulses with a period of $1/\Delta f$. . 36

2.3 Antenna equivalent circuit. . . . . 37

2.4 The block diagram of a generic WPT system. . . . . 41

2.5 Diagram of multiuser MIMO downlink system. . . . . 46

3.1 System model. . . . . 63

3.2 Alternating Optimization in one iteration. . . . . 68

3.3 $\mathbf{D}_{q,0}$ is the sum of red block diagonal, $\mathbf{D}_{q,1}$ is the sum of yellow block diagonal and $\mathbf{D}_{q,-1}$ is the sum of green block diagonal in the above matrix. 69

3.4 The definition of $\mathbf{B}_q$. . . . . 77

3.5 Simulation layout. . . . . 85

3.6 Average $i_{dc}$ versus $N$ for $L=1,20,40$. . . . . 86

3.7 Average $i_{dc}$ versus $L$ for $N=16, 32, 64$. . . . . 87

3.8 Average $i_{dc}$ versus $N$. . . . . 88

3.9 The average weighted sum of $i_{\mathrm{dc}}$ versus $N$ for $K=1$, 3, 5 and $L=20$. . 89

3.10 Average $i_{\mathrm{dc}}$ versus number of iterations for $N=4$, 16, 64, 128 and $L=30$. 90

3.11 Average $i_{\mathrm{dc}}$ versus number of iterations for $L=10$, 20, 30, 40 and $N=64$. 91

3.12 Average $i_{\mathrm{dc}}$ versus bandwidth for $L=20$ and $N=16$. . . . . 92

3.13 Average $i_{\mathrm{dc}}$ versus bandwidth for $L=30$ and $N=16$. . . . . 93

3.14 $i_{\mathrm{dc}}$ region versus $L$ for $N=32$. . . . . 94

3.15 $i_{\mathrm{dc}}$ region versus $D_{\mathrm{h}}$ for $N=32$ and $L=20$. . . . . 95

3.16 Average $i_{\mathrm{dc}}$ versus $N$ for $L=10$, 20, 30 and $K=1$. . . . . 96

3.17 Average weighted sum $i_{\mathrm{dc}}$ versus $N$ for $L=20$, number of transmit antennas $M_t=1$, 2, 4 and $K=2$. . . . . 97

3.18 Average $i_{\mathrm{dc}}$ versus $L$ for $N=32$ and resolution bits $M=1$, 2, 3. . . . 98

3.19 Average weighted sum $i_{\mathrm{dc}}$ versus $N$ for $M=1$, 2, 3, $L=20$ and $K=3$. 99

4.1 Markov decision process of Q-learning. . . . . 108

4.2 The downlink training and uplink feedback of proposed DRL framework. The detail structure of the distributed-learning-distributed-processing framework is shown in Fig. 4.4. . . . . 110

4.3 Timing of time slot $t-1$. . . . . 111

4.4 The framework of distributed-learning-distributed-processing scheme. 111

4.5 Average information rate versus the number of time slots with different codebook sizes $(S_{\mathrm{t}},S_{\mathrm{r}})$. . . . . 124

4.6 Average achievable information rate versus the number of time slots with different correlation coefficients. . . . . 125

4.7 Average information rate versus number of time slots with users rescheduling at 5e4th, 1e5th and 15e4th time slot. . . . . 126

4.8 Average information rate versus number of time slots with different number of users $K$. . . . . 127

4.9 Average information rate versus the number of time slots with different number of transmit antennas $M$ when $N=1,N_s=1,K=1$. . . . . 129

4.10 Average information rate versus number of time slots with different number of transmit antennas $M$ when $N=4,N_s=2,K=4$. . . . . 130

4.11 Average information rate versus the number of time slots with different number of streams $N$. . . . . 131

4.12 Average information rate versus number of time slots with different DRL-based schemes. . . . . 132

4.13 Sum reward versus the number of time slots with different number of users $K$. . . . . 133

4.14 Average information rate versus the number of time slots with different penalty $\lambda$. . . . . 134

4.15 cumulative distribution function (CDF) of the average information rate over different DRL-based methods and benchmarks. . . . . 135

4.16 Boxplot of the average information rate over different DRL-based methods and benchmarks. The bottom and top of each box are the 25th and 75th percentiles of the information rate values, respectively. The whisker length is set to infinity to ensure there are no outliers. . . . . 136

5.1 Proposed DDPG structure including the critic neural network and the actor neural network. . . . . 142

5.2 Set of beam directions with discrete phases when $S_t = 1,2,4,8$. . . . . 144

5.3 Set of beam directions with continuous phases. . . . . 145

5.4 Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N = 4, N_s = 1, K = 1$. . . . . 151

5.5 Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N = 4, N_s = 2, K = 1$. . . . . 152

5.6 Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N = 4, N_s = 1, K = 2$. . . . . 153

5.7 Boxplot of the average information rate over DDRL and DDPG with different numbers of transmit antennas $M$ when $N = 4, N_s = 1, K = 2$. The bottom and top of each box are the 25th and 75t percentiles of the information rate values, respectively. The whisker length is set to infinity to ensure there are no outliers. . . . . 154

5.8 Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N = 4, N_s = 1, K = 2$ with different reward function. . . . . 155

# List of Tables

4.1 Comparison Between Strategies . . . . . 116

4.2 Comparison Between Algorithms . . . . . 120

4.3 Running Time For Each Network Topology . . . . . 125

4.4 Total Execution Time For Each Channel Realization With Different $(K,M,N,N_s)$ In DDRL($\epsilon=0$) . . . . . 125

5.1 Comparison Between DDPG-based and DQN-based Algorithms . . . 147

# Chapter 1

# Introduction

Wireless networks have completely revolutionized our lives and our society. The currently deployed fourth generation (4G)/ fifth generation (5G) wireless networks have enabled billions of Internet of Things (IoT) devices such as smart phones, smart manufacturing, smart cities, wearables, healthcare, etc [1, 2]. It is noticed that these devices might be numerous, small-sized, deployed in unreachable environments or impossible for battery replacement which poses challenges on normal functioning, data throughput, spectual efficiency, energy supply, and connectivity. To address these challenges, next generation wireless networks, e.g., sixth generation (6G) should be multi-functional networks that not only provide communication-only services but also offer a wide range of new functionalities such as powering, intelligence, and computation [3].

### Multi-antenna Wireless Networks

To pave the way for multi-functionality, one of the most basic and powerful technologies is the multi-antenna technology where multiple antennas work cooperatively to transmit or receive radio waves. This technology, now standardized and integrated into current wireless systems, has become increasingly indispensable to fulfill the demands of future wireless networks [4]. Actually, multi-antenna technology utilizes

the spatial dimensions of the wireless propagation channel through multi-antenna signal processing techniques to increase the capacity of wireless networks. Such property is captured by so-called power gain in wireless powering networks and spatial multiplexing gain in wireless communication networks. To achieve these gains, a significant multi-antenna signal processing technique that can be exploited for both networks is spatial beamforming [5] which is defined as vectors that form a single narrow beam steering or tracking to one single target/user [6]. Although beamforming is very important in energizing low-power devices and low-interference communication, this technique is highly dependent on the ratio environment, especially channel state information (in the form of the frequency response) at the transmitter (CSIT), specifically in the downlink mode [7], which is not always available [4].

In this thesis, we make progress toward the understanding of multi-antenna beamforming techniques for different functionalities of wireless networks with perfect/imperfect CSIT. We will refine the beamforming optimization strategies of different functionalities or various forms of functionality integration, i.e, joint design of beamforming and waveform for powering networks to make better use of wireless resources, integration of communication and artificial intelligence to obtain robust beamforming and interference management, where robustness is intended with respect to the imperfect CSIT.

## 1.1 Motivation and Outline

### 1.1.1 Motivation

Inspired by the promising opportunities enabled by multi-antenna beamforming for multi-functional wireless networks, in this thesis, we push the boundary of this technology for different scenarios and analyze the impact of the beamforming on the performance limits. In particular, we aim to demonstrate that a good beamforming de-

sign not only enables larger active/passive beamforming gain, frequency diversity gain, and gain from the rectifier nonlinearity in WPN but also offers benefits over massive MIMO with imperfect CSIT in WCN. Specifically, The contribution is divided into three parts. The first part is devoted to the beamforming design for WPT using convex optimization. The second part is the beamforming optimization for massive MIMO with channel aging using DQN. The third part is the beamforming optimization for massive MIMO with channel aging using DDPG. This thesis is supposed to show that beamforming by convex optimization and reinforcement learning are very powerful for emerging multi-antenna wireless systems, e.g., improving the EPTE of WPN and the spectral efficiency of WCN to meet the high demand for power supply and explosive growth of data traffic in future wireless networks. In the next section, the detailed motivation is presented for emerging multi-antenna WPN and WCN systems, respectively.

For WPN, existing IRS-aided WPT works focus on single tone, frequency-flat IRS, frequency-flat channel and jointly optimize transmit beamforming and passive IRS beamforming phases [8–11]. The limitation of these designs is that they focused on the spatial domain beamforming-only and that the rectenna nonlinearity has been ignored. Ignoring the rectenna nonlinearity is equivalent to assuming that $e_3$ is constant and independent of the rectenna input signal power and shape [12,13]. Unfortunately, such an assumption is well documented to be invalid and to lead to inefficient designs [14]. Rectenna nonlinearity is indeed known to be a crucial feature to be modeled and accounted for in WPT system design as it significantly influences the WPT architecture designs [15,16]. Modeling the nonlinearity is particularly crucial in the low power regime (below 1mW), which is the primary deployment scenario of WPT in wireless networks. Also, the non-linear model of the rectifier is expected to reveal the benefits of multisine-based waveforms, otherwise inaccessible with the linear model. Hence, the nonlinearity opens the door to more efficient signal strategies, e.g. joint waveform and beamforming, for WPT that significantly enhance the performance over beamforming-only approaches.

In this thesis (Chapter 3), beyond the transmit beamforming-only approaches in wireless power networks, we particularly focus on an efficient IRS-aided WPT design by leveraging the progress made in the past few years on the modeling of the rectenna nonlinearity and corresponding efficient signal design for WPT on one hand and IRS on the other hand. Motivated by the significant benefits of a joint waveform and beamforming design to expand the range and the output DC power of WPT [16], [17], [14], this work pushes the limit of IRS-aided WPT by sufficiently exploiting the presence of IRS through a joint waveform and beamforming design for both single-user and multi-user deployments.

For WCN, accounting for reinforcement learning for massive MIMO, existing works [18–28] assume that perfect CSIT or instantaneous channel gain via receiver feedback is known at the transmitter. Unfortunately, such an assumption is impractical in both time division duplexing (TDD) and frequency division duplexing (FDD) with CSI feedback/acquisition delay and user mobility [29,30]. In TDD systems, the BS acquires CSI by uplink training and channel reciprocity. The uplink training and estimating approach is subject to antenna miscalibration and pilot contamination, leading to an imperfect CSIT. In FDD systems, the downlink channel is estimated by each user during the downlink training phase and then fed back to the BS via uplink signaling. This approach requires a high-rate low-latency feedback link between the user and the BS to keep the latter continuously and instantaneously informed about the CSI, which is not always available. In addition, beamforming is not limited to the transmitter and can also be used at the receiver to enhance interference management. Predicting the beamformers of both transmitter and receiver with channel aging is never considered in DRL-based works. Instead, all the existing work focuses on high-level multi-cell single-use (SU) single-input-single-output (SISO) [18, 20] (no transmit and receive beamforming) and multi-input-single-output (MISO) (only transmit beamforming) [19, 22, 26, 31] scenarios without considering multiple receive antenna cases.

In this thesis (Chapter 4, 5), beyond communication-only designs, both communication and reinforcement learning are combined in a joint signal-processing framework, such design leads to a trade-off between computation complexity and system performance and also calls for robust interference-management strategies. Driven by the appealing benefits of DRL, the main objective of this thesis on communication networks is to investigate the application DRL to the massive MIMO scenarios and enable technologies to combat interference caused by channel aging, namely DRL for massive MIMO with channel aging, which are envisioned to play key roles in next-generation wireless communication networks.

### 1.1.2 Outline

The remainder of this thesis is organized as follows:

In Chapter 2, we review the fundamentals and related works of this thesis by introducing the principles of waveform and rectenna design for WPT, channel model and CSIT types for WCN, and some mathematical tools.

In Chapter 3, we study the waveform and passive beamforming design for IRS-aided wireless power networks. Generalized multi-user and low-complexity single-user algorithms are demonstrated to maximize the weighted sum output DC. Performance evaluations confirm the significant benefits of a joint waveform and passive beamforming design accounting for the energy harvester nonlinearity to boost the performance of single-user and multi-user WPT systems.

In Chapter 4, we investigate the application of DRL for joint transmit and receive beamforming design in massive MIMO downlink transmission with channel aging. Leveraging this DRL-based framework, interference management is explored and three DRL-based schemes, namely the DDRL scheme, PDRL scheme, and CDRL scheme, are proposed and analyzed.

In Chapter 5, we extend the scope of research on the discrete beamformer selection scheme in Chapter 4 to continuous action spaces. Performance evaluation confirms that, compared with the result in Chapter 4, the performance can be further improved by beamformer generation strategies with continuous action space.

Finally, in Chapter 6, we conclude the thesis by summarising the achievements and discussing possible directions for future works.

## 1.2 List of Contributions

The novel contributions of this thesis are listed as follows:

### 1.2.1 For Wireless Power Networks

- A novel architecture based on the joint waveform and passive beamforming for IRS-aided WPT is presented. The proposed architecture can provide transmit beamforming gain, passive beamforming gain, frequency diversity gain, and gain from energy harvester nonlinearity in both single-user and multi-user deployments.

- Two different problem formulations of the joint waveform and beamforming design based on a frequency-selective IRS (FS-IRS) and a frequency-flat IRS (FF-IRS) are presented. High complexity multi-user and low complexity single user algorithms based on AO, SCA, and SDR are developed to solve the problems.

- The scaling law is derived with respect to number of passive beamforming elements, frequencies and transmit power which is cross-validated by the result evaluations.

- The DC current performance of the proposed IRS-aided wireless power transfer (WPT) is compared with conventional strategies in continuous/discrete beamforming phases, single/multi-transmit antennas, and single/multi-user scenarios.

### 1.2.2 For Wireless Communication Networks

- A novel multi-agent DRL-based framework for massive MIMO downlink transmission is studied. The proposed framework demonstrates robust beamforming and interference management in channel aging scenarios.

- A multi-user sum rate problem is formated and three DQN-based codebook-based beamforming schemes, namely distributed-learning-distributed-processing DRL-based scheme (DDRL), partial-distributed-learning-distributed-processing DRL-based scheme (PDRL), and central-learning-distributed-processing DRL-based scheme (CDRL), are proposed, and analyzed.

- The average rate performance of different schemes is compared with several conventional baseline strategies. The DDRL scheme is not sensitive to the dynamic and fast-changing wireless environment. Detailed convergence and scalability analysis are shown.

- A multi-agent DDPG-based beamforming design algorithm is proposed. Compared with DQN-based algorithms, DDPG-based algorithm exhibits better interference management.

- The pros and cons of DDPG-based beamforming strategy is compared with the approaches in Chapter 4 in the aspect of convergence, sum rate, variance of reward, stability, parameter tuning, computation complexity, and communication overhead.

## 1.3 Publications

The material presented in this thesis has led to the following publications:

1. Z. Feng, B. Clerckx and Y. Zhao, "Waveform and Beamforming Design for Intelligent Reflecting Surface Aided Wireless Power Transfer: Single-User and Multi-User Solutions," in IEEE Transactions on Wireless Communications, vol. 21, no. 7, pp. 5346-5361, July 2022, doi: 10.1109/TWC.2021.3139440.

2. Z. Feng and B. Clerckx, "Deep Reinforcement Learning for Multi-User Massive MIMO With Channel Aging," in IEEE Transactions on Machine Learning in Communications and Networking, vol. 1, pp. 360-375, 2023, doi: 10.1109/TMLCN.2023.3325299.

3. Z. Feng and B. Clerckx, "Deep Reinforcement Learning for Multi-User Massive MIMO: A DDPG Approach," (working title) *to be submitted*

## 1.4 Notation

The following notation is used throughout the thesis. Vectors and matrices are denoted by bold and lower letters and bold and upper letters, respectively. $(\cdot)^H$,$(\cdot)^*$ and $(\cdot)^T$ denote conjugate transpose, complex conjugate and transpose. $\|\cdot\|_F$ and $\|\cdot\|$ denote Frobenius norm and $L_2$-norm. We use $|\cdot|$ for absolute value, $\mathrm{Tr}\{\cdot\}$ for trace and $\Re\{\cdot\}$ and $\Im\{\cdot\}$ for real part operators and imaginary part operators. The notation $\succeq$ denotes positive-semidefinite and $\mathrm{diag}\{\mathbf{x}\}$ denotes diagonalization operation with elements of $\mathbf{x}$ in the diagonal. Imaginary unit is denoted as $j=\sqrt{(-1)}$. $\mathcal{E}\{\cdot\}$ represents statistical expectation. $\mathbf{I}_M$ denotes an $M \times M$ identity matrix. $\mathbf{0}$ denotes an all-zero matrix. $||\mathbf{a}||$ denotes the norm of a vector $\mathbf{a}$. $|a|$ denotes the norm of a variable $a$.

# Chapter 2

# Fundamentals and Literature Review

This chapter introduces some fundamental knowledge and a literature review of multi-antenna wireless networks. First, we revisit the waveform design and rectenna design for WPT systems. Second, we introduce the multi-user MIMO system. Finally, we briefly discuss some mathematical tools for a better understanding of later chapters.

## 2.1 Some Fundamentals of WPT

WPT uses RF signals, transmitted at a certain frequency and with a certain amount of average power to power up devices up to large distances. The amount of transmitted power is known and usually constant, which allows performing optimized design of some key components, e.g., waveform and rectenna, to maximize the RF-to-DC conversion efficiency.

### 2.1.1 WPT Waveform Design

For WPT, an important way to improve the power transfer is to design a special signal waveform at the transmitter. The conventional waveform design method for WPT comes from the EPCglobal, class 1, generation 2 ("gen2") standard, which specifies a continuous caveform (CW) power transmission having a constant envelope with a single-tone carrier. This method is not efficient and operates with efficiencies of less than 60% [32]. This efficiency is improved by using power-optimized waveforms (POW) with multiple-tone carriers since such waveforms have high peak-to-average-ratio (PAPR) which increases RF-to-DC conversion efficiency. The PAPR(dB) of a waveform can be defined as

$$\text{PAPR} = 10\log_{10}\left(\frac{\max[x(t)^2]}{<x(t)^2>}\right) \tag{2.1}$$

where $x(t)$ is the time domain waveform of the signal of interest and $<>$ refers to the time average operator.

The first development of POW can be found in the works of M.S. Trotter and G.D. Durgin [33], wherein the authors propose the use of multiple-tone carriers instead of a single-tone carrier to provide a higher charge pump efficiency and thus increase the range of Radio Frequency Identification (RFID) readers. As can be seen from Fig 2.1 and 2.2, a POW is characterized by a high PAPR. The envelope of the transmitted POW signal is designed so that there are large peaks, while the average power is kept the same as in the single-carrier CW case. The time-domain baseband representation of the $N$-POW is given by a simple summation of $N$ cosines, which is then up-converted to the carrier frequency:

$$x_{POW}(t) = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\cos(2\pi n\Delta f t) \tag{2.2}$$

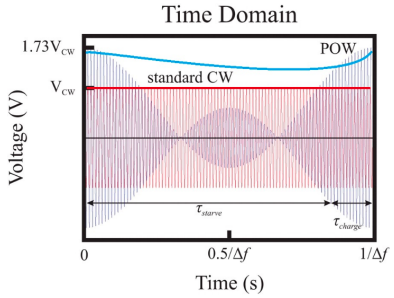



Figure 2.1: Time domain of a typical 3-subcarrier multisine and single sine CW.

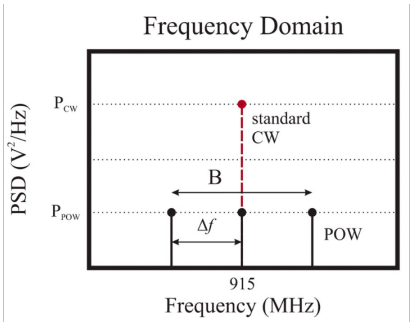



Figure 2.2: Frequency domain of a typical 3-subcarrier multisine and single sine CW. The multisine signal appears as pulses with a period of $1/\Delta f$.

The normalization factor $\sqrt{1/N}$ is used to ensure that the time-averaged power of the POW remains equal to that of a CW carrier of unity amplitude. Thus, each subcarrier has an average power of $1/2N$, and the average conducted power of the transmitted signal is the sum of the powers of all subcarriers and equals $1/2$. The root mean square (RMS) voltage of the multisine signal is also the same as in the case of a unity voltage amplitude single carrier CW signal ($1/\sqrt{2}$). Peak voltage of the $N$-POW is equal to $\sqrt{N}$ and peak power is given by $N$, and it occurs when each subcarrier is at its maximum value, provided all the subcarriers are in phase [32]. The benefit of using multisine is that high-PAPR signals can increase the RF-to-DC conversion efficiency which has been simulated and experimentally proved in [34].

In (2.2), the waveform weights are constant and are not adaptive to the CSI and multipath fading. Hence, in Chapter 3, we design multi-antenna multisine WPT waveform for transmission over a multipath channel and formulate an optimization

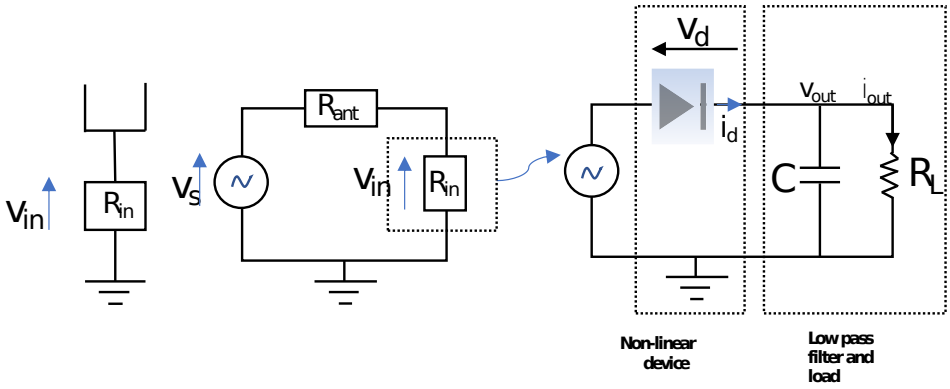



Figure 2.3: Antenna equivalent circuit.

problem to adaptively change the waveform weights as a function of the CSI so as to maximize the rectenna output DC current.

### 2.1.2 Rectenna

The design of RF EH is also very important in a WPT system. The basic RF EH is a rectenna – a device consisting of an antenna, a matching network and a rectifying circuit [35]. The antenna collects the energy emitted from an RF source, whereas the rectifying element, most often the diode, rectifies the AC voltage induced at the antenna terminals by the electromagnetic wave and an output low pass filter converts the rectified signal into the DC voltage.

In this thesis, to have a more tractable model of the nonlinear process of the EH, we consider a simple model of the rectifier which is shown in Fig. 2.3.

Assume that the impinging received signal $y(t)$ has a average power $P_{\mathrm{r}} = \mathcal{E}\{|y(t)|^2\}$. The receive antenna is assumed to be a lossless antenna which can be modeled as a voltage source $v_{\mathrm{s}}(t)$. Then, the rectenna architecture can be seen as a voltage source $v_{\mathrm{s}}(t)$ connected with a rectifier input impedance $Z_{\mathrm{in}} = R_{\mathrm{in}} + jX_{\mathrm{in}}$ and an antenna impedance $Z_{\mathrm{ant}} = R_{\mathrm{ant}} + jX_{\mathrm{ant}}$. Under the assumption of the lossless antenna and perfect matching

$(R_{\text{ant}}=R_{\text{in}},X_{\text{ant}}=-X_{\text{in}})$, the receive power $P_{\text{r}}$ is completely transferred to rectenna impedance which builds the relation between $P_{\text{r}}$ and $v_{\text{in}}(t)$ such that

$$P_{\text{r}}=\mathcal{E}\{|y(t)|^2\}=\mathcal{E}\{|y_q(t)|^2\}\frac{R_{\text{ant}}}{R_{\text{in}}}=\frac{\mathcal{E}\{|v_{\text{in}}(t)|^2\}}{R_{\text{in}}} \tag{2.3}$$

Under the assumption of perfect matching, we have $v_{\text{s}}(t)=2v_{\text{in}}(t)$. Hence, $v_{\text{in}}(t)$ can be associated with the received signal $y(t)$ with $v_{\text{in}}(t)=y_q(t)\sqrt{R_{\text{ant}}}$ [16]. To better understand the rectifier DC output current $i_{\text{out}}$, we focus on the diode I-V characteristic

$$i_{\text{d}}(t)=i_{\text{s}}(e^{\frac{v_{\text{in}}(t)-v_{\text{out}}(t)}{n'v_{\text{t}}}}-1), \tag{2.4}$$

where $i_{\text{s}}$,$n'$ and $v_{\text{t}}$ are the reverse bias saturation current, diode ideal factor, and thermal voltage, respectively. The diode nonlinearity is rooted in the term $e^{\frac{v_{\text{d}}(t)}{n'v_{\text{t}}}}$. [16] [36] illustrates the significance of the higher order ($\geq 4$) truncation of the Taylor series for the term $e^{\frac{v_{\text{d}}(t)}{n'v_{\text{t}}}}$ through

$$i_{\text{d}}(t)\approx\sum_{i=0,\text{even}}^{n_0}\beta_i R_{\text{ant}}^{i/2}y_q(t)^i,\beta_i\triangleq i_{\text{s}}/i!(n'v_{\text{t}})^i. \tag{2.5}$$

After the low pass filter (LPF), the approximated output DC $i_{\text{out}}$ is given by

$$i_{\text{out}}=\mathcal{E}\{i_{\text{d}}(t)\}\approx\sum_{i=0,\text{even}}^{n_0}\beta_i R_{\text{ant}}^{i/2}\mathcal{E}\{y(t)^i\}. \tag{2.6}$$

**Linear Model**

Given $n_0=2$, we have

$$i_{\text{out}}=\beta_2 R_{\text{ant}}\mathcal{E}\{y(t)^2\}. \tag{2.7}$$

The name "linear model" stems from the fact that $i_{\text{out}}$ is directly proportional to the average incident power $\mathcal{E}\{y(t)^i\}$. This second-order truncation model is equivalent to the linear model.

Note that (2.7) is derived by truncating the Taylor series of the diode equation to the second-order term and assuming that the converted power is linearly proportional to the diode current $i_{\text{out}}$. It is shown that the linear model favors narrowband signal transmission, where the transmit power is allocated to a single frequency component [37].

### Nonlinear Model

In multi-sine WPT, the linear rectenna model-based design allocates transmit power to a single frequency component. Unfortunately, this contradicts experimental results that show the benefits of a power allocation over multiple frequency components to boost the RF-to-DC conversion efficiency [38–40]. As it is revealed in the RF literature [40] that the 4th-order truncation is necessary for characterizing the basic nonlinear rectification process, [37] sets $n_0 = 4$ to achieve a fourth-order truncation nonlinear model, i.e.

$$i_{\text{out}} = \beta_2 R_{\text{ant}} \mathcal{E}\{y(t)^2\} + \beta_4 R_{\text{ant}}^2 \mathcal{E}\{y(t)^4\}. \tag{2.8}$$

It is worth noting that if the amplitude of $y(t)$ is quite large such that the diode behaviour is dominated by the diode series resistance, the diode I-V characteristic would be linear [38], and the nonlinear model is inapplicable. Circuit simulations in [16] validate the nonlinear model in (2.8), highlight the inaccuracy of the linear model and confirm the superiority of optimized waveforms over various baselines. The nonlinear model leads to a completely different spectrum usage compared to the linear model [37, 38].

To get a tractable analytical model, one can truncate to $n_0 = 4$ while retaining the main source of nonlinearity as part of the fourth-order term. Higher order analysis can be found in [16] [36]. In this thesis, we consider $n_0 = 4$ in rectenna design in Chapter 3.

## 2.2 Literature Review for WPT

### 2.2.1 Multi-antenna Wireless Power Networks

As one of the most promising functionalities of wireless networks, powering, or say WPT has drawn great attention because of its potential to energize billions of future low-power devices. It is deemed as the enabler of the 1G of mobile power networks [12], and is divided into near-field Inductive Power Transfer and far-field Radiative Power Transfer based on serving distance [41].

#### Near-field VS Far-field WPT

Near-field Inductive Power Transfer needs to precisely align coils at energy transmitter (ET) and energy receiver (ER). These interconnected coils function as an air gap transformer. When alternating current flows through the primary coil (at the ET), it generates a changing magnetic field. This field, in turn, induces a varying voltage in the secondary coil (at the ER). However, as the distance increases, the magnetic field strength diminishes rapidly [42].

Far-field Radiative Power Transfer is a promising WPT approach. In radiative WPT, ER gather wireless power from electromagnetic waves. As the distance between an ET and an ER increases, the inductive and capacitive effects caused by the antennas at both ends diminish. The advantage of this vanishing effect is that the ET and ER become electrically separated for distances over a few to hundreds of meters, allowing for independent design of their circuits [38]. In this thesis,we consider far-field radiative power transfer in Chapter 3, and the terminology "WPT" refers to far-field radiative WPT unless otherwise stated.

#### Optimization Strategies of WPT




Figure 2.4: The block diagram of a generic WPT system.

In order to demonstrate the advantage of far-field radiative WPT, recall first the block diagram of a typical WPT system illustrated in (2.4). The end-to-end power transfer efficiency (EPTE) $e$ can be decomposed as

$$e = \underbrace{\frac{P_{\text{RF,Tx}}}{P_{\text{DC,Tx}}}}_{e_1} \cdot \underbrace{\frac{P_{\text{RF,Rx}}}{P_{\text{RF,Tx}}}}_{e_2} \cdot \underbrace{\frac{P_{\text{DC,Rx}}}{P_{\text{RF,Rx}}}}_{e_3} \tag{2.9}$$

where $e_1$, $e_2$, and $e_3$ denote the DC-to-RF, RF-to-RF, and RF-to-DC power conversion efficiency (CE), respectively. An ET can have the same structure as a communication transmitter, sending electromagnetic waves, i.e. ratio frequency (RF) signals. An ER harvests energy from RF signals by exploiting a rectenna, which consists of an antenna and a rectifier [38]. RF power input into a rectenna can be rectified and converted to DC power. In the past few years, four different strategies have been proposed and optimized to boost the EPTE.

A *first* strategy is the design of energy *waveforms*[1] to utilize the nonlinearity characteristic of the rectenna and boost jointly the RF-to-RF CE $e_2$ and the RF-to-DC CE $e_3$ [16]. Such design originates from the fact that the output of the ER is a nonlinear function of the rectenna input signal. Hence, the transmit waveform design has a significant influence on the EPTE. It not only affects the $e_2$ and the signal at the input of the rectenna, but also the $e_3$ of the rectifier. A stepping stone in such design was made in [16] where a

[1]A more fundamental illustration for waveform is shown in Chapter 2, here we focus on literature review only.

systematic methodology was derived to design and optimize waveforms for WPT. Different from prior non-optimized waveform designs [43], the optimal waveform design in [16] was adapted to the frequency selective channel (with frequency flat channel being a special case) and was rooted in the tradeoff between maximizing $e_2$ and $e_3$. Such optimized waveform was shown to provide significant benefits over conventional continuous-wave signal and non-optimized waveforms by leveraging the channel frequency diversity gain and a gain originating from the rectifier nonlinearity. Since the optimal signal design is the solution of computationally involved problems, low complexity methods have been proposed in [17,44], and have been experimentally validated in [14]. Furthermore, since the CSIT is needed to design the optimal waveform, waveform designs based on limited CSI feedback have been proposed in [45], and experimentally demonstrated in [46].

A *second* strategy is the design of multi-antenna *beamforming* to boost the input power of ER and enhance the RF-to-RF CE $e_2$. Due to the hardware and other limitation of ER, the acquisition of CSIT is challenging. An efficient method was proposed in [47] to overcome this challenge by designing beamforming with limited feedback. Other directive or energy focusing solutions, such as time-reversal techniques [48] and time-modulated arrays [49], have also been used in WPT system to exploit the real-time antenna beamforming strategies. Beamforming is not limited to the transmitter and can also be used at the receiver subject to a proper design of the combiner scheme [50]. Waveform and beamforming can also be combined into a joint waveform and beamforming design [16, 17, 51] to further boost the output DC power and the range of WPT, as demonstrated experimentally in [52]. Joint waveform and beamforming enables to simultaneously harvest three different gains, namely a beamforming gain, a frequency diversity gain and a gain related to the rectifier nonlinearity, and therefore offers additional opportunities over spatial domain processing/beamforming-only.

A *third* strategy is to design *modulation* for WPT. Different from [16] where energy waveforms consist of deterministic unmodulated multi-carrier signals, modulation

induces random fluctuation of the input rectifier signal to increase $e_3$. For example, real Gaussian modulation outperforms circularly symmetric complex Gaussian (CSCG) modulation despite the same average input power to the rectifier [53]. Similarly, [54] has observed that an on-off keying based modulation strategy outperforms real Gaussian modulation in terms of output DC power.

A *fourth* strategy is to apply phase sweeping *transmit diversity* in multi-antenna settings [55]. Compared to the beamforming strategy, transmit diversity does not rely on CSIT but still increases $e_3$ and therefore output DC power.

### Intelligent Reflecting Surfaces aided WPT

In parallel to the active research area on signal design strategies for WPT, a new paradigm named intelligent reflecting surfaces (IRS) has recently attracted the attention of the wireless research community to further boost the efficiency of wireless communications [56–60]. Compared with the active beamforming of (massive) multi-antenna systems, IRS benefits from the presence of passive beamforming reflection elements (REs) integrated into the propagation environment, which is optimized to adapt the wireless channel so as to be favorable to communication receivers.

The key advantages of IRS are the tunability of the phase shifters at each RE (so-called passive beamforming phases), the flexibility of the deployment on arbitrarily shaped surfaces and the sustainability due to low cost and power consumption. By collaboratively adjusting the passive beamforming phases, the incident signal is reflected at each RE such that the reflected signals are constructively accumulated at the target receiver to increase the received signal power, therefore enabling a passive beamforming gain. Nevertheless, in contrast to active antenna arrays where the amplitudes and phases can be adjusted freely at each antenna and at each frequency, REs may be subject to less flexibility due to the passive nature of the IRS and the hardware constraints.

Moreover, the IRS is commonly assumed to be consistent with frequency-flat in the sense that passive beamforming phases are kept constant across frequencies, (so-called frequency-flat IRS in the sequel). Under those constraints, IRS-assisted wireless communications have focused on a single carrier, frequency-flat channel and frequency-flat IRS by jointly designing the active transmit beamforming at the transmitter and passive reflection beamforming at the IRS for SU [61] and multi-user (MU) scenarios [58, 62, 63]. Multi-carrier orthogonal frequency division multiplexing (OFDM) waveform designs were also studied in some papers using frequency-flat IRS [64, 65]. Aside from communications, IRS brings some natural benefits to wireless power since IRS can help to increase the power level at the input of the rectenna which focuses on single tone, frequency-flat IRS, frequency-flat channel and jointly optimizes transmit beamforming and passive IRS beamforming phases [8–11].

## 2.3 Some Fundamentals of Multiuser MIMO

MIMO technology stands as an essential part of 3G/4G/5G wireless communication standards. In contrast to SISO, MIMO facilitates spatial multiplexing and diversity gain through the utilization of multiple antennas at both the transmitting and receiving ends. This setup enables base stations (BS) to simultaneously serve numerous users within the same time-frequency resource, enhancing both uplink and downlink transmissions. However, achieving optimal multiuser transmission, particularly in the downlink, necessitates precise CSIT to mitigate multiuser interference, given the mobile users across the service area. Yet, obtaining perfect CSIT poses significant challenges in practical systems and has become progressively difficult due to the increasing number of antennas and access points in 5G.

So far, many techniques have been developed under the assumption of perfect CSIT, then adapted for real-world scenarios with imperfect CSIT. Consequently, system

performance suffers varying degrees of degradation due to CSIT imperfections. There exists an increasing need to design wireless networks that can accommodate imperfect CSIT and reduce multiuser interference.

### 2.3.1 Multi-user MIMO Systems

Consider multiuser MIMO downlink systems, which is illustrated in Fig 2.5. A Base Station (BS) equipped with multiple antennas serves a group of multi-antenna users. These users can be served by Time-Division-Multiple Access (TDMA) or Frequency-Division-Multiple-Access (FDMA) techniques with time or frequency sharing among users. However, this transmission strategy (e.g., TDMA/FDMA) is limited by low spectral efficiency.

### 2.3.2 Broadcast Channel

There exist a number of technologies that the BS (i.e., transmitter) uses to simultaneously serve those users within a time/frequency slot. To name just a few, for instance, Broadcast Channel (BC) refers to that the BS simultaneously sends dedicated (i.e., private) messages to each user [66]. A BC models the downlink of a wireless system consisting of one transmitter (or information source) and multiple receivers. The terminology reflects the fact that communication takes place over a shared broadcast medium. A BC where the transmitter and (possibly) the receivers are equipped with multiple antennas is known as a MIMO-BC. In this thesis, we mainly focus on broadcast channels in Chapter 3, 4, 5.

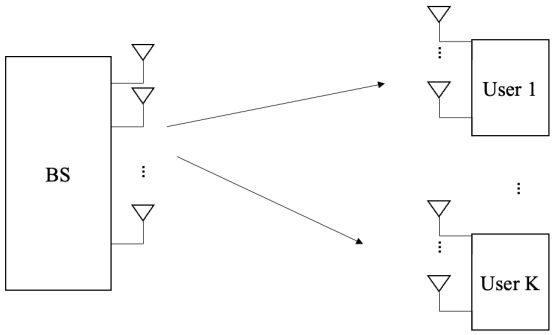



Figure 2.5: Diagram of multiuser MIMO downlink system.

### 2.3.3 Signal Model

Consider the MIMO-BC equipped with one $M$-antenna BS and $K$ $N$-antenna users indexed by $\mathcal{K}=\{1,...,K\}$ [67] . The BS aims to deliver $M_s$ streams in the time instant of interest. A number of $KN_s$ streams are transmitted simultaneously from the $M$ antennas of the BS. Each group of $N_s$ streams indexed by $N_s \in \mathcal{N} = \{1,...,N_s\}$ is targeted at one of the $K$ users. The transmitted signal for user $k$, i.e., the precoded data vector, at time slot $t$ can be written as

$$\mathbf{x}_k(t)=\sum_{n=1}^{N_s}\mathbf{p}_{k,n}(t)s_{k,n}(t) \tag{2.10}$$

where $s_{k,n},\forall k\in\mathcal{K},\forall n\in\mathcal{N}$, is the encoded message from message $W_{k,n}$ with zero mean and $\mathcal{E}(|s_{k,n}|^2)=1$, and precoder $\mathbf{p}_{k,n}(t)\in\mathbb{C}^{M\times 1}$ is subject to $\|\mathbf{p}_{k,n}(t)\|^2=1$. The received signal at user $k$ can be expressed as

$$\mathbf{y}_k(t)=\mathbf{H}_k(t)\sum_{n=1}^{N_s}P_{k,n}\mathbf{p}_{k,n}(t)s_{k,n}(t)+\mathbf{H}_k(t)\sum_{j\neq k,j=1}^{K}\sum_{i=1}^{N_s}P_{j,i}\mathbf{p}_{j,i}(t)s_{j,i}(t)+\mathbf{n}_k(t) \tag{2.11}$$

where $\mathbf{H}_k(t)$ is the channel matrix between the transmitter and the $k$th receiver,the noise vector $\mathbf{n}_k\in\mathbb{C}^{N\times 1}$ is assumed to follow a complex normal distribution, i.e., $\mathbf{n}_k\in\mathcal{CN}(\mathbf{0},\sigma_n^2\mathbf{I}_N)$. At the user side, the combiner vector for each stream is denoted

as $\mathbf{w}_{k,n}(t) \in \mathbb{C}^{N\times 1}$, $\|\mathbf{w}_{k,n}(t)\|^2 = 1, \forall k \in \mathcal{K}, n \in \mathcal{N}$.

### 2.3.4 Channel Model

Channel model is a simple representation of the communication channel between the BS and the user. The channels vary randomly, or fade, according to a statistical distribution [4]. Fast fading refers to that the channel impulse response changes rapidly within the symbol duration due to reflections of local objects and the motion of the objects relative to those objects. By contrast, slow fading refers to that the channel impulse response remains within the symbol duration. In this thesis, we focus on fast fading channels and briefly introduce independently and identically distributed (i.i.d.) Rayleigh channel and finite scatterer (or Ray-based geometric) channel.

#### Rayleigh Fading Channel

The i.i.d. Rayleigh fading channel indicates that each entry of the channel vector e.g. $\mathbf{H}$ is i.i.d. zero-mean unit-variance complex Gaussian distributed, i.e., $h \in \mathcal{CN}(0,1)$. It is also known as uncorrelated Rayleigh channel and occurred when the antenna spacings and/or the angular spreading of the energy at both sides of the wireless link are sufficiently large such that the various channel correlation becomes very small and negligible [4]. The detail application of Rayleigh fading channel is discussed in Chapter 3.

#### Finite Scattering Channel

The finite scattering channel model has been widely used in Massive MIMO systems [68]. In this model, the channel vector from a given user is defined as the multiplication of path gain, and the angle-of-departure (AoD). The path gain has i.i.d. $h \in \mathcal{CN}(0,1)$ entries and varies independently across different time slots. AoD contains steering

vectors and is closely related to the antenna array structure, operating frequency and antenna spacing, etc. The mathematical definition of the finite scattering channel is demonstrated in Chapter 4.

### 2.3.5 CSIT Acquisition

So far, we have introduced a number of multiuser MIMO system where the BS serves a group of users by precoding techniques in order to achieve certain performance. The performance of these systems and corresponding precoding techniques heavily depends on the acquisition of CSIT which can be divided into frequency division duplexing (FDD)[2] systems and time division duplexing (TDD) systems.

In FDD systems, the downlink channel is estimated by each user during the downlink training phase and then fed back to the BS via uplink signalling. In TDD systems, the BS acquires CSI by uplink training and channel reciprocity. Perfect CSIT requires accurate channel estimation and ideal channel reciprocity.

However, obtaining high quality instantaneous CSIT represents a challenge, especially in fast fading channels. In FDD systems, the feedback link suffers from error, delay, etc. In TDD systems, the BS obtains CSI by channel reciprocity. Still, only a partial current CSIT is attainable due to antenna miscalibration and pilot contamination. An intermediate solution should be adaptive to different CSIT types.

### 2.3.6 CSIT Types

To realize multiuser transmission performance in the downlink, CSIT is required to deal with multiuser interference as the receivers (i.e., mobile users) are usually scattered over the service area and cannot cooperate. Depending on the acquisition of

[2]While some results developed in this work can be applied to various imperfect CSIT scenarios either in TDD or in FDD mode, we mainly focus on FDD system for clarity.

CSIT, CSIT can be divided into perfect CSIT, statistical CSIT [69], delayed CSIT [70] and partial CSIT [71].

#### Perfect CSIT

For perfect CSIT, in FDD systems, the downlink channel is estimated by each user during the downlink training phase and then fed back to the BS via uplink signaling. This approach requires a high-rate low-latency feedback link between the user and the BS to keep the latter continuously and instantaneously informed about the CSI. In TDD systems, the BS acquires CSI by uplink training and channel reciprocity. Perfect CSIT requires accurate channel estimation and ideal channel reciprocity. In Chapter 3, we focus on perfect CSIT.

#### Delayed CSIT

However, in FDD systems, high-quality CSI can be obtained on the user side whereas the BS may acquire this information with a delay such that the reported CSI known at the BS is different with the current CSI. Delayed CSIT occurs when the coherence time of the channel becomes short due to high mobility for example. This phenomenon, which is also known as channel aging, describes the mismatch between estimated and update-to-date channels. That is to say, it represents the divergence arising between the channel estimation happening in the BS/users and the actual channel through which the data transmission occurs. In this thesis, we focus on delayed CSIT in Chapter 4, and 5.

### 2.3.7 Performance Metric

To evaluate the system performance, a list of throughput metrics are of considerable interest, including but not limited to sum DC power, weighted sum DC power, sum

rate, weighted sum rate and minimum rate. Specifically, DC power is defined as the output signal of the diode that is passed through a low-pass filter, which removes high-frequency harmonic components and outputs the DC power. Weighted DC power indicates the sum of a weighted individual power of each user. Sum rate refers to the sum of individual data rate that each user achieves, which measures the system performance as a whole. Weighted sum rate indicates the sum of a weighted individual data rate of each user, which maintains various user priorities or fairness. Minimum rate refers to the minimum individual rate out of all users which measures the worst data rate that users can experience. In this thesis, we mainly focus on the weighted sum power metric in Chapter 3, and the sum rate in Chapter 4, and 5.

### 2.3.8 Precoding Techniques

Let us consider multiuser MIMO broadcast channels and momentarily assume perfect channel state information at the transmitter (CSIT). Precoding is a processing technique that exploits CSIT to deal with multiple streams transmission.

From the perspective of design, the precoding techniques can be categorized into nonlinear and linear methods. The non-linear precoding technique, Dirty Paper Coding (DPC), achieves the sum-rate capacity of broadcast channel [72].

Typical linear precoding techniques include matched precoding, and zero-forcing (ZF) [4]. Specifically, matched precoding designs the precoder along with the channel direction, maximizing the desired signal power. ZF projects each users signal in the nullspace of all other users' channels to minimise the multiuser interference. To perform interference suppression during data transmission, the Zero Forcing Channel Inversion (ZF-CI) strategy is usually used.

**Zero Forcing Channel Inversion**

The purpose of ZF-CI is to identify the precoding matrix $\mathbf{P}_k = [\mathbf{p}_{k,1}, \mathbf{p}_{k,2}, ..., \mathbf{p}_{k,N_s}] \in \mathbb{C}^{M \times N_s}$ and combiner matrix $\mathbf{W}_k = [\mathbf{w}_{k,1}, \mathbf{w}_{k,2}, ..., \mathbf{w}_{k,N_s}] \in \mathbb{C}^{N_s \times N}$ for user $k$. To suppress the multi-user interference (MUI), we introduce the following constraint

$$\mathbf{H}_j \mathbf{P}_k = 0, \forall j \neq k, \tag{2.12}$$

in the sense that $\mathbf{P}_k$ should be in the null space of $\hat{\mathbf{H}}_k \in \mathbb{C}^{N(K-1) \times M}$ which is defined as

$$\hat{\mathbf{H}}_k = [\mathbf{H}_1, ..., \mathbf{H}_{k-1}, ..., \mathbf{H}_{k+1}, ..., \mathbf{H}_K]^H. \tag{2.13}$$

We denote the rank of $\hat{\mathbf{H}}_k$ as $\hat{L}_k$, the single value decomposition (SVD) of $\hat{\mathbf{H}}_k$ is given below

$$\hat{\mathbf{H}}_k = \hat{\mathbf{U}}_k \hat{\mathbf{\Lambda}}_k [\hat{\mathbf{V}}_k^{(a)}, \hat{\mathbf{V}}_k^{(b)}]^H \tag{2.14}$$

where $\hat{\mathbf{U}}_k \in \mathbb{C}^{N(K-1) \times N(K-1)}$ and $\hat{\mathbf{\Lambda}}_k \in \mathbb{C}^{N(K-1) \times M}$ contains left singular vectors and ordered singular values of $\hat{\mathbf{U}}_k$, respectively. $\hat{\mathbf{V}}_k^{(a)} \in \mathbb{C}^{M \times \hat{L}_k}$ contains first $\hat{L}_k$ right singular vectors and $\hat{\mathbf{V}}_k^{(b)} \in \mathbb{C}^{M \times (M - \hat{L}_k)}$. Since $\hat{\mathbf{V}}_k^{(b)}$ forms an orthogonal basis for the null space of $\hat{\mathbf{H}}_k$, user $k$ has a non-interference block channel $\hat{\mathbf{H}}_k \hat{\mathbf{V}}_k^{(b)}$, and it's columns are candidates for the precoder matrix $\mathbf{P}_k$.

The next step is to decouple the non-interfering block channel $\hat{\mathbf{H}}_k \hat{\mathbf{V}}_k^{(b)}$ into $N$ parallel sub-channels, we compute the SVD of $\hat{\mathbf{H}}_k \hat{\mathbf{V}}_k^{(b)}$ as

$$\hat{\mathbf{H}}_k \hat{\mathbf{V}}_k^b = \hat{\mathbf{U}}_k^{(c)} \hat{\mathbf{\Lambda}}_k^{(c)} \hat{\mathbf{V}}_k^{(c)H}. \tag{2.15}$$

Then, we denote the precoder matrix $\mathbf{P}_k = \hat{\mathbf{V}}_k^{(b)} \hat{\mathbf{V}}_k^{(c)}$ and combiner matrix as $\mathbf{W}_k = \hat{\mathbf{U}}_k^{(c)H}$. In this thesis, ZF-CI serves as a performance upperbound in Chapter 4, and 5.

Generally speaking, non-linear precoding outperforms linear precoding in terms of the

rate performance. Nevertheless, linear precoding methods are of more interest due to their implementation simplicity in practical systems. In this thesis, we mainly focus on linear precoding methods. We also develop novel precoding strategies to enable good performance in Chapter 3, 4, and 5.

From the perspective of implementation, the precoding techniques can be categorised into digital precoding, analog precoding and hybrid precoding methods. Digital precoding is realised when the BS has a dedicated RF chain for each antenna element. The digital precoding vector (or matrix) can be arbitrarily constructed in the complex domain. When the BS has multiple transmit antenna but only one RF chain, analog precoding as in [73] can be implemented by analog phase shifters. However, the analog phase shifter has only control in its phase and is subject to a constant modulus constraint, leading to reduced precoding flexibility. When the BS has less RF chains than the number of transmit antennas, a hybrid precoding is implemented by a high-dimensional analog precoder using cost-efficient analog phase shifters network, cascaded with a reduced-dimensional digital precoder [74]. In this thesis, we mainly consider digital precoding in Chapter 3, 4, and 5.

## 2.4 Literature Review of Multi-antenna Wireless Communication Networks

One of the most developed and powerful functionalities of wireless networks is WCN. The use of multiple antennas can tremendously increase the performance of wireless communication networks and it has been one of the driving factor behind the success of the 4G/5G.

### From MIMO to massive MIMO

Multiple-Input-Multiple-Output (MIMO) technology has become an essential component in 4G/5G wireless communication standards. Compared to Single-Input-Single-Output (SISO), MIMO provides spatial multiplexing gain by employing multiple antennas at the transmitter and receiver. This gain can be roughly defined as the number of streams that can be simultaneously multiplexed over the channel. With a massive number of antennas at the BS, MIMO system is gradually evolved into a massive MIMO system. [75–80].

To enable a high throughput in the massive MIMO system, the base station (BS) relies on the huge demand for global and instantaneous channel state information (CSI) based on efficient channel estimation techniques [81,82]. Nevertheless, the ground/air/space platforms such as high-speed trains/unmanned aerial vehicles (UAV)/satellites have a common characteristic of 3D mobility which leads to a stringent time constraint on CSI acquisition and even causes misalignment of narrow beams. Therefore, in future WCN, how to maintain good connectivity and system capacity without perfect CSIT (so-called imperfect CSIT) is regarded as an important problem that yearns for prompt solutions.

### From CSIT Imperfection to Channel Aging

As already mentioned earlier, acquiring an accurate and instantaneous CSIT is very challenging in practical systems. For instance, in the current wireless networks, the users estimate their CSI using pilot symbols, and the estimated CSI are then quantized and reported to the transmitter over a standardized number of bits. It is clear that this leads to a knowledge of the channel by the transmitter which is not perfect. The imperfect CSIT is usually caused by the drastic change of the propagation environment due to user mobility [29] and CSI feedback/acquisition delay between the BS and users [30] for downlink transmission. The CSI delay due to user mobility or feedback or acquisition delay is the time gap between the time point when the downlink training happens and

the BS starts downlink data transmission with the estimated channel. Such delay can be on the order of milliseconds which can cause the estimated channels to be outdated when actually downlink transmission happens. This delay becomes more catastrophic at high user mobility since rapid channel variation inevitably causes performance degradation in massive MIMO systems [83]. This phenomenon, which is known as *channel aging*, describes the mismatch between estimated and update-to-date CSIT. That is to say, it represents the divergence arising between the channel estimation happening in the BS/users and the actual channel through which the data transmission occurs.

**Massive MIMO with Channel Aging**

Since CSIT is pivotal to the performance of systems that account for a very large number of antennas and users, a great research effort has been made to massive MIMO systems with channel aging [29, 30, 83–89]. In [29], the impact of channel aging due to mobility is partially overcome through finite impulse response Wiener predictor without considering hardware phase noise, which is further studied in [83]. To tackle the CSI feedback/acquisition delay, one strategy is to use space-time interference alignment to optimize the degree of freedom (DoF) [3] with delayed CSIT [84, 85]. Another method investigates the channel prediction based on the channel correlation [86] and past CSI [87]. In addition, to maintain the multi-user connectivity and mitigate the degrading effect of user mobility, low complexity power allocation methods are derived in [88] for Space Division Multiple Access (SDMA) which is outperformed by Rate-Splitting Multiple Access (RSMA) in [89] in terms of ergodic sum-rate.

On the one hand, the channel prediction approaches in the papers cited above [29,30,83–87] demonstrate good performance but experience extremely high complexity in channel prediction algorithms due to the increasing dimension of the antenna arrays. On the other hand, power allocation strategies in [88, 89] exhibit lower complexity but

[3]The spatial multiplexing gain is re-branded as Degrees of Freedom (DoF) when the signal-to-noise-ratio (SNR) is let go to infinity and the system becomes interference-limited [4].

sacrifice performance for tractability. To maintain a better balance of performance and complexity, an alternative strategy with lower complexity and looser CSI requirements needs to be developed urgently.

### Integrated Artificial Intelligence and Communications

As an emerging functionality of the wireless network, artificial intelligence including machine learning (ML), deep learning (DL), and reinforcement learning (RL) [90] has demonstrated great usefulness in WCN [18–28, 31, 91–99]. To cope with complex problems in a large-dimensional MIMO system, DL has drawn research interest in not only beamforming design [93, 94] by feeding CSI to the neural network but also channel prediction [95–99] by treating the time-varying channel as a time series, taking advantage of the strong representation capability of the deep neural network (DNN). Nevertheless, under stringent time constraints in mobility scenarios, the excellent generalization performance of DNN can not be fully exploited due to an insufficient number of data samples. In view of it, by elaborately treating the time-varying channel problem as a Markov decision process (MDP), DRL has been regarded as a useful technology to design wireless communication systems by leveraging fast convergence of DL frameworks as well as continuous improvement characteristic in RL algorithms [18–28, 31, 100]. Systematically, a comprehensive tutorial in [100] reveals the applications of DRL for 5G and beyond. DRL is used to solve the power allocation problem in time-varying channels in [18] for single transmit antenna scenarios, and is further studied in [19] for multi-antenna beamforming and in [25] for multi-user conditions. In [21–23, 28], DRL is utilized to tackle the passive beamforming design problem in reconfigurable intelligent surfaces (RIS)-aided communications and help reduce the computations compared to alternative frameworks. In terms of active beamforming using DRL, several efforts have been made to design low complexity algorithms based on DQN [19, 25–27] and partially observed MDP [31] frameworks.

## 2.5 Mathematical Tools

To solve the problems formulated in this thesis, a number of mathematical tools are required, including but not limited to optimization techniques and data science techniques. We here briefly describe convex optimization technique and reinforcement learning technique. For more information, please refer to [101, 102] and references therein.

### 2.5.1 Convex Optimization

Convex optimization is a special class of mathematical optimization problems which typically includes least-squares and linear programming problems. Typically, the convex optimization problem is written as [101]

$$\min \quad f_0(x) \tag{2.16a}$$

$$\text{s.t.} \quad f_i(x) \leq b_i, \quad i = 1,...,m \tag{2.16b}$$

where the functions $f_0,...,f_m$ are convex, i.e., satisfying

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y), \forall \lambda \in [0,1] \tag{2.17}$$

After interior-point methods was developed to efficiently and optimally solve convex optimization problem, the convex optimization technique can be applied to a large class of problems such as semidefinite programming (SDP).

However, problems are usually non-convex, For non-convex problems, obtaining the optimal solution is a NP-hard problem. Some of these problems can be approximated in a relaxed or conservative manner as a convex problem. Then, the optimal solution of the approximated/relaxed problem is suboptimal but may be a good solution to the original problem. The SDR [103] and successive convex approximation (SCA) [104]

are two examples of such relaxation/approximation techniques. In Chapter 3, two different problem formulations of the joint waveform and beamforming design based on an FS-IRS and FF-IRS are solved based on SCA,and SDR.

### 2.5.2 Reinforcement Learning

In RL, an agent learns the optimal action policy to maximize the reward through trial-and-error interactions with the environment. RL is always formalized as an approach for MDP problems, which consists of $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, $\mathcal{P}$, and $\gamma$ referring to a set of states, a set of actions, a reward function, a state transition function, and the discount factor. To be specific, at time $t$, with the environment in state $s_t \in \mathcal{S}$, an agent takes an action $a_t \in \mathcal{A}$ according to policy $\pi(a_t|s_t)$, obtains a reward $r_t = \mathcal{R}(a_t,s_t)$ and next state $s_{t+1} \in \mathcal{S}$ with probability $\mathcal{P}(s_t,a_t,s_{t+1})$ in return for the action taken. Formally, each transition can be written as a tuple below

$$e_t = \langle s_t,a_t,r_t,s_{t+1} \rangle. \tag{2.18}$$

The optimal policy $\pi^*(a_t|s_t)$ is a mapping function between state and action to maximize the future accumulated reward

$$R_t(s_t,a_t) = \sum_{\tau=0}^{\infty} \gamma^\tau \mathcal{R}(s_{t+\tau+1},a_{t+\tau+1}) \tag{2.19}$$

where discount factor $\gamma \in [0,1]$ balances the significance between immediate and future rewards.

#### Multi-agent Reinforcement Learning

MARL, a subfield of RL, has been widely used in real-world scenarios, which focuses on the interaction with the environment and multiple agents. Extending from a

single-agent domain to a multi-agent environment, this method arises from the need to develop intelligent systems that can interact with other intelligent agents in complex and dynamic environments. It combines the principles of RL and multi-agent systems to enable agents to learn how to interact with other agents and the environment to achieve their goals [105]. The main idea behind MARL is to model the behavior of a group of agents, who can cooperate, compete, and even negotiate. More intuitively, different training schemes, e.g., fully decentralized, fully centralized, and centralized training and decentralized execution (CTDE), are envisioned as promising paradigms to adapt to different environments. Moreover, the existing RL algorithms can be divided into value-based algorithms, e.g, Deep Q-network (DQN) [106] and actor-critic algorithms, e.g., deep deterministic policy gradient (DDPG) [107]. In Chapter 4, three multi-agent DQN-based beamforming design algorithms are presented, followed by a multi-agent DDPG-based beamforming design algorithm in Chapter 5.

# Chapter 3

# Waveform and Beamforming Design for IRS-aided Wireless Power Networks

This chapter is concerned with the waveform and passive beamforming design for IRS-aided WPT. Generalized multi-user and low complexity single-user algorithms are demonstrated based on AO framework to maximize the weighted sum output DC, subject to the transmit power constraints and passive beamforming modulus constraints. The input signal waveform and IRS passive beamforming phase shifts are jointly designed as a function of users' individual frequency-selective CSI. The energy harvester nonlinearity is explored and two IRS deployment schemes, namely FS-IRS and FF-IRS, are modeled and analyzed. This chapter highlights the fact that IRS can provide an extra passive beamforming gain on output DC power over conventional WPT designs and significantly influence the waveform design by leveraging the benefit of passive beamforming, frequency diversity and energy harvester nonlinearity. Even though FF-IRS exhibits lower output DC than the ideal FS-IRS, it still achieves substantially increased DC power over conventional WPT designs. Performance eval-

uations confirm the significant benefits of a joint waveform and passive beamforming design accounting for the energy harvester nonlinearity to boost the performance of single-user and multi-user WPT systems.

## 3.1 Introduction

In the last decade, far-field WPT has drawn great attention because of its potential to energize billions of future autonomous low-power devices. The major challenge of far-field WPT is to maximize the output DC power or equivalently the EPTE without increasing the transmit power, and for distances over a few to hundreds of meters. Spurred by such challenges, far-field WPT has attracted widespread research attention. The traditional solution to resolve this challenge has focused on the design of efficient energy harvesters, so-called rectennas, with high RF-to-DC CE. More recently, in addition to rectenna design, a new paradigm named IRS has recently attracted the attention of the wireless research community. IRS provides a passive beamforming gain that can significantly increase the output DC power of WPT.

In this chapter, we derive an efficient IRS-aided WPT design by leveraging the progress made in the past few years on the modeling of the rectenna nonlinearity and its corresponding efficient signal design for WPT on one hand and IRS on the other hand. Motivated by the significant benefits of a joint waveform and beamforming design to expand the range and the output DC power of WPT [16], [17], [14], this work pushes the limit of IRS-aided WPT by better exploiting the presence of IRS through a joint waveform and beamforming design for both single-user and multi-user deployments.

The contributions of this chapter are summarized as follows.

*First*, we design a novel architecture based on the joint waveform and passive beamforming for IRS-aided WPT. We leverage the rectenna nonlinear model of [16] and

use it to derive a joint multisine waveform and beamforming optimization framework applicable to both single-user and multi-user IRS-aided WPT deployments. This is the first work tackling nonlinearity and waveform design for (single-user and multi-user) IRS-aided WPT. By leveraging jointly the frequency and spatial domains, and properly modeling the harvester nonlinearity, the proposed architecture exploits the combined benefits of the passive beamforming gain, the channel frequency selectivity and the rectenna nonlinearity. This contrasts with existing IRS-aided WPT papers [8–11] whose architecture only benefits from the passive beamforming gain. This also contrasts with existing papers on waveform for WPT [16,17], whose designs assume active antennas and are therefore not transferable to IRS-aided WPT due to the hardware constraints imposed by IRS.

*Second*, we develop two different formulations of the joint waveform and beamforming design based on a FS-IRS and a FF-IRS. The FF-IRS assumes that the passive beamforming phases are constant across frequencies, while the FS-IRS is flexible enough to adjust the passive beamforming phases at each frequency of the multisine waveform, which thus renders the weighted sum current problem more difficult to be solved. The performance of the FS-IRS always provides an upperbound on the FF-IRS performance and helps assess how much loss is incurred by constraining the passive beamforming phases to be the same across frequencies. This comparison sheds light on the benefits of developing more advanced IRS hardware enabling frequency-selective passive beamforming phases. This is particularly important in the context of wideband multi-band/carrier WPT since the choice of the passive beamforming phases across frequencies severely influences how much the channel frequency selectivity and the rectenna nonlinearity can be exploited by the waveform design to boost the harvested DC power. This also contrasts with existing IRS-aided WPT papers [8–11] that ignore such consideration since their designs assume narrowband transmissions and frequency-flat channels.

*Third*, under the assumption of perfect CSI, we develop an optimization framework for a general multi-user IRS-aided WPT setup based on AO, SCA, and SDR to suboptimally optimize the IRS passive beamforming phases and waveform weights with the others being fixed. An efficient solution is derived in two subproblems of AO by optimizing one of these variables in closed form. Conventional WPT waveform optimization strategies mainly rely on Reverse Geometric Program (GP) [16] and Semi-definite Program (SDP) [45,50,51]. In contrast, for the FF-IRS, both subproblems of AO are solved by formulating the problem into a standard SDP with SCA in an iterative manner. This is different from conventional WPT research because the output DC is not only influenced by the waveform weights on each frequency (which is systematically discussed in [16]) but also impacted by the passive beamforming phases in spatial and frequency domains (which is firstly promoted in this thesis). In addition, low complexity solutions are demonstrated in multi-user (MU) and single-user (SU) FS-IRS algorithms with the element-wise updating method and prior determined passive beamforming phases, respectively.

*Fourth*, numerical results are displayed to validate the theoretical discussion on FS-IRS and FF-IRS WPT algorithms. Although FS-IRS always provides an upperbound performance to FF-IRS, FF-IRS still outperforms conventional WPT papers which emphasizes the significance of passive beamforming designs. Then, FS-IRS brings more opportunities to broadband WPT while FF-IRS is more suitable for narrowband WPT. Both FS-IRS and FF-IRS algorithms can exhibit fast convergence which strongly reveals the effectiveness of the SCA-based algorithms. Furthermore, the IRS should be deployed closed to users or BS to obtain a larger current region. Moreover, a near-optimal result can be achieved with low-resolution discrete-phase IRS in both FF-IRS and FS-IRS. Numerical results confirm the inefficiency of the linear-based model and demonstrate the advancement of IRS-aided WPT on leveraging the nonlinearity of rectenna, waveform weights and passive beamforming phases. In addition, this model can be extended to the multi-antenna case as in [16].

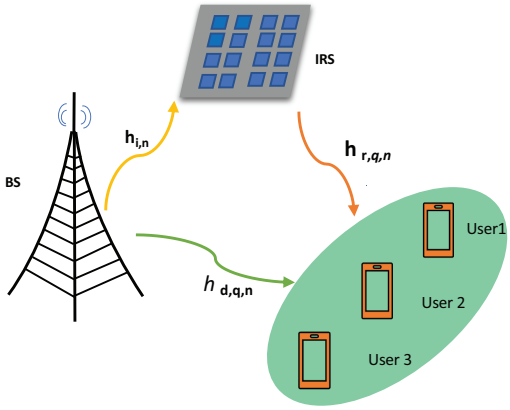



Figure 3.1: System model.

## 3.2 System Model

### 3.2.1 Transmit Signal

As is illustrated in Fig. 3.1, we consider an IRS-aided downlink single-input-single-output (SISO) WPT system including $K$ single-antenna users, one IRS with $L$ REs and one BS equipped with single antenna at the transmitter. The multisine waveform at time $t$ is given as

$$x(t)=\Re\left\{\sum_{n=1}^{N}s_n e^{j2\pi f_n t}\right\}=\sum_{n=1}^{N}w_n\cos(2\pi f_n t+\gamma_n) \tag{3.1}$$

where $s_n=w_n e^{j\gamma_n}$ $(n=1,...,N)$ denotes the complex weight of the multisine waveform with $w_n$ and $\gamma_n$ accounting for the amplitude and phase at frequency $n$, respectively. For simplicity, we assume that the frequencies are equally spaced, i.e. $f_n=f_0+n\Delta_f$. The transmit power constraint at the transmitter is written as $\frac{1}{2}\|\mathbf{s}\|^2\leq P$ with $\mathbf{s}=[s_1,s_2,...,s_N]^T\in\mathbb{C}^{N\times 1}$ and $\frac{1}{2}$ referring to the average power of cosine function in (3.1).

### 3.2.2 Channel Model and Reflection Pattern

The multisine waveform propagates through wireless channels. The channel from the BS to user $q$, from the BS to the IRS and from the IRS to user $q$ at the frequency $n$ are denoted as $h_{\mathrm{d},q,n}$, $\mathbf{h}_{\mathrm{i},n} \in \mathbb{C}^{L\times 1}$ and $\mathbf{h}_{\mathrm{r},q,n} \in \mathbb{C}^{1\times L}$, respectively. To facilitate the reading, they are also termed as direct channel, incident channel ,and reflected channel, respectively. To be specific, we denote the entries of $\mathbf{h}_{\mathrm{i},n}$ as $h_{\mathrm{i},l+(n-1)L}$ and the entries of reflected channel $\mathbf{h}_{\mathrm{r},q,n}$ as $h_{\mathrm{r},q,l+(n-1)L}$ with $\mathbf{h}_{\mathrm{i},n} = [h_{\mathrm{i},1+(n-1)L},...,h_{\mathrm{i},L+(n-1)L}]^T$ and $\mathbf{h}_{\mathrm{r},q,n} = [h_{\mathrm{r},q,1+(n-1)L},...,h_{\mathrm{r},q,L+(n-1)L}]$, respectively. The channels above are assumed quasi-static and the CSIT is assumed to be perfect.

Two IRS deployment scenarios are studied here, namely a FS-IRS scenario and a FF-IRS scenario. FS-IRS is flexible enough to adjust the passive beamforming phases in different frequencies. In contrast, FF-IRS assumes that passive beamforming phases are constant across frequencies.

#### FS-IRS

The complex phasors $e^{j\psi_{l+(n-1)L}}, l = 1,2,...,L$, at frequency $n$ are collected on the diagonal entries of the diagonal matrix $\mathbf{\Theta}_n \in \mathbb{C}^{L\times L}$ with $\mathbf{\Theta}_n = \mathrm{diag}\{e^{j\psi_{1+(n-1)L}},...,e^{j\psi_{L+(n-1)L}}\}, \forall n$. We assume that $\psi_{l+(n-1)L} \in [0,2\pi)$ and $\theta_{l+(n-1)L} = e^{j\psi_{l+(n-1)L}}, l = 1,2,...,L$.

#### FF-IRS

We design a common set of passive beamforming phases that cater to all subcarriers i.e. $\mathbf{\Theta}_n = \mathbf{\Theta} = \mathrm{diag}\{e^{j\psi_1},e^{j\psi_2},...,e^{j\psi_L}\}, \forall n$. With $\psi_l \in [0,2\pi)$, we have $\theta_l = e^{j\psi_l}, \forall l$ in the remainder of this work.

By superposing the auxiliary channel ($\mathbf{h}_{\mathrm{r},q,n}\mathbf{\Theta}_n\mathbf{h}_{\mathrm{i},n}$) to the direct channel, we can

obtain the composite channel between BS and user $q$ at frequency $n$ as

$$h_{q,n} = h_{\mathrm{d},q,n} + \mathbf{h}_{\mathrm{r},q,n} \boldsymbol{\Theta}_n \mathbf{h}_{\mathrm{i},n}. \tag{3.2}$$

**Remark 1.** *This work does not deal with the implementation of FS-IRS but rather uses FS-IRS as an upper bound on FF-IRS to assess how much performance could be obtained if FS-IRS could be prototyped by RF engineers. According to [60], the phase shift essentially depends on the reflection coefficient of a reconfigurable impedance. Then, the frequency response of phase shift depends on the frequency response of a particular impedance. Generally, once the circuit topology is fixed, the frequency response is fixed so that it is more complex to implement FS-IRS than FF-IRS. A practical IRS model is illustrated in [108] which emphasizes that the amplitudes and phases of passive beamforming phases vary with the frequencies of transmitted signals. Nevertheless, there is still a lack of a general model that can accurately characterize this nonlinear relationship between the practical frequency-dependent passive beamforming phases and nonlinear EH by capturing all factors. Meanwhile, due to the different nature of systems, the objective functions and the receiver architecture, the benefits of FS-IRS over FF-IRS can be much different in WPT and in wideband OFDM communications. This means that any conclusions from IRS-aided communications may not be transferable to WPT. Hence, FF-IRS with unit modulus constraints are adopted in this work as widely utilized in existing literature [64, 65]. As this is the first work proposing FS-IRS and FF-IRS with a nonlinear energy harvester model, we would like to keep the system model as clear and simple as possible such that researchers can understand the fundamental benefits of the two proposed strategies and carry on their studies in more practical senarios in the future.*

### 3.2.3 Receive Signal

The received signal by user $q$ at time $t$ for FS-IRS can be expressed as[1]

$$y_q(t) = \Re\left\{ \sum_{n=1}^{N} (h_{\mathrm{d},q,n} + \mathbf{h}_{\mathrm{r},q,n} \boldsymbol{\Theta}_n \mathbf{h}_{\mathrm{i},n}) s_n e^{j2\pi f_n t} \right\} \tag{3.3}$$

where $q = 1,2,...,K$ and $y_q(t)$ is the input of the rectenna which is discussed in the next part. In FF-IRS, $\boldsymbol{\Theta}_n$ is replaced with $\boldsymbol{\Theta}, \forall n$.

## 3.3 Joint Design of Waveform and Reflection Elements for Multi-user Scenario

The purpose of the multi-user designs is to jointly optimize the frequency domain complex weights of the waveform and passive beamforming phases of IRS to maximize the $K$-user weighted sum output DC, subject to the input power constraint ($\frac{1}{2}\|\mathbf{s}\|^2 \leq P$) as well as the unit modulus constraints ($\left|\theta_{l+(n-1)L}\right| = 1$). Using the model of (2.6) [16] and truncating to order 4, i.e. $n_0 = 4$, the approximated DC of user $q$ is $i_{\mathrm{dc},q} = \sum_{n,\ \text{even}\ ,n\geq 2}^{4} \beta_n R_{ant}^{n/2} \mathcal{E}\{y_q(t)^n\}$. $i_{\mathrm{dc},q}$ can be found as

$$i_{\mathrm{dc},q} = k_2 \mathcal{E}\{y_q(t)^2\} + k_4 \mathcal{E}\{y_q(t)^4\} \tag{3.4}$$

where $k_2 = \beta_2 R_{\mathrm{ant}} = 0.17$ and $k_4 = \beta_4 R_{\mathrm{ant}}^2 = 957.25$ [16] (assuming that $i_{\mathrm{s}} = 5\mu\mathrm{A}, n' = 1.05$ and $v_{\mathrm{t}} = 25.86\mathrm{mV}$). Then, $\mathcal{E}\{y(t)^2\}$, $\mathcal{E}\{y(t)^4\}$ and the DC output current $i_{\mathrm{dc},q}$ are calculated as

$$\mathcal{E}\{y_q(t)^2\} = \frac{1}{2} \sum_{n=1}^{N} s_n^* h_{q,n}^* h_{q,n} s_n, \tag{3.5}$$

[1] In this thesis, we ignore the signals which are reflected by IRS for two or more times due to substantial path loss.

$$\mathcal{E}\{y_q(t)^4\} = \frac{3}{8} \sum_{\substack{n_1,n_2,n_3,n_4 \\ n_1+n_3=n_2+n_4}} s_{n_1}^* h_{q,n_1}^* h_{q,n_2} s_{n_2} s_{n_3}^* h_{q,n_3}^* h_{q,n_4} s_{n_4}, \tag{3.6}$$

$$i_{\mathrm{dc},q}(\mathbf{s},\{\mathbf{\Theta}_n\}_{n=1}^N) = \frac{1}{2}k_2 \sum_{n=1}^N s_n^* h_{q,n}^* h_{q,n} s_n + \frac{3}{8}k_4 \sum_{\substack{n_1,n_2,n_3,n_4 \\ n_1+n_3=n_2+n_4}} s_{n_1}^* h_{q,n_1}^* h_{q,n_2} s_{n_2} s_{n_3}^* h_{q,n_3}^* h_{q,n_4} s_{n_4}, \tag{3.7}$$

respectively, where the composite channel $h_{q,n}$ is given by (3.2). [2]. The corresponding problem can be formulated as a weighted sum of DC

$$\max_{\mathbf{s},\{\mathbf{\Theta}_n\}_{n=1}^N} \quad \sum_{q=1}^K \xi_q i_{\mathrm{dc},q}(\mathbf{s},\{\mathbf{\Theta}_n\}_{n=1}^N) \tag{3.8a}$$

$$\text{s.t.} \quad \|\mathbf{s}\|^2 \leq 2P, \tag{3.8b}$$

$$|\theta_{l+(n-1)L}| = 1, n = 1,2,...,N, l = 1,2,...,L \tag{3.8c}$$

where $\xi_q$ is the weight for the $q$th user.

Due to the intricate coupled variables in problem (3.8), we apply the SCA based AO method to separate problem (3.8) into two subproblems to update 1) the passive beamforming phases, 2) the waveform complex weights, in an iterative manner until convergence for both FF-IRS and FS-IRS. 3) the convergence can be guaranteed in several iterations which is shown in Section 3.8.2 with a suboptimal solution proved in Section 3.3.5.

In the next three sections, at iteration $i$, the passive beamforming designs for MU FF-IRS and MU FS-IRS and the waveform design for both MU IRS deployments scenarios are sequentially studied. In the passive beamforming subproblem, optimal $\{\mathbf{\Theta}_n^{(i)}\}_{n=1}^N$ is obtained with waveform weight vector $\mathbf{s}^{(i-1)}$ being fixed. In the waveform design subproblem, $\mathbf{s}^{(i)}$ is solved by using $\{\mathbf{\Theta}_n^{(i)}\}_{n=1}^N$ which is achieved in the previous subproblem.

[2]When it comes to the multi-antenna scenario, the objective function (3.7) becomes $i_{dc}(\mathbf{s},\mathbf{\Theta}) = \frac{1}{2}k_2 \sum_{n=1}^N \mathbf{s}_n^H \mathbf{h}_{k,n}^H \mathbf{h}_{k,n} \mathbf{s}_n + \frac{3}{8}k_4 \sum_{\substack{n_1,n_2,n_3,n_4 \\ n_1+n_3=n_2+n_4}} \mathbf{s}_{n_1}^H \mathbf{h}_{k,n_1}^H \mathbf{h}_{k,n_2} \mathbf{s}_{n_2} \mathbf{s}_{n_3}^H \mathbf{h}_{k,n_3}^H \mathbf{h}_{k,n_4} \mathbf{s}_{n_4}$ where $\mathbf{s}_n = [s_{1+(n-1)M_t}, s_{2+(n-1)M_t}, ..., s_{M_t+(n-1)M_t}]$ and $\mathbf{h}_{k,n} = \mathbf{h}_{\mathrm{d},q,n} + \mathbf{h}_{\mathrm{r},q,n} \mathbf{\Theta}_n \mathbf{H}_{\mathrm{i},n} \in \mathbb{C}^{1\times M_t}$.

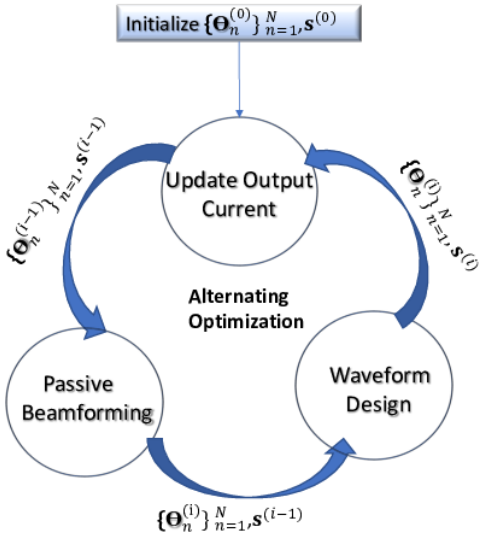



Figure 3.2: Alternating Optimization in one iteration.

### 3.3.1 Passive Beamforming for Multi-User FF-IRS

In FF-IRS, we optimize $L$ passive beamforming phases which are constrained to be constant across all $N$ frequencies. To build this weighted sum current subproblem on passive beamforming phases, we denote $\mathbf{v}_{q,n} = \begin{bmatrix} \text{diag}\{\mathbf{h}_{\text{r},q,n}\}\mathbf{h}_{\text{i},n} \\ h_{\text{d},q,n} \end{bmatrix} \in \mathbb{C}^{(L+1)\times 1}$. Denote $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_L, m_1] \in \mathbb{C}^{1\times(L+1)}$ with $m_1$ referring to an auxiliary variable. Then $(h_{\text{d},q,n} + \mathbf{h}_{\text{r},q,n}\boldsymbol{\Theta}\mathbf{h}_{\text{i},n})s_n = \boldsymbol{\theta}\mathbf{v}_{q,n}s_n$ and we let $\mathbf{z}_{q,n} = \mathbf{v}_{q,n}s_n$ which is grouped into $\mathbf{z}_q = [\mathbf{z}_{q,1}^T, \mathbf{z}_{q,2}^T, \ldots, \mathbf{z}_{q,N}^T]^T \in \mathbb{C}^{(L+1)N\times 1}$. The $i_{\text{dc},q}(\mathbf{s}, \{\boldsymbol{\Theta}_n\}_{n=1}^N)$ in (3.8a) can be equivalently rewritten as $i_{\text{dc},q}(\boldsymbol{\theta})$ below

$$i_{\text{dc},q}(\boldsymbol{\theta}) = \tfrac{1}{2}k_2 \sum_{n=1}^{N} \boldsymbol{\theta}\mathbf{z}_{q,n}\mathbf{z}_{q,n}^H\boldsymbol{\theta}^H + \tfrac{3}{8}k_4 \sum_{\substack{n_1,n_2,n_3,n_4 \\ n_1+n_3=n_2+n_4}} \boldsymbol{\theta}\mathbf{z}_{q,n_1}\mathbf{z}_{q,n_2}^H\boldsymbol{\theta}^H\boldsymbol{\theta}\mathbf{z}_{q,n_3}\mathbf{z}_{q,n_4}^H\boldsymbol{\theta}^H. \tag{3.9}$$

To transform (3.9) into a more compact form, we define $N(L+1)$-by-$N(L+1)$ matrix $\mathbf{D}_q$ and $(L+1)$-by-$(L+1)$ matrices $\mathbf{D}_{q,k}$ where $\mathbf{D}_q = \mathbf{z}_q\mathbf{z}_q^H$ which is shown in Fig. 3.3. $k = 0$ refers to the index of the main block diagonal of $\mathbf{D}_q$. $k \in \{1, \ldots, N-1\}$ denotes the index

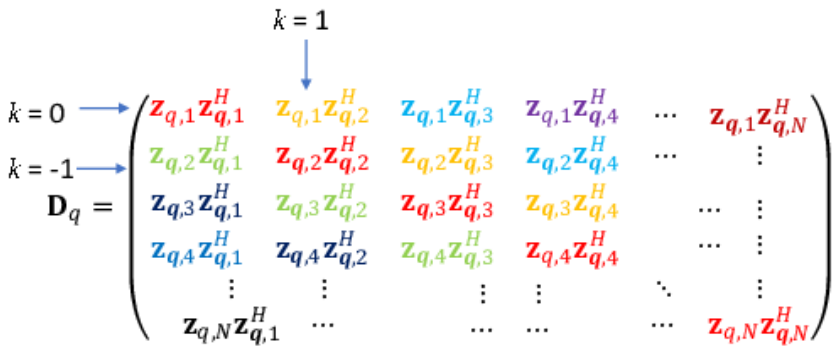



Figure 3.3: $\mathbf{D}_{q,0}$ is the sum of red block diagonal, $\mathbf{D}_{q,1}$ is the sum of yellow block diagonal and $\mathbf{D}_{q,-1}$ is the sum of green block diagonal in the above matrix.

of the $k$th block diagonal above the main block diagonal of the matrix $\mathbf{D}_q$. $k \in \{-(N-1),\ldots,-1\}$ denotes the index of the $|k|$th block diagonal below the main block diagonal. Matrix $\mathbf{D}_{q,k}$ is obtained by summing the block matrices of the $k$th block diagonal of $\mathbf{D}_q$ e.g. $\mathbf{D}_{q,0} = \mathbf{z}_{q,1}\mathbf{z}_{q,1}^H + \mathbf{z}_{q,2}\mathbf{z}_{q,2}^H + \cdots + \mathbf{z}_{q,N}\mathbf{z}_{q,N}^H$, $\mathbf{D}_{q,1} = \mathbf{z}_{q,1}\mathbf{z}_{q,2}^H + \mathbf{z}_{q,2}\mathbf{z}_{q,3}^H + \cdots + \mathbf{z}_{q,N-1}\mathbf{z}_{q,N}^H$ and $\mathbf{D}_{q,N-1} = \mathbf{z}_{q,1}\mathbf{z}_{q,N}^H$. Luckily, we can get a compact form of (3.9) for the $q$th user as

$$i_{\mathrm{dc},q}(\boldsymbol{\theta}) = \tfrac{1}{2}k_2\boldsymbol{\theta}\mathbf{D}_{q,0}\boldsymbol{\theta}^H + \tfrac{3}{8}k_4\boldsymbol{\theta}\mathbf{D}_{q,0}\boldsymbol{\theta}^H(\boldsymbol{\theta}\mathbf{D}_{q,0}\boldsymbol{\theta}^H)^H + \tfrac{3}{4}k_4\sum_{k=1}^{N-1}\boldsymbol{\theta}\mathbf{D}_{q,k}\boldsymbol{\theta}^H(\boldsymbol{\theta}\mathbf{D}_{q,k}\boldsymbol{\theta}^H)^H. \tag{3.10}$$

Then, we formulate weighted sum current problem which is subject to the unit modulus constraints

$$\max_{\boldsymbol{\theta}}\{\sum_{q=1}^{K}\xi_q i_{\mathrm{dc},q}(\boldsymbol{\theta}) : |\theta_l| = 1, l = 1,2,\ldots,L\}. \tag{3.11}$$

However, this problem is a quartic polynomial (due to term $\frac{3}{4}k_4\sum_{k=1}^{N-1}\boldsymbol{\theta}\mathbf{D}_{q,k}\boldsymbol{\theta}^H(\boldsymbol{\theta}\mathbf{D}_{q,k}\boldsymbol{\theta}^H)^H$) which is NP-hard in general. To unveil this problem, we introduce auxiliary variable $d_{q,k}$ where $k \in [0,\ldots,N-1]$ and $d_{q,k} = \boldsymbol{\theta}\mathbf{D}_{q,k}\boldsymbol{\theta}^H$. However, for $k \neq 0$, $\mathbf{D}_{q,k}$ is non-Hermitian matrix and the problem becomes a quadratic polynomial which is also NP-hard. To deal with this, we introduce a rank-one positive semidefinite matrix variable

$\mathbf{X}=\boldsymbol{\theta}^H\boldsymbol{\theta}\in\mathbb{C}^{(L+1)\times(L+1)}$ to linearize the problem with $d_{q,k}=\boldsymbol{\theta}\mathbf{D}_{q,k}\boldsymbol{\theta}^H=\text{Tr}\{\mathbf{D}_{q,k}\boldsymbol{\theta}^H\boldsymbol{\theta}\}=\text{Tr}\{\mathbf{D}_{q,k}\mathbf{X}\}$. Then, the problem becomes a linear problem with non-convex constraint $\text{rank}(\mathbf{X})=1$. After defining $\mathbf{d}_q=[d_{q,0},d_{q,1},...,d_{q,N-1}]^T$, $\mathbf{K}_0=\text{diag}\{\frac{3}{8}k_4,\frac{3}{4}k_4,...,\frac{3}{4}k_4\}\succeq 0$ and the diagonal element of $\mathbf{X}$ as $\mathbf{X}_{l',l'}$ with $l'=1,2,...,L+1$, problem (3.11) can be transformed to an equivalent problem as

$$\min_{t_0,\{\mathbf{d}_q\}_{q=1}^K} \quad t_0 \tag{3.12a}$$

$$\text{s.t.} \quad \sum_{q=1}^{K}\xi_q(-\frac{1}{2}k_2 d_{q,0}-\mathbf{d}_q^H\mathbf{K}_0\mathbf{d}_q)-t_0\leq 0, \tag{3.12b}$$

$$\text{Tr}\{\mathbf{D}_{q,k}\mathbf{X}\}=d_{q,k},\forall q,k, \tag{3.12c}$$

$$\mathbf{X}_{l',l'}=1,\forall l', \tag{3.12d}$$

$$\mathbf{X}\succeq 0, \tag{3.12e}$$

$$\text{rank}(\mathbf{X})=1 \tag{3.12f}$$

which is still a non-convex problem because of the existence of rank constraint (3.12f). We apply SDR to relax (3.12f) and focus on relaxed problems (3.12a)-(3.12e). As the term $\mathbf{d}_q^H\mathbf{K}_0\mathbf{d}_q$ in (3.12b) is a non-convex quadratic constraint, SCA can be applied to address this problem to approximate the non-convex constraint into a convex constraint and solve the approximated convex problem iteratively, the performance of this iteration is explained shown in Fig. 3.8.2. At iteration $i$, the optimal $\mathbf{d}_q^{(i)}$ can be approximated by the optimal $\mathbf{d}_q^{(i-1)}$ via Taylor expansion [109]. With $f(\mathbf{d}_q)=\mathbf{d}_q^H\mathbf{K}_0\mathbf{d}_q$, we have

$$f\left(\mathbf{d}_q,\mathbf{d}_q^{(i-1)}\right)\triangleq 2\Re\left\{\mathbf{d}_q^{(i-1)H}\mathbf{K}_0\mathbf{d}_q\right\}-\mathbf{d}_q^{(i-1)H}\mathbf{K}_0\mathbf{d}_q^{(i-1)}. \tag{3.13}$$

From [104], (3.13) has the pattern that $f(\mathbf{d}_q^{(i)},\mathbf{d}_q^{(i)})=f(\mathbf{d}_q^{(i)})\geq f(\mathbf{d}_q^{(i)},\mathbf{d}_q^{(i-1)})$. Accordingly, non-convex constraint (3.12b) has the property that $\sum_{q=1}^K\xi_q(-\frac{1}{2}k_2 d_{q,0}-f(\mathbf{d}_q^{(i)}))\leq\sum_{q=1}^K\xi_q(-\frac{1}{2}k_2 b_{q,0}-f(\mathbf{d}_q^{(i)},\mathbf{d}_q^{(i-1)}))\leq t_0$. Since inequation $\sum_{q=1}^K\xi_q(-\frac{1}{2}k_2 b_{q,0}-f(\mathbf{d}_q^{(i)},\mathbf{d}_q^{(i-1)}))\leq t_0$ is a convex constraint at iteration $i$, relaxed problems (3.12a)-(3.12e)

can be formulated as a standard SDP as

$$\min_{t_0,\mathbf{d}_q,\mathbf{X}\succeq 0} \quad t_0 \tag{3.14a}$$

$$\text{s.t.} \quad \sum_{q=1}^{K}\xi_q(-\frac{1}{2}k_2 d_{q,0}-f(\mathbf{d}_q,\mathbf{d}_q^{(i-1)}))-t_0\leq 0, \tag{3.14b}$$

$$(3.12c),(3.12d) \text{ and } (3.12e). \tag{3.14c}$$

By substituting (3.12c) into (3.14b), it follows that

$$\min \quad \text{Tr}\{\mathbf{K}_1\mathbf{X}\} \tag{3.15a}$$

$$\text{s.t.} \quad \mathbf{X}_{l',l'}=1,l'=1,2,...,L+1, \tag{3.15b}$$

$$\mathbf{X}\succeq 0 \tag{3.15c}$$

where $\mathbf{K}_1=\mathbf{J}_1+\mathbf{J}_1^H$ is a Hermitian matrix and

$$\mathbf{J}_1=\sum_{q=1}^{K}\xi_q\big(-\tfrac{k_2}{4}\mathbf{D}_{q,0}-\tfrac{3}{8}k_4 d_{q,0}^{(i-1)}\mathbf{D}_{q,0}-\tfrac{3}{4}k_4\textstyle\sum_{k=1}^{N-1}d_{q,k}^{(i-1)*}\mathbf{D}_{q,k}\big). \tag{3.16}$$

This standard SDP problem (3.15) can be solved by invoking existing softwares, e.g., CVX MATLAB [110]. Denoting the solution as $\mathbf{X}^\star$, if $\text{rank}(\mathbf{X}^\star)=1$, the SDR is tight and $\mathbf{X}^\star$ is a stationary point of (3.15) so that a local optimal solution can be extracted by $\mathbf{X}^\star=\boldsymbol{\theta}^{\star(i)}\boldsymbol{\theta}^{H\star(i)}$ [111]. If $\text{rank}(\mathbf{X}^\star)>1$, we can use the Gaussian randomization method in [112, 113] to extract a suboptimal rank-one solution $\boldsymbol{\theta}'^{(i)}=[\theta_1'^{(i)},\theta_2'^{(i)},...,\theta_L'^{(i)},m_1^{(i)}]$, retrieve the phase shifts by $\boldsymbol{\theta}^{\star(i)}=\boldsymbol{\theta}'^{(i)}/m_1^{(i)}$ and group the first $L$ elements into the diagonal of $\boldsymbol{\Theta}^{\star(i)}=\text{diag}\{\boldsymbol{\theta}^{\star(i)}[1:L]\}$. Then, we have the composite channel $h_{q,n}=(h_{\text{d},q,n}+\mathbf{h}_{\text{r},q,n}\boldsymbol{\Theta}^{\star(i)}\mathbf{h}_{\text{i},n})$.

**Remark 2.** *Note that different from [8] where a predefined minimum harvested power threshold is addressed, problem (3.15) is a standard SDP in each iteration with only unit modulus constraint. The known approximation accuracy for this complex constant*

*modulus problem is $\pi/4$ which guarantees that a good approximation can be found if we have a rank-one solution [111]. When we evaluated the results in Section 3.8, we found that all channel realizations generated a rank-one $\mathbf{X}^\star$. This is particularly important since Algorithm 2 can provide a stationary point for problem (3.8) with a rank-one solution in the simulations (see Appendix A.2). Even though the performance is degraded by the Gaussian randomization method later, the performance loss is still negligible if $\mathbf{X}^\star$ is rank-one for all tested channels [58].*

### 3.3.2 Passive Beamforming for Multi-User FS-IRS

In MU FS-IRS, different from problem (3.11), we aim to maximize the weighted sum current subject to $NL$ unit modulus constraints and the problem can be written as

$$\max_{\{\boldsymbol{\Theta}_n\}_{n=1}^N} \{\sum_{q=1}^{K} \xi_q i_{\mathrm{dc},q}(\{\boldsymbol{\Theta}_n\}_{n=1}^N): \tag{3.17}$$
$$|\theta_{l+(n-1)L}|=1, n\in 1,2,...,N, l\in 1,2,...,L\}.$$

To obtain a compact form , we define $\mathbf{v}_{q,n}=\begin{bmatrix}\mathrm{diag}\{\mathbf{h}_{\mathrm{r},q,n}\}\mathbf{h}_{\mathrm{i},n}\\ h_{\mathrm{d},q,n}\end{bmatrix}\in\mathbb{C}^{(L+1)\times 1}$. We denote $\boldsymbol{\theta}_n=[\theta_{1+(n-1)L},\dots,\theta_{L+(n-1)L},m_n]\in\mathbb{C}^{1\times(L+1)},\forall n$ with $m_n$ referring to an auxiliary variable. We have $(h_{\mathrm{d},q,n}+\mathbf{h}_{\mathrm{r},q,n}\boldsymbol{\Theta}_n\mathbf{h}_{\mathrm{i},n})s_n=\boldsymbol{\theta}_n\mathbf{v}_{q,n}s_n$. Denote $\mathbf{z}_{q,n}=\mathbf{v}_{q,n}s_n\in\mathbb{C}^{(L+1)\times 1}$ which is grouped into $\mathbf{z}_q=[\mathbf{z}_{q,1}^T,\mathbf{z}_{q,2}^T,...,\mathbf{z}_{q,N}^T]^T\in\mathbb{C}^{(L+1)N\times 1}$. (3.7) can be rewritten as

$$i_{\mathrm{dc},q}(\{\boldsymbol{\theta}_n\}_{n=1}^N)=\tfrac{1}{2}k_2\textstyle\sum_{n=1}^N\boldsymbol{\theta}_n\mathbf{z}_{q,n}\mathbf{z}_{q,n}^H\boldsymbol{\theta}_n^H+\tfrac{3}{8}k_4\sum_{\substack{n_1,n_2,n_3,n_4\\ n_1+n_3=n_2+n_4}}\boldsymbol{\theta}_n\mathbf{z}_{q,n_1}\mathbf{z}_{q,n_2}^H\boldsymbol{\theta}_n^H\boldsymbol{\theta}_n\mathbf{z}_{q,n_3}\mathbf{z}_{q,n_4}^H\boldsymbol{\theta}_n^H. \tag{3.18}$$

To formulate (3.18) into a more tractable compact form, we denote $\boldsymbol{\theta}=[\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_N]\in\mathbb{C}^{1\times N(L+1)}$ and define $(L+1)N$-by-$(L+1)N$ matrices $\mathbf{E}_q$ and $\mathbf{E}_{q,k}$ with $\mathbf{E}_q=\mathbf{z}_q\mathbf{z}_q^H$. $k=0$ denotes the index of the main block diagonal in $\mathbf{E}_q$. $k\in\{1,...,N-1\}$ denotes the index of the $k$th block diagonal above main diagonal of $\mathbf{E}_q$. $k\in\{-(N-1),...,-1\}$

denotes the index of the $|k|$th block diagonal below the main block diagonal. $\mathbf{E}_{q,k}$ is obtained by retaining the $k$th block diagonal of $\mathbf{E}_q$ and setting other block matrices as $\mathbf{0}_{(L+1)\times(L+1)}$. Then, the output current for user $q$ can be equivalently written as

$$i_{\mathrm{dc},q}(\boldsymbol{\theta})=\tfrac{1}{2}k_2\boldsymbol{\theta}\mathbf{E}_{q,0}\boldsymbol{\theta}^H+\tfrac{3}{8}k_4\boldsymbol{\theta}\mathbf{E}_{q,0}\boldsymbol{\theta}^H(\boldsymbol{\theta}\mathbf{E}_{q,0}\boldsymbol{\theta}^H)^H+\tfrac{3}{4}k_4\sum_{k=1}^{N-1}\boldsymbol{\theta}\mathbf{E}_{q,k}\boldsymbol{\theta}^H(\boldsymbol{\theta}\mathbf{E}_{q,k}\boldsymbol{\theta}^H)^H. \tag{3.19}$$

To linearize the quartic function (3.19), we take $\boldsymbol{\theta}\mathbf{E}_{q,k}\boldsymbol{\theta}^H=e_{q,k}$ which is collected into $\mathbf{e}_q=[e_{q,0},e_{q,1},...,e_{q,N-1}]^T$. (3.19) can be written as $i_{\mathrm{dc},q}(\mathbf{e}_q)=\frac{1}{2}k_2e_{q,0}+\mathbf{e}_q^H\mathbf{K}_0\mathbf{e}_q$. The problem can be formulated as

$$\min_{t_1,\{\mathbf{e}_q\}_{q=1}^K} \quad t_1 \tag{3.20a}$$

$$\text{s.t.} \quad \sum_{q=1}^{K}\xi_q(-\frac{1}{2}k_2e_{q,0}-\mathbf{e}_q^H\mathbf{K}_0\mathbf{e}_q)-t_1\le 0, \tag{3.20b}$$

$$\boldsymbol{\theta}\mathbf{E}_{q,k}\boldsymbol{\theta}^H=e_{q,k},\forall q,k, \tag{3.20c}$$

$$\left|\theta_{l+(n-1)L}\right|=1,\forall n,l. \tag{3.20d}$$

We still approximate the non-convex constraint (3.20b) iteratively by the SCA which is similar with that in Section 3.3.1 and the problem can be reformulated as

$$\begin{aligned}\min_{\boldsymbol{\theta}} \quad & \boldsymbol{\theta}\mathbf{K}_2\boldsymbol{\theta}^H \\ \text{s.t.} \quad & (3.20c) \text{ and } (3.20d)\end{aligned} \tag{3.21}$$

where $\mathbf{K}_2=\mathbf{J}_2+\mathbf{J}_2^H$ is a Hermitian matrix and $\mathbf{J}_2$ is

$$\mathbf{J}_2=\sum_{q=1}^{K}\xi_q\big(-\tfrac{k_2}{4}\mathbf{E}_{q,0}-\tfrac{3}{8}k_4e_{q,0}^{(i-1)}\mathbf{E}_{q,0}-\tfrac{3}{4}k_4\textstyle\sum_{k=1}^{N-1}[e_{q,k}^{(i-1)}]^*\mathbf{E}_{q,k}\big). \tag{3.22}$$

If we apply the same approach as in Section 3.3.1 to problem (3.21), that is, by denoting $\mathbf{X}=\boldsymbol{\theta}^H\boldsymbol{\theta}$ and formulating problem (3.21) into a SDP, solving this intricate SDP will render a complexity around $\mathrm{O}((NL)^6)$ which demonstrates low feasibility for

large number of frequencies. To tackle this issue, here, we illustrate a low complexity strategy to optimize the $NL$ elements in $\boldsymbol{\theta}$ with a complexity around $\mathrm{O}((NL)^2)$.

*Element-Wise Updating Method*: Element-Wise Updating Method (EWU) sequentially optimizes $N(L+1)$ variables ($NL$ passive beamforming phases variables and $L$ auxiliary variables) in $\boldsymbol{\theta}$ which is also applied in [114] and [62]. To facilitate the reading, in this method, we denote $\mathbf{K}=-\mathbf{K}_2$ and the element at the $i$th row and the $j$th column of $\mathbf{K}$ as $k_{i,j}$,$i=1,2,...,N(L+1)$,$j=1,2,...,N(L+1)$. In addition, we denote the $m$th variable in $\boldsymbol{\theta}$ as $\theta_m$ in this part for simplicity (instead of $\theta_{l+(n-1)L}$ in the preceding discussion). minimization of $\boldsymbol{\theta}\mathbf{K}_2\boldsymbol{\theta}^H$ in problem (3.21) becomes maximization of $\boldsymbol{\theta}\mathbf{K}\boldsymbol{\theta}^H$ with the same unit modulus constraints (3.20d). Since $\theta_m$ is updated with other $N(L+1)-1$ elements in $\boldsymbol{\theta}$ being fixed, $\boldsymbol{\theta}\mathbf{K}\boldsymbol{\theta}^H$ can be written as an element-wise function $f(\theta_m)$ below

$$f(\theta_m)=\theta_m k_{m,m}\theta_m^*+\sum_{j=1,j\neq m}^{N(L+1)}\theta_m k_{m,j}\theta_j^*+\sum_{i=1,i\neq m}^{N(L+1)}\theta_i k_{i,m}\theta_m^*+\sum_{i\neq m,j\neq m}\theta_i k_{i,j}\theta_j \tag{3.23}$$

where $m=1,...,N(L+1)$.

Due to the fact that $\mathbf{K}$ is a Hermitian matrix, we substitute $k_{i,j}=k_{j,i}^*$ and yield

$$f(\theta_m)=\theta_m k_{m,m}\theta_m^*+2\Re\{\sum_{j=1,j\neq m}^{N(L+1)}\theta_m k_{m,j}\theta_j^*\}+\sum_{i\neq m,j\neq m}\theta_i k_{i,j}\theta_j. \tag{3.24}$$

In order to optimize $\theta_m$, we abandon other irrelevant elements. The objective function $g(\theta_m)$ becomes

$$g(\theta_m)=k_{m,m}|\theta_m^*|^2+2\Re\{\sum_{j=1,j\neq m}^{N(L+1)}\theta_m k_{m,j}\theta_j^*\}. \tag{3.25}$$

Since $|\theta_m^*|^2=1$, maximizing $g(\theta_m)$ is to maximize the term $\Re\{\sum_{j=1,j\neq m}^{N(L+1)}\theta_m k_{m,j}\theta_j^*\}$. After denoting $\sum_{j=1,j\neq m}^{N(L+1)}k_{m,j}\theta_j^*=C_m$, the optimal $\theta_m^\star$ is

$$\angle\theta_m^\star=-\angle C_m,\theta_m^\star=e^{-j\angle C_m}. \tag{3.26}$$

**Algorithm 1** EWU Algorithm

1: **Initialize**: Set $t=0$, tolerance $\tau$, feasible $\theta_m^{(0)}$ and corresponding objective function value $f(\boldsymbol{\theta}^{(0)})$;
2: **Repeat**
3: $t=t+1$;
4: Compute $g(\theta_m^{(t)})$ in (3.25) and update $\theta_m^{(t)}$ in (3.26). Note that $\theta_m=1$ when $m$ is integer multiple of $L+1$;
5: Compute $f(\boldsymbol{\theta}^{(t)})$ in (3.23);
6: **Until** $|f(\boldsymbol{\theta}^{(t)})-f(\boldsymbol{\theta}^{(t-1)})|/|f(\boldsymbol{\theta}^{(t)})|\leq\tau$ or $t\geq t_{\max}$
7: $\boldsymbol{\theta}^{\star}=\boldsymbol{\theta}^{(t)}$.

Therefore, $\theta_m^{\star}$ can be determined one by one until all $N(L+1)$ variables are updated. [3].

After this SCA based EWU method, similarly, the composite channel for user $q$ is obtained by grouping the first $L$ elements of $\boldsymbol{\theta}_n$ into the diagonal matrix $\boldsymbol{\Theta}_n^{\star(i)}=\text{diag}\{\boldsymbol{\theta}_n^{\star(i)}[1:L]\}$ and substituting the $\boldsymbol{\Theta}_n^{\star(i)}$ into the auxiliary channel to get $h_{q,n}=h_{\text{d},q,n}+\mathbf{h}_{r,q,n}\boldsymbol{\Theta}_n^{\star(i)}\mathbf{h}_{\text{i},n}$. [4].

### 3.3.3 Waveform Design for Multi-User FF-IRS and Multi-User FS-IRS

In this section, we aim to maximize the weighted sum current subject to the power constraint at the transmitter for the composite channel $h_{q,n}$ obtained in the previous sections by optimizing the waveform weights at different frequencies. To formulate this subproblem into a compact form, we introduce $N$-by-$N$ matrices $\mathbf{B}_q=\mathbf{h}_q^H\mathbf{h}_q$ with $\mathbf{h}_q=[h_{q,1},h_{q,2},...,h_{q,n}]$, which is shown in Fig. 3.4. $k\in\{1,...,N-1\}$ denotes the index of the $k$th diagonal above the main diagonal (with index $k=0$) of the matrix $\mathbf{B}_q$. $k\in\{-(N-1),...,-1\}$ denotes the index of the $|k|$th diagonal below the main diagonal. $\mathbf{B}_{q,k}$ is obtained by retaining the $k$th diagonal of $\mathbf{B}_q$ with other entries being zeros.

[3]Note that $\theta_m$ is assigned to be 1 when $m$ is an integer multiple of $L+1$.
[4]This SCA-based EWU method can be directly implemented in multi-antenna scenario since the number of variables that need to be optimized is still $N(L+1)$ due to the nature of FS-IRS, the performance with different number of transmit antennas $M_t$ is illustrated in Fig. 3.17.

The compact form of the output current at the $q$th user $i_{\mathrm{dc},q}(\mathbf{s})$ can be recast as

$$i_{\mathrm{dc},q}(\mathbf{s}) = \tfrac{1}{2}k_2\mathbf{s}^H\mathbf{B}_{q,0}\mathbf{s} + \tfrac{3}{8}k_4\mathbf{s}^H\mathbf{B}_{q,0}\mathbf{s}(\mathbf{s}^H\mathbf{B}_{q,0}\mathbf{s})^H + \tfrac{3}{4}k_4\sum_{k=1}^{N-1}\mathbf{s}^H\mathbf{B}_{q,k}\mathbf{s}(\mathbf{s}^H\mathbf{B}_{q,k}\mathbf{s})^H. \quad (3.27)$$

The weighted sum current of all users can be written as $\sum_{q=1}^{K}\xi_q i_{\mathrm{dc},q}(\mathbf{s})$. Denote the auxiliary variable $b_{q,k} = \mathbf{s}^H\mathbf{B}_{q,k}\mathbf{s} = \mathrm{Tr}\{\mathbf{B}_{q,k}\mathbf{s}\mathbf{s}^H\} = \mathrm{Tr}\{\mathbf{B}_{q,k}\mathbf{Y}\}$ for $q = 1,...,K, k = 0,...,N-1$, such that $\mathbf{b}_q = [b_{q,0},...,b_{q,N-1}]^T$. We have $i_{\mathrm{dc}}(\{\mathbf{b}_q\}_{q=1}^K) = \sum_{q=1}^K \xi_q(\frac{1}{2}k_2 b_{q,0} + \mathbf{b}_q^H\mathbf{K}_0\mathbf{b}_q)$. The problem becomes

$$\min_{t_2,\{\mathbf{b}_q\}_{q=1}^K} \quad t_2 \quad (3.28a)$$

$$\text{s.t.} \quad \sum_{q=1}^{K}\xi_q(-\frac{1}{2}k_2 b_{q,0} - \mathbf{b}_q^H\mathbf{K}_0\mathbf{b}_q) - t_2 \le 0, \quad (3.28b)$$

$$\mathrm{Tr}\{\mathbf{B}_{q,k}\mathbf{Y}\} = b_{q,k}, \forall q,k, \quad (3.28c)$$

$$\mathrm{Tr}\{\mathbf{Y}\} \le 2P, \quad (3.28d)$$

$$\mathrm{rank}\{\mathbf{Y}\} = 1. \quad (3.28e)$$

Similar to the approach in Section 3.3.1, we relax the rank constraint (3.28e) and use the SCA to solve it. Let $g(\mathbf{b}_q, \mathbf{b}_q^{(i-1)}) = 2\Re\{\mathbf{b}_q^{(i-1)H}\mathbf{K}_0\mathbf{b}_q\} - \mathbf{b}_q^{(i-1)H}\mathbf{K}_0\mathbf{b}_q^{(i-1)}$. We have $-\frac{1}{2}k_2 b_{q,0} - g(\mathbf{b}_q^{(i)}, \mathbf{b}_q^{(i)}) \le -\frac{1}{2}k_2 b_{q,0} - g(\mathbf{b}_q^{(i)}, \mathbf{b}_q^{(i-1)}) \le t_2$. Therefore, $-\frac{1}{2}k_2 b_{q,0} - g(\mathbf{b}_q, \mathbf{b}_q^{(i-1)}) \le t_2$ can be seen as an upper bound convex constraint of initial non-convex constraint (3.28b). The problem can be rewritten as

$$\min_{t_2,\mathbf{b}_q,\mathbf{Y}\succeq 0} \quad t_2 \quad (3.29a)$$

$$\text{s.t.} \quad -\frac{1}{2}k_2 b_{q,0} - g(\mathbf{b}_q, \mathbf{b}_q^{(i-1)}) - t_2 \le 0, \quad (3.29b)$$

$$(3.28c) \text{ and } (3.28d) \quad (3.29c)$$

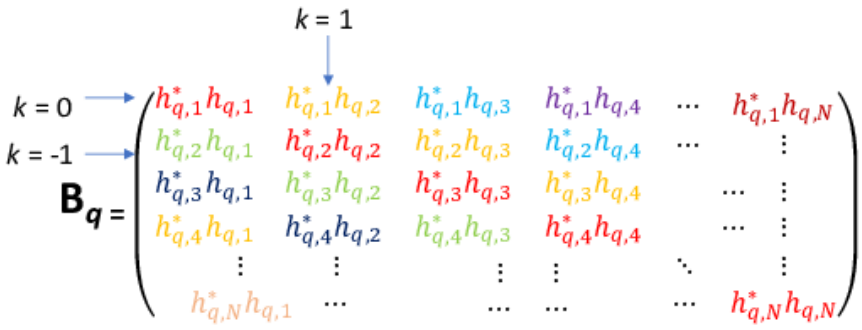



Figure 3.4: The definition of $\mathbf{B}_q$.

which can be written into a SDP as

$$\min_{\mathbf{Y}}\{\mathrm{Tr}\{\mathbf{K}_3\mathbf{Y}\}:\mathrm{Tr}\{\mathbf{Y}\}\leq 2P\} \tag{3.30}$$

where $\mathbf{K}_3=\mathbf{J}_3+\mathbf{J}_3^H$ is a Hermitian matrix and

$$\mathbf{J}_3=\sum_{q=1}^{K}\xi_q\big(-\tfrac{k_2}{4}\mathbf{B}_{q,0}-\tfrac{3}{8}k_4 b_{q,0}^{(i-1)}\mathbf{B}_{q,0}-\tfrac{3}{4}k_4\textstyle\sum_{k=1}^{N-1}b_{q,k}^{(i-1)*}\mathbf{B}_{q,k}\big). \tag{3.31}$$

### 3.3.4 Alternating Optimization

For all the channel trails, we iteratively update the passive beamforming phases and waveform weights until convergence, which is shown in Algorithm 2 and Algorithm 3. In both algorithms, $[\mathbf{V}_{\mathbf{K}_3}]_{\min}$ refers to the eigenvector corresponding to the smallest eigenvalue which can be obtained via the eigenvalue decomposition (EVD) [104].

**Remark 3.** *Interestingly, (3.10), (3.19) and (3.27) are essentially closed-form transformations of same objective formula (3.7) with different variables and this is the key reason that both subproblems of AO can converge to the same output current solutions.*

**Algorithm 2** MU FF-IRS WPT Algorithm

1: **Initialize**: $i=0$, $\{\mathbf{d}_q^{(0)},\mathbf{b}_q^{(0)},\mathbf{B}_q^{(0)},\mathbf{D}_q^{(0)}\}_{q=1}^K,\mathbf{K}_1^{(0)},\mathbf{K}_3^{(0)},\boldsymbol{\Theta}^{(0)}$, $\mathbf{y}^{(0)}$ and corresponding $t_2^{(0)}$;
2: **Repeat**
3: $i=i+1$;
4: Update $\mathbf{K}_1^{(i)}$ and compute $\boldsymbol{\theta}^{\star(i)}$ using Gaussian randomization method in [112], $\boldsymbol{\Theta}^{(i)\star}=\text{diag}\{\boldsymbol{\theta}^{\star(i)}(1:L)\}$;
5: Update $\mathbf{B}_{q,k}^{(i)}$ and $b_{q,k}^{(i)},\forall q,k$;
6: Update $\mathbf{K}_3^{(i)}$, get eigenvector $\mathbf{y}^\star=\sqrt{2P}[\mathbf{V}_{\mathbf{K}_3}]_{\min}$;$\mathbf{Y}^\star=\mathbf{y}^\star\mathbf{y}^{\star H}$;
7: Update $\mathbf{Y}^{(i)}=\mathbf{Y}^\star$. Then update $d_{q,k}^{(i)},\mathbf{D}_{q,k}^{(i)},\forall q,k$, compute $t_2^{(i)}$;
8: **Until** $|t_2^{(i)}-t_2^{(i-1)}|/|t_2^{(i)}|\leq\varepsilon$ or $i\geq i_{\max}$
9: $\mathbf{s}^\star=\mathbf{y}^\star$, output $\boldsymbol{\Theta}^\star$.

**Algorithm 3** MU FS-IRS WPT Algorithm

1: **Initialize**: $i=0$, $\{\mathbf{e}_q^{(0)},\mathbf{b}_q^{(0)},\mathbf{B}_q^{(0)},\mathbf{E}_q^{(0)}\}_{q=1}^K,\mathbf{K}_2^{(0)}$, $\mathbf{K}_3^{(0)}$, $\{\boldsymbol{\Theta}_n^{(0)}\}_{n=1}^N$, $\mathbf{y}^{(0)}$ and corresponding $t_2^{(0)}$;
2: **Repeat**
3: $i=i+1$;
4: Update $\mathbf{K}_2^{(i)}$ and compute $\boldsymbol{\theta}^{*(i)}$ using EWU and generate $\{\boldsymbol{\Theta}_n^{(i)\star}\}_{n=1}^N=\{\text{diag}\{\boldsymbol{\theta}_n^{\star(i)}(1:L)\}\}_{n=1}^N$;
5: Update $\mathbf{B}_{q,k}^{(i)}$ and $b_{q,k}^{(i)},\forall q,k$;
6: Update $\mathbf{K}_3^{(i)}$, get eigenvector $\mathbf{y}^\star=\sqrt{2P}[\mathbf{V}_{\mathbf{K}_3}]_{\min}$;$\mathbf{Y}^\star=\mathbf{y}^\star\mathbf{y}^{\star H}$;
7: Update $\mathbf{Y}^{(i)}=\mathbf{Y}^\star$ and $e_{q,k}^{(i)},\mathbf{E}_{q,k}^{(i)},\forall q,k$ and compute $t_2^{(i)}$;
8: **Until** $|t_2^{(i)}-t_2^{(i-1)}|/|t_2^{(i)}|\leq\varepsilon$ or $i\geq i_{\max}$
9: $\mathbf{s}^\star=\mathbf{y}^\star$, output $\{\boldsymbol{\Theta}_n^\star\}_{n=1}^N$.

### 3.3.5 Convergence

**Proposition 1.** *For any feasible initial points, Algorithm 2 provides a $\boldsymbol{X}^\star$ which satisfies the Karush-Kuhn-Tucker (KKT) conditions of problem (3.15). Note that $\boldsymbol{X}^\star$ is not guaranteed to be a rank-one solution due to the non-convex modulus constraints.*

*Proof*: See Appendix A.1. □

**Proposition 2.** *If the $\boldsymbol{X}^\star$ in Proposition 1 is a rank-one solution, $\boldsymbol{X}^\star$ is also the local optimal solution of problem (3.12). Then, Algorithm 2 converges to a local optimal solution in (3.8).*

*Proof*: See Appendix A.2. □

**Proposition 3.** *For any feasible initial points*[5]*, the passive beamforming vectors* $\boldsymbol{\theta}^{\star}$ *generated by Algorithm 3 in problem (3.21) can guarantee the non-decreasing property of* $\sum_{q=1}^{K}\xi_q i_{dc,q}(\boldsymbol{\theta}^{\star})$*. The Algorithm 3 can converge to a suboptimal point of (3.8).*

*Proof*: The SCA firstly ensures the non-decreasing property as $\boldsymbol{\theta}^{\star(i-1)}$ is also a feasible solution in iteration $i$. In the EWU method, (3.25) and (3.26) double guarantee the non-decreasing procedure by applying the complementary phasor of $\sum_{j=1,j\neq m} k_{m,j}\theta_j^*$ in (3.25). Due to the unit modulus constraint, the SCA based EWU strategy is upper bounded and finally converge. The Algorithm 3 is suboptimal due to the performance loss incurred by the EWU method. This performance loss is demonstrated in Fig. 3.16. □

**Proposition 4.** *For any feasible initial points, the waveform design subproblem in Section 3.3.3 can converge to the stationary points of problem (3.28) with* $\sum_{q=1}^{K}\xi_q(i_{dc,q}(\mathbf{s}^{\star(i-1)}) \leq \sum_{q=1}^{K}\xi_q(i_{dc,q}(\mathbf{s}^{\star(i)})$.

*Proof*: See [17] for detail. □

## 3.4 Low Complexity Design for Single-user Scenario

In this section, we show an extra algorithm for SU FS-IRS for three reasons. First, Algorithm 3 can only provide suboptimal solutions due to the EWU strategy, but Algorithm 4 exhibits global optimal solutions. Second, Algorithm 4 further reduces the complexity compared with Algorithm 3 with $K=1$ without the necessity of AO strategy. Third, Algorithm 3 cannot be reduced to Algorithm 4 when $K=1$ since, in Algorithm 3, perfect alignment between the auxiliary channel and the direct channel

[5]For example, uniform power allocation with random beaforming.

may not be achieved, which incurs a performance loss compared with Algorithm 4. This performance loss is demonstrated in Fig. 3.16.

Because of the issues above, we propose an efficient algorithm for SU FS-IRS by priorly determining the passive beamforming phases and waveform phases. For simpler notation, $h_{\mathrm{d},q,n}$ and $\mathbf{h}_{\mathrm{r},q,n}$ boil down to $h_{\mathrm{d},n}$ and $\mathbf{h}_{\mathrm{r},n}$ respectively. We denote $h_{\mathrm{d},n}=A_{\mathrm{d},n}e^{\phi_{\mathrm{d},n}}$, $h_{\mathrm{i},l+(n-1)L}=A_{\mathrm{i},l+(n-1)L}e^{\phi_{\mathrm{i},l+(n-1)L}}$ and $h_{\mathrm{r},l+(n-1)L}=A_{\mathrm{r},l+(n-1)L}e^{\phi_{\mathrm{r},l+(n-1)L}}$. Then, we denote the entries of $\mathbf{A}_{\mathrm{i},n}$ as $A_{\mathrm{i},l+(n-1)L}$ and the entries of $\mathbf{A}_{\mathrm{r},n}$ as $A_{\mathrm{r},l+(n-1)L}$ with $\mathbf{A}_{\mathrm{i},n}=[A_{\mathrm{i},1+(n-1)L},\ldots,A_{\mathrm{i},L+(n-1)L}]^T$ and $\mathbf{A}_{\mathrm{r},n}=[A_{\mathrm{r},1+(n-1)L},\ldots,A_{\mathrm{r},L+(n-1)L}]$, respectively. The passive beamforming and frequency domain power allocation for SU FS-IRS are sequentially demonstrated in the next sections.

### 3.4.1 Passive Beamforming for Single-User FS-IRS

Since FS-IRS is assumed to have an independent reflection for all frequencies with totally $NL$ variables. When $K=1$, each passive beamforming phase in FS-IRS can idealy align the auxiliary channel to the direct channel and the optimal passive beamforming phases and waveform phases have the values below

$$\gamma_n^\star=-\phi_{\mathrm{d},n},\forall n, \tag{3.32}$$

$$\psi_{l+(n-1)L}^\star=-(\gamma_n^\star+\phi_{\mathrm{i},l+(n-1)L}+\phi_{\mathrm{r},l+(n-1)L}),\forall n,l. \tag{3.33}$$

**Remark 4.** *Recall that in Section 3.3.3, we directly optimize the waveform weight* $\mathrm{s}_n$ *which consists of the amplitude* $w_n$ *and phase* $\gamma_n$*. In contrast, due to the prior determination of* $\gamma_n$*, the joint passive beamforming and waveform design in SU FS-IRS is essentially the frequency domain power allocation at different frequencies, i.e.* $w_n$*, and naturally exhibits lower complexity than SCA based AO strategy in MU FS-IRS.*

### 3.4.2 Frequency Domain Power Allocation for Single-User FS-IRS

Due to the ideal alignment of composite channel in SU FS-IRS, we denote the corresponding composite channel $A_n = A_{\mathrm{d},n} + \mathbf{A}_{r,n}\mathbf{A}_{\mathrm{i},n}$ which is collected into $\mathbf{A} = [A_1,...,A_N]^T$ and group $w_n$ into the power allocation vector $\mathbf{p} = [w_1,w_2,...,w_N]^T$. Objective function (3.7) is transformed to

$$i_{\mathrm{dc}}(\mathbf{p}) = \tfrac{1}{2}k_2\textstyle\sum_{n=1}^{N} w_n^2 A_n^2 + \tfrac{3}{8}k_4 \sum_{\substack{n_1,n_2,n_3,n_4 \\ n_1+n_3=n_2+n_4}} w_{n_1}A_{n_1}A_{n_2}w_{n_2}w_{n_3}A_{n_3}A_{n_4}w_{n_4}. \tag{3.34}$$

We can formulate (3.34) into a more compact form by introducing $N$-by-$N$ matrices $\mathbf{B}$ and $\mathbf{B}_k$ where $\mathbf{B} = \mathbf{A}\mathbf{A}^T$. $k \in \{1,...,N-1\}$ denotes the index of the $k$th diagonal above the main diagonal (with the index $k=0$) of the matrix $\mathbf{B}$. $k \in \{-(N-1),...,-1\}$ denotes the index of the $|k|$th diagonal below the main diagonal. $\mathbf{B}_k$ is obtained by retaining the $k$th diagonal of $\mathbf{B}$ with other entries as zeros. The compact form of (3.34) can be expressed as

$$i_{\mathrm{dc}}(\mathbf{p}) = \tfrac{1}{2}k_2\mathbf{p}^H\mathbf{B}_0\mathbf{p} + \tfrac{3}{8}k_4\mathbf{p}^H\mathbf{B}_0\mathbf{p}(\mathbf{p}^H\mathbf{B}_0\mathbf{p})^H + \tfrac{3}{4}k_4\sum_{k=1}^{N-1}\mathbf{p}^H\mathbf{B}_k\mathbf{p}(\mathbf{p}^H\mathbf{B}_k\mathbf{p})^H. \tag{3.35}$$

To maximize the output DC under the power constraint at transmitter, the problem can be formulated as

$$\max_{\mathbf{p}}\{i_{\mathrm{dc}}(\mathbf{p}) : \|\mathbf{p}\|^2 \le 2P\}. \tag{3.36}$$

Similar to that in Section 3.3.3, we introduce a rank-one positive semidefinite matrix variable $\mathbf{P} = \mathbf{p}\mathbf{p}^H$ to linearize the problem with $b_k = \mathbf{p}^H\mathbf{B}_k\mathbf{p} = \mathrm{Tr}\{\mathbf{B}_k\mathbf{P}\}$. Then, denote $\mathbf{b} = [b_0, b_1, \ldots, b_{N-1}]^T$. Finally, we have $i_{\mathrm{dc}}(\mathbf{b}) = \frac{1}{2}k_2b_0 + \frac{3}{8}k_4b_0b_0^* + \frac{3}{4}k_4\sum_{k=1}^{N-1}b_kb_k^* =$

$\frac{1}{2}k_2 b_0 + \mathbf{b}^H \mathbf{K}_0 \mathbf{b}$. The problem (3.36) can be equivalently written as

$$\min_{t_3, \mathbf{b}, \mathbf{P} \succeq 0} \quad t_3 \tag{3.37a}$$

$$\text{s.t.} \quad -i_{\text{dc}}(\mathbf{b}) - t_3 \leq 0, \tag{3.37b}$$

$$\text{Tr}\{\mathbf{B}_k \mathbf{P}\} = b_k, \forall k, \tag{3.37c}$$

$$\text{Tr}\{\mathbf{P}\} \leq 2P, \tag{3.37d}$$

$$\text{rank}\{\mathbf{P}\} = 1. \tag{3.37e}$$

After the relaxation of rank constraint and SCA approach, the compact form of problem (3.36) can be formulated as

$$\min_{\mathbf{P} \succeq 0} \{\text{Tr}\{\mathbf{K}_4 \mathbf{P}\} : \text{Tr}\{\mathbf{P}\} \leq 2P\} \tag{3.38}$$

where $\mathbf{K}_4 = \mathbf{J}_4 + \mathbf{J}_4^H$ is a Hermitian matrix and $\mathbf{J}_4 = -\frac{k_2}{4}\mathbf{B}_0 - \frac{3}{8}k_4 b_0^{(i-1)}\mathbf{B}_0 - \frac{3}{4}k_4 \sum_{k=1}^{N-1} b_k^{(i-1)*}\mathbf{B}_k$. The whole procedure is summarized in Algorithm 4.

**Remark 5.** *According to [17], problem (3.38) has a rank-one global optimal solution* $\mathbf{P}^\star = \mathbf{p}^\star \mathbf{p}^{\star H}$ *with* $\mathbf{p}^\star = \sqrt{2P}[\mathbf{V}_{\mathbf{K}_4}]_{\min}$*, which is different from that in generalized MU scenario as the EWU based strategy can only provide suboptimal solutions. In each iteration, the rank-one global optimal solution not only guarantees to demonstrate convergence to the stationary points of the relaxed problem (3.38) but also the original problem (3.37).*

## 3.5 Large Scale Analysis of SU FS-IRS

This subsection proposes a brief analysis of the performance of SU FS-IRS in the limit of a large number of REs $L$ and frequencies $N$. We assume that all the channels are uncorrelated. Given the large scale fading for $h_{\text{d},n}, h_{\text{i},l+(n-1)L}$ and $h_{\text{r},l+(n-1)L}$ as

**Algorithm 4** SU FS-IRS WPT Algorithm

1: **Initialize**: $i=0$ and $\mathbf{b}^{(0)}$, $\mathbf{p}^{(0)}$ and $t_3^{(0)}$;
2: **Repeat**
3: $i=i+1$
4: Compute $\mathbf{K}_4=\mathbf{J}_4+\mathbf{J}_4^H$, get eigenvector $\mathbf{p}^\star=\sqrt{2P}[\mathbf{V}_{\mathbf{K}_4}]_{\min}$;$\mathbf{P}^\star=\mathbf{p}^\star\mathbf{p}^{\star H}$;
5: Update $\mathbf{P}^{(i)}=\mathbf{P}^\star$; Update $b_k^{(i)}=\mathrm{Tr}\{\mathbf{B}_k\mathbf{P}^{(i)}\},\forall k$;
6: Update $t_3^{(i)}$;
7: **Until** $|t_3^{(i)}-t_3^{(i-1)}|/|t_3^{(i)}|\leq\varepsilon$ or $i\geq i_{\max}$
8: $s_n^\star=[\mathbf{p}^\star]_{n,1}\cdot h_n^*/|h_n|,\forall n$ .

$\Lambda_{\mathrm{d}}^{1/2}$,$\Lambda_{\mathrm{i}}^{1/2}$ and $\Lambda_{\mathrm{r}}^{1/2}$, respectively, under the perfect alignment in SU FS-IRS, we have

$$\mathcal{E}\{|h_n|\}=\mathcal{E}\left\{\sqrt{|h_{\mathrm{d},n}+\sum_{l=1}^{L}h_{\mathrm{r},l+(n-1)L}\theta_{l+(n-1)L}h_{\mathrm{i},l+(n-1)L}|^2}\right\}\overset{L\to\infty}{\approx}\sqrt{\Lambda_{\mathrm{d}}+\Lambda_{\mathrm{i}}\Lambda_{\mathrm{r}}L^2}. \tag{3.39}$$

According to [17], with uniform power allocation across frequencies, we can approximate the output current as

$$i_{dc}\overset{N\to\infty}{\approx}k_2P(\Lambda_{\mathrm{d}}+\Lambda_{\mathrm{i}}\Lambda_{\mathrm{r}}L^2)+\tfrac{3}{2}k_4P^2(\Lambda_{\mathrm{d}}+\Lambda_{\mathrm{i}}\Lambda_{\mathrm{r}}L^2)^2+3k_4P^2(\Lambda_{\mathrm{d}}+\Lambda_{\mathrm{i}}\Lambda_{\mathrm{r}}L^2)^2N. \tag{3.40}$$

Therefore, if the number of $L$ is large, the output current will scale up with the square number of $L^2$ (i.e. $L^4$) which is shown in (3.40) and validated in Fig. 3.7. This is due to the nonlinearity of the rectenna model which is truncated to $n_0=4$. In addition, if $N$ is large enough, the output DC linearly scales with $N$ which is validated in Section 3.8[6]. The observation above is inline with the scaling law in [16].

## 3.6 Discussion on Multi-Antenna Scenario

The main contribution of this work is to maximize the output DC power by leveraging the frequency diversity gain, the passive beamforming gain and the gain from EH

[6]Based on the rectenna model in Section 2.1.2, increasing the number of subcarriers improves the harvested DC power as long as the EH does not operates in the saturation region. In a practical prototype of No-IRS WPT, due to the Peak-to-Average Power Ratio (PAPR) limits of the transmitter, 16 is a good number which was experimentally observed in [14, 52].

nonlinearity. The active beamforming will not fundamentally change the observation and novelty of this research. To keep the explanation simpler without being overwhelmed by the variables of a multi-antenna scenario, a multi-user SISO transmission model is chosen in our work. However, it is easy to scale to the multi-antenna system with the same AO framework in the next section and the benefits of transmit active beamforming are shown in Fig. 3.17.

## 3.7 Discussion on Discrete-Phase

For discrete IRS problems, one widely used strategy is to relax the discrete IRS problems to their continuous counterparts and obtain quantized phases $\tilde{\psi}_{l+(n-1)L} \forall n,l$ in phase set $\mathcal{S}$ by mapping the continuous values to their closest discrete value in $\mathcal{S}$ with mapping function $\mathcal{F}$ [114]. The mapping function is given below

$$\tilde{\psi}_{l+(n-1)L} = \mathcal{F}(\psi_{l+(n-1)L}), \forall l,n, \tag{3.41}$$

$$\tilde{\psi}_{l+(n-1)L} \in \mathcal{S}, \forall l, \tag{3.42}$$

where $\mathcal{S} = \{0, \Delta\psi, ..., \Delta\psi(2^M - 1)\}$ denotes the equally spaced phase set with $\Delta\psi = 2\pi/2^M$ and $M$ denotes the resolution bits.

## 3.8 Simulation Result and Analysis

We now evaluate the performance in a typical open space WiFi-like environment under a transmit power constraint of 36 dBm at a central frequency of 5.18 GHz with signal bandwidth being fixed to 10MHz as a baseline unless otherwise stated. A uniform linear array (ULA) at the IRS with half wavelength spacing are considered in our

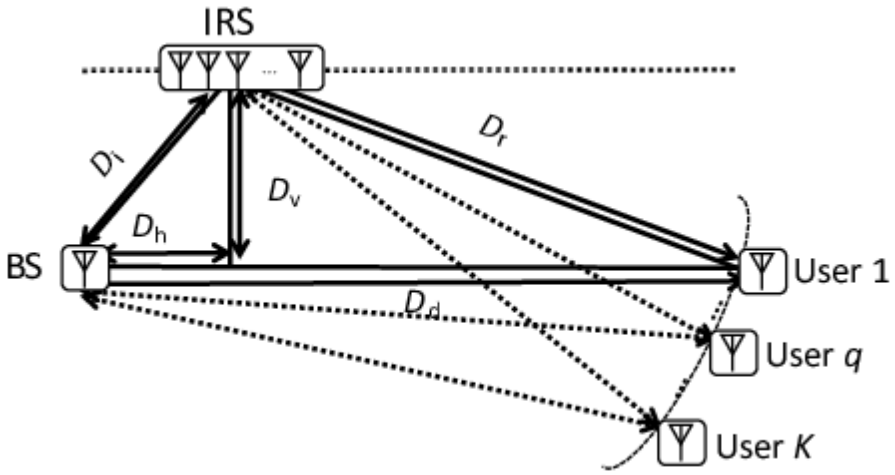



Figure 3.5: Simulation layout.

model. The pathloss model is shown below

$$T_j(D_j) = r_0(\frac{D_j}{d_0})^{-\nu_j} \tag{3.43}$$

where $r_0$ is the large scale fading parameter at reference distance $d_0 = 1$m with $D_j$ and $\nu_j$ referring to the distance and pathloss exponent, respectively, for $j \in \{\text{d,i,r}\}$. All the channels are modeled as Rayleigh fading NLOS channels with the path loss exponents and power delay profiles coming from model D in [115]. 18 taps are modeled as i.i.d CSCG random variables to generate uncorrelated frequency-selective fading channels. For simplicity, all users experience the same distances of direct channel ($D_\text{d}$), incident channel ($D_\text{i}$) and reflected channel ($D_\text{r}$) with 0dB transmit/receive antenna gain on isotropic radiation pattern. Accordingly, $K$ users experience the same large scale fading but different independent small scale fading in power signal transmission. According to Fig. 3.5, $D_\text{i} = \sqrt{D_\text{h}^2 + D_\text{v}^2}$ and $D_\text{r} = \sqrt{D_\text{v}^2 + (D_\text{d} - D_\text{h})^2}$ with $D_\text{h}$ and $D_\text{v}$ referring horizontal distance and vertical distance, respectively. Parameters are assigned as $D_\text{v} = 2$m, $D_\text{h} = 2$m and $D_\text{d} = 15$m as a baseline unless specified later. For all the numerical figures, SU FS-IRS and MU FS-IRS simulation results come

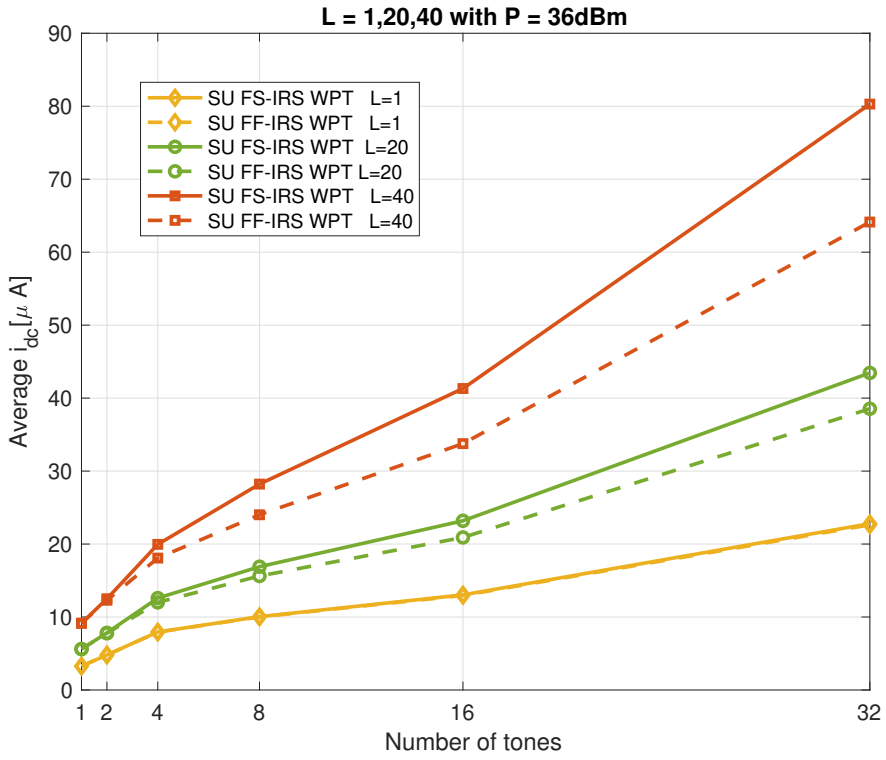



Figure 3.6: Average $i_{dc}$ versus $N$ for $L=1,20,40$.

from Algorithm 4 and 3, respectively. Both SU FF-IRS and MU FF-IRS results are generated from Algorithm 2 for choices of $K=1$ and $K>1$, respectively. The sufficient small tolerance of stopping threshold in all algorithms is set as $10^{-4}$ and each point in the following figures is acquired via averaging over 1000 independent realizations. The number of candidates for Gaussian randomization in Algorithm 2 is 1000.

### 3.8.1 Single-User

We first characterize the average output DC versus the number of sinewaves $N$ with different numbers of passive reflecting elements $L$ in Fig. 3.6 and 3.7. A *first* observation is that the output DC of both SU FS-IRS and SU FF-IRS increases with the number of sinewaves $N$ and SU FS-IRS can be seen as a performance upper bound

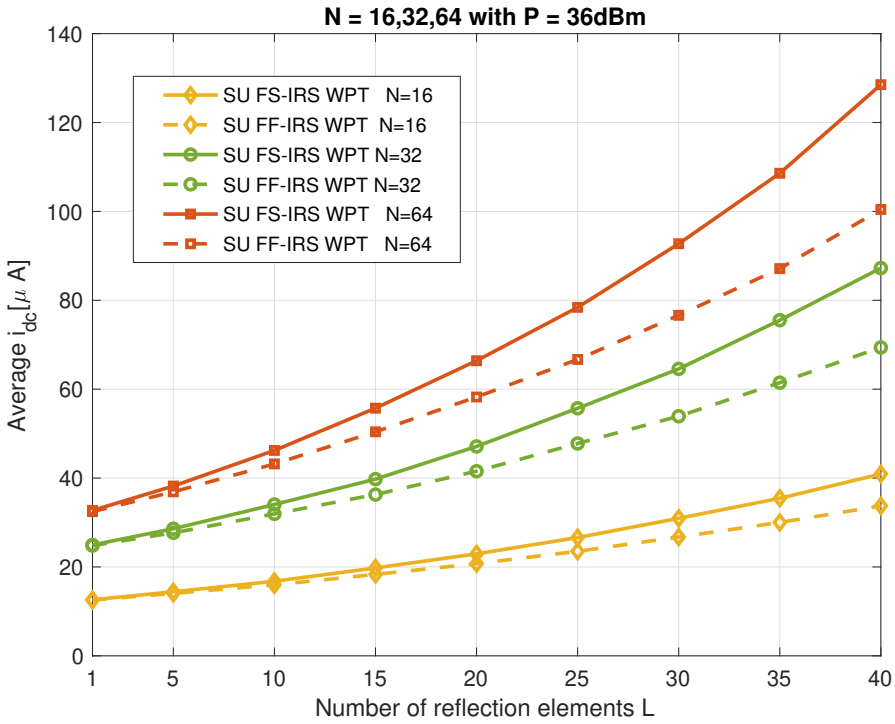



Figure 3.7: Average $i_{dc}$ versus $L$ for $N=16$, 32, 64.

of the corresponding SU FF-IRS scenario. When $N=1$, SU FF-IRS displays the same current as SU FS-IRS. A *second* observation is that the output current approximately scales up with the $L^4$. This is, thus, envisioned to revolutionize large-scale design since a large $L$ can compensate for the small number of transmit antennas. A *third* observation is that the performance gap between SU FS-IRS and SU FF-IRS is gradually increased with $N$ and SU FS-IRS observes a gain of 28% over SU FF-IRS when $N=64$ and $L=40$. The reason behind such a phenomenon is that, in SU FS-IRS, ideal passive beamforming phases can always align with the composite channel. While, in SU FF-IRS, more performance loss is incurred by the misalignment between the auxiliary channel and the direct channel with a larger $N$.

An overview of average output DC versus the number of sinewaves $N$ is exhibited in Fig. 3.8. SU FS-IRS, SU FF-IRS and SU Rand-Phase-IRS WPT refer to nonlinear-based

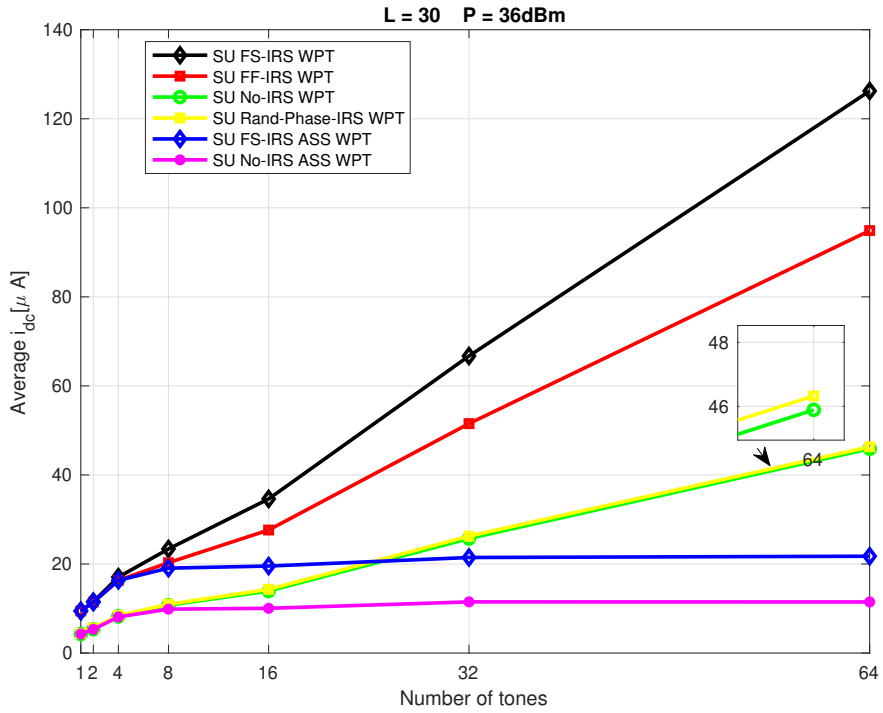



Figure 3.8: Average $i_{dc}$ versus $N$.

rectenna model with different IRS scenarios. SU No-IRS WPT refers to conventional WPT designs in [16]. SU FS-IRS ASS and SU No-IRS ASS denote the ASS-based linear rectenna model (a truncating order of 2) [16] [17] with FS-IRS and without IRS, respectively[7].

We make the following observations.

*First*, if taking SU FS-IRS ASS WPT as a benchmark, SU FS-IRS and SU FF-IRS can exceed the benchmark on around 550% and 400%, respectively, when $N=64$. This result strongly embodies the superiority of nonlinear based rectenna model over ASS based linear rectenna model, showing that the output DC can be effectively boosted by leveraging the gain from rectenna nonlinearity for both FS-IRS and FF-IRS scenarios.

[7]The output current for user $q$ under a linear based model (a truncating order of 2) is given as $i_{\mathrm{dc},q}(\mathbf{s},\{\boldsymbol{\Theta}_n\}_{n=1}^N)=\frac{1}{2}k_2\sum_{n=1}^N s_n^* h_{q,n}^* h_{q,n} s_n$. Compared with (3.7) , the maximization of output current can be efficiently solved by allocating all the power to the strongest frequency-selective channel which leads to the ASS strategy in Fig. 3.8.

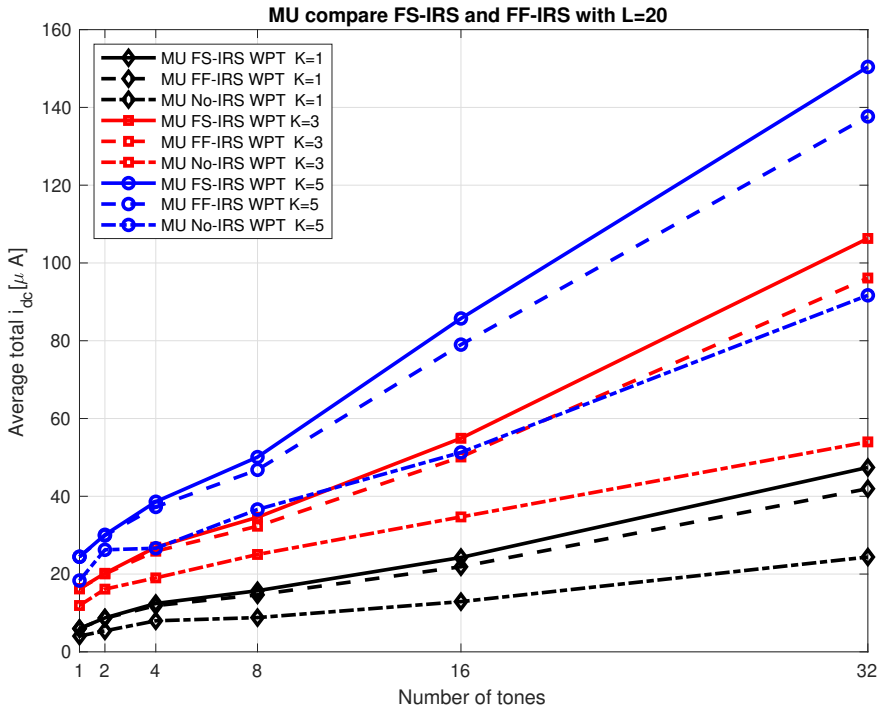



Figure 3.9: The average weighted sum of $i_{\mathrm{dc}}$ versus $N$ for $K=1$, 3, 5 and $L=20$.

*Second*, both SU FS-IRS and SU FF-IRS observe a great performance gain over No-IRS scenario on around 195% and 155%, respectively, when $N=64$ and $L=30$. An explanation is that both of them consolidate the composite channel strength with passive beamforming phases manipulation. Furthermore, it is worth pointing out that although SU FF-IRS is outperformed by SU FS-IRS, it still achieves non-negligible performance improvement over No-IRS conventional WPT waveform designs in [16] [17].

*Third*, SU Rand-Phase-IRS displays a similar output current with the No-IRS scenario which emphasizes the significance of the passive beamforming design. Without the optimization of the passive beamforming phases, the signal attenuation can not be effectively alleviated by IRS as different cascaded channels superpose randomly instead of constructively.

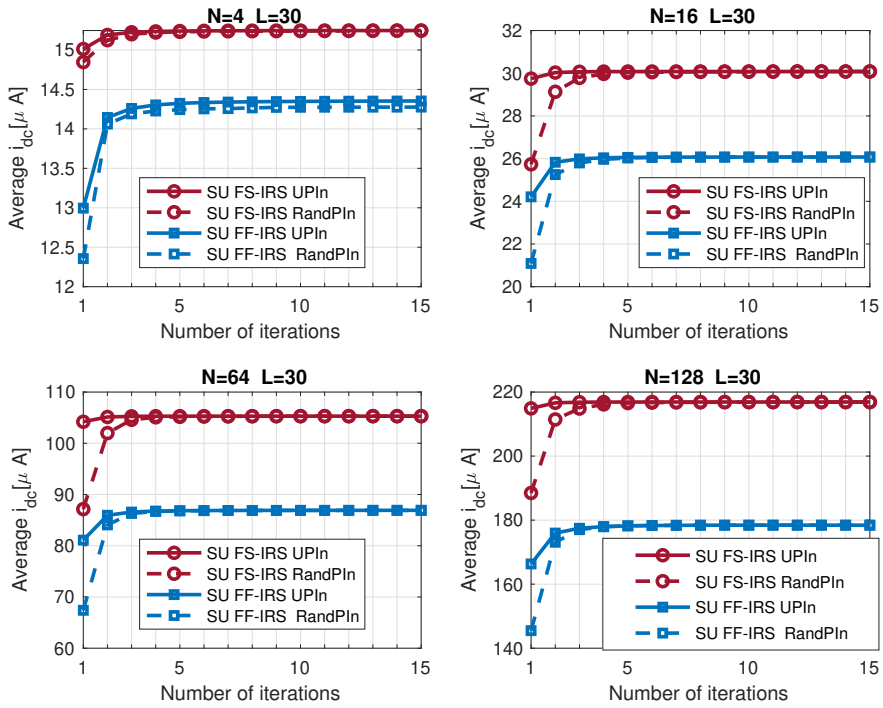



Figure 3.10: Average $i_{\mathrm{dc}}$ versus number of iterations for $N=4$, 16, 64, 128 and $L=30$.

### 3.8.2 Convergence Analysis

Fig. 3.10 and 3.11 illustrate the average output DC versus the number of iterations for different initialization strategies. The initial passive beamforming phases are randomly chosen in [0,2$\pi$). The frequency domain power allocation strategies are "UPIn" and "RandPIn" referring to uniform power allocation and random power allocation, respectively. First, "UPIn" exhibits a faster convergence than "RandPIn", which suggests the potential benefits of uniform power allocation in large-scale SU FS-IRS and SU FF-IRS WPT. Second, both "UPIn" and "RandPIn" converge to nearly the same objective function values. An explanation is that considering the non-convexity in problem (3.12), various initial points may lead to different solutions in Algorithm 2. However, solution $\mathbf{X}^{\star}$ is demonstrated to be a rank-one solution in all tested channels and strongly reduces the performance loss incurred by Gaussian randomization method, which contributes

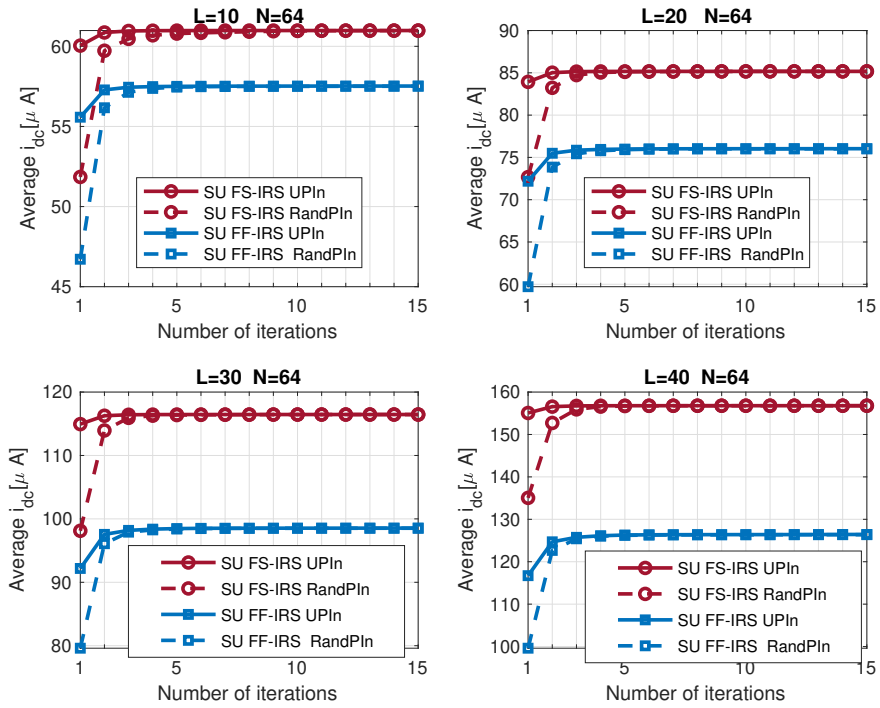



Figure 3.11: Average $i_{\mathrm{dc}}$ versus number of iterations for $L=10$, 20, 30, 40 and $N=64$.

to the insensitivity of initial power allocation and initial passive beamforming phases.

### 3.8.3 Bandwidth

Fig. 3.12 and Fig. 3.13 explore the sensitivity of $i_{\mathrm{dc}}$ to different bandwidths from 1MHz to 10 MHz for $L=20$ and $L=30$, respectively. A *first* observation is that the gap between SU FS-IRS and SU FF-IRS in both figures increase with the bandwidth which emphasizes the superiority of SU FS-IRS over SU FF-IRS on output DC by favoring the strongest sinewaves more efficiently. A *second* observation is that the output DC of SU FS-IRS increases slowly with the bandwidth and the opposite behavior is observed in SU FF-IRS. An explanation is that when $L=20$, both SU FS-IRS and SU FF-IRS benefit from increasing frequency selectivity which contributes to an ascending trend. However, when $L=30$, the misalignment of the composite channel in SU FF-IRS becomes more

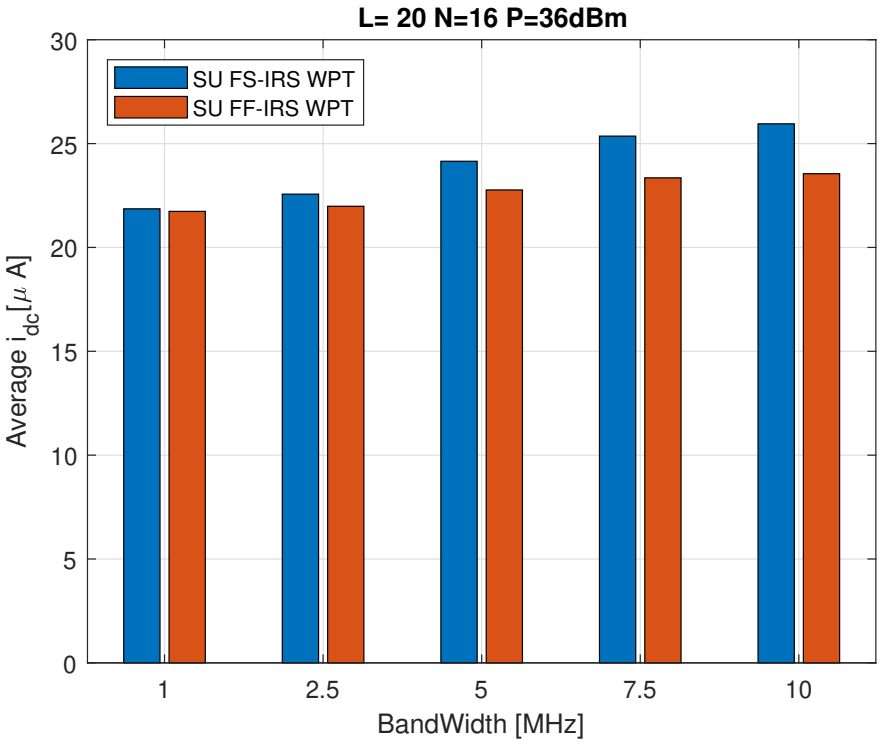



Figure 3.12: Average $i_{\mathrm{dc}}$ versus bandwidth for $L=20$ and $N=16$.

pronounced with increasing bandwidth, which renders a descending trend.

To conclude, SU FF-IRS is not equally suitable for broadband and narrowband WPT, namely, SU FF-IRS works well in narrowband systems but incurs some loss in a higher bandwidth transmission. In contrast, SU FS-IRS offers a promising gain over SU FF-IRS by flexibly favoring passive beamforming phases in a frequency-selective design but also incurs a higher hardware complexity.

### 3.8.4 Multi-User

The average weighted sum current versus the number of sinewaves $N$ for the different number of users $K$ is shown in Fig. 3.9. The weight $\xi_q$ for each user is 1. It is first observed that MU FF-IRS incurs a performance loss compared with MU FS-IRS and

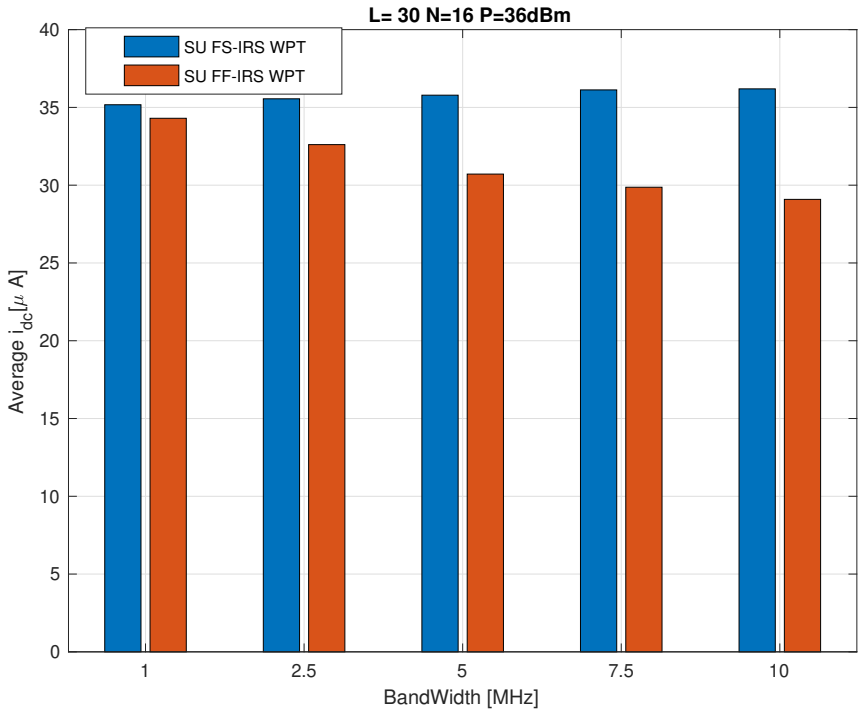



Figure 3.13: Average $i_{\mathrm{dc}}$ versus bandwidth for $L=30$ and $N=16$.

the performance gap increases with $K$. This is expected since MU FF-IRS suffers more misalignment on different user's composite channels than MU FS-IRS by constraining the passive beamforming phases to be the same across frequencies, which stresses the natural benefits of developing more flexible IRS hardware to enable adaptive frequency-selective passive beamforming phase shifts.

Fig. 3.14 draws a useful insight into the average $i_{\mathrm{dc}}$ region (obtained by solving problem (3.8) with varying weights $(\xi_1,\xi_2)$) versus the number of reflecting elements $L$ in a two-user system. It is observed that the current regions of the two IRS strategies gradually increase with $L$. This is intuitive since both FS-IRS and FF-IRS can compensate for the distance-dependent path loss by strengthening the composite channel of each user. Due to the additional design flexibility in the frequency domain, FS-IRS provides better channel alignment and enlarges the region more than FF-IRS. This

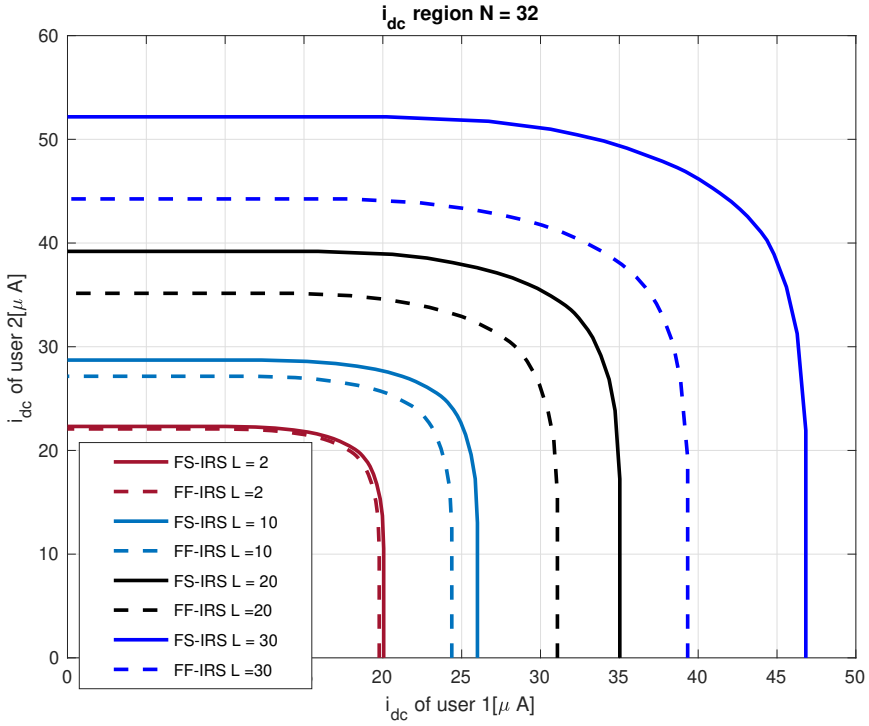



Figure 3.14: $i_{\mathrm{dc}}$ region versus $L$ for $N=32$.

characteristic highlights the advancement of IRS in frequency-selective design. To further unveil the impact of the IRS position, based on the same user weight pairs as Fig. 3.14, Fig. 3.15 demonstrates the current region versus the horizontal distances, i.e. $D_{\mathrm{h}}$, with $D_{\mathrm{v}}=2$m as a constant. One can observe that the regions are effectively enlarged when IRS is installed either near the BS or near the users. The comparison between Algorithm 3 and Algorithm 4 with $K=1$ is demonstrated in Fig. 3.16. Due to the suboptimal operation of the EWU strategy, Algorithm 3 exhibited a slight loss in contrast with Algorithm 4. However, this loss is small enough to be negligible (less than 0.1%) which, instead, reveals the generality and superiority of Algorithm 3.

The impact of joint waveform, active and passive beamforming is demonstrated in Fig. 3.17 which explores the average weighted sum current versus the number of sinewaves $N$ for the different number of transmit antenna $M_t$. It is observed that the weighted

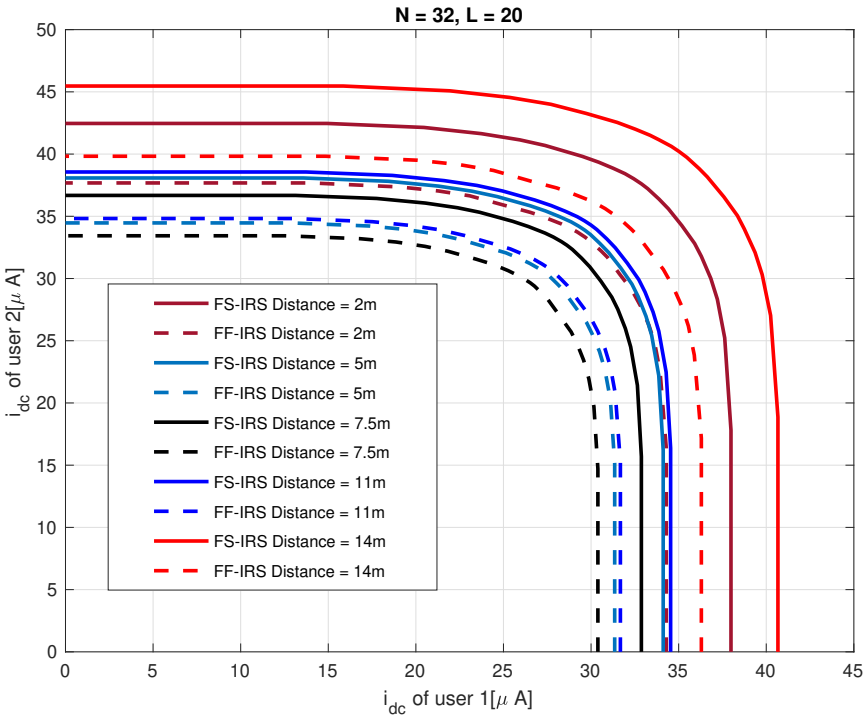



Figure 3.15: $i_{\mathrm{dc}}$ region versus $D_{\mathrm{h}}$ for $N=32$ and $L=20$.

sum output current of both MU FF-IRS and MU FS-IRS approximately scale up with the square number of $M_t$. Compared with the single-antenna scenario in Fig. 3.7, active beamforming is envisioned to further boost the output DC power.

### 3.8.5 Discrete-Phase

To extend to the discrete-phase IRS problem, one widely used strategy is to relax the discrete IRS problems to their continuous counterparts and obtain quantized phases $\tilde{\psi}_{l+(n-1)L}\forall n,l$ in phase set $\mathcal{S}$ by mapping the continuous values to their closest discrete value in $\mathcal{S}$ with mapping function $\mathcal{F}$ [114] which is given as

$$\tilde{\psi}_{l+(n-1)L}=\mathcal{F}(\psi_{l+(n-1)L}),\forall l,n, \tag{3.44}$$

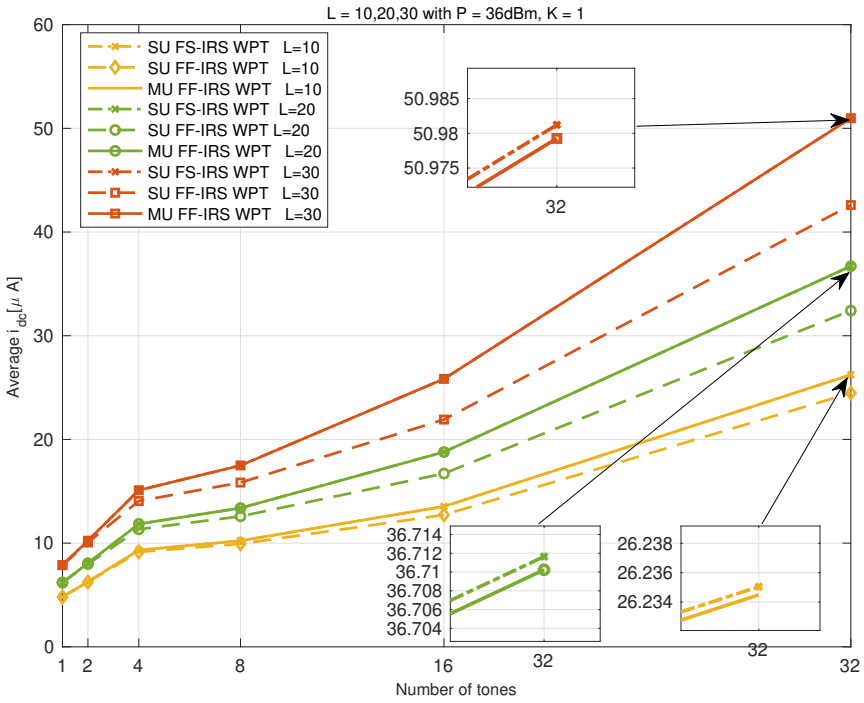



Figure 3.16: Average $i_{\mathrm{dc}}$ versus $N$ for $L=10$, 20, 30 and $K=1$.

$$\tilde{\psi}_{l+(n-1)L} \in \mathcal{S}, \forall l, n, \tag{3.45}$$

where $\mathcal{S}=\{0,\Delta\psi,...,\Delta\psi(2^M-1)\}$ denotes the equally spaced phase set with $\Delta\psi=2\pi/2^M$ and $M$ denotes the resolution bits. Leveraging this discrete-phase strategy in [114], Fig. 3.18 and 3.19 are plotted for single-user and multi-user conditions, respectively.

**Remark 6.** *Note that the rounding method used in this thesis is not the only way to address the discrete-phase problems. [116] proposes an optimization way using the SNR distribution and scaling laws in the presence of channel fading and phase noise but with high computation complexity. The complexity is greatly reduced by selecting scaling laws in the presence of channel fading and phase noise in [117]. The complexity can be further reduced with graph neural network in [118].*

Fig. 3.18 demonstrates the average output DC versus $L$ with $N=32$ and quantization

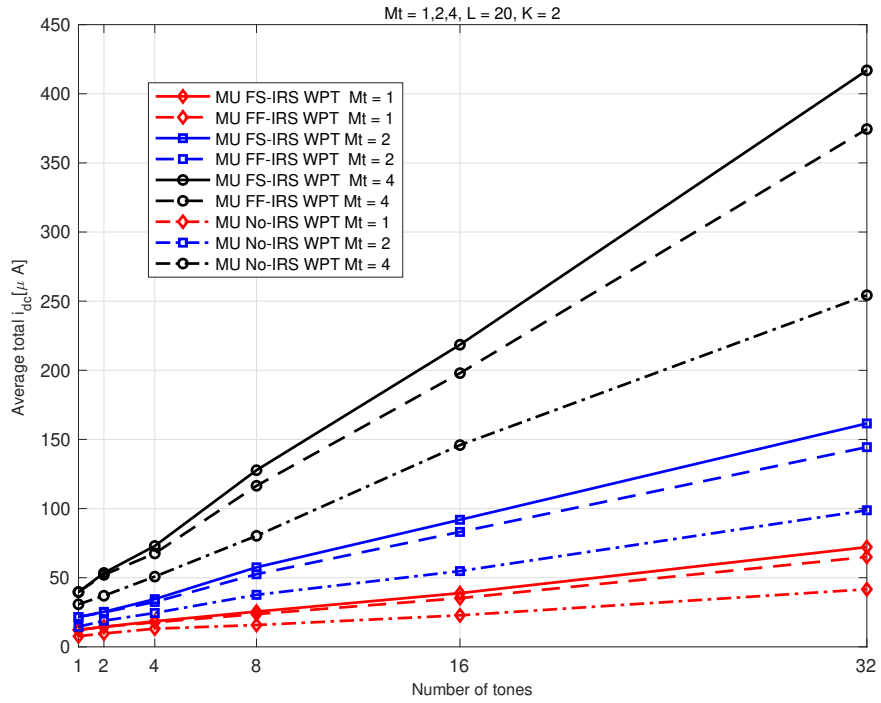



Figure 3.17: Average weighted sum $i_{\mathrm{dc}}$ versus $N$ for $L = 20$, number of transmit antennas $M_t = 1$, 2, 4 and $K = 2$.

scheme $M = 1,2,3$ bits for single-user scenario. *First*, it is observed that the $i_{\mathrm{dc}}$ for both SU FF-IRS and SU FS-IRS can be greatly improved compared with the No-IRS condition even with a 1-bit phase shifter. *Second*, one can observe that the performance loss incurred by discrete phases decreases with an increasing number of resolution bits. This is intuitive since a larger number of bits enables a better alignment between channels. Under the same quantization scheme, the right figure illustrates the weighted sum current versus $N$ with $L = 20$ and $K = 3$. The performance gain of the 1-bit resolution current over the No-IRS benchmark increases with the number of subcarriers $N$, which suggests the effectiveness of deploying discrete-phase IRS in wideband WPT. Moreover, the performance loss incurred by discrete-IRS is low with 3-bit resolution phase shifters, which confirms that directly quantizing the optimized continuous RE phases can achieve near-optimal performance in a multi-user

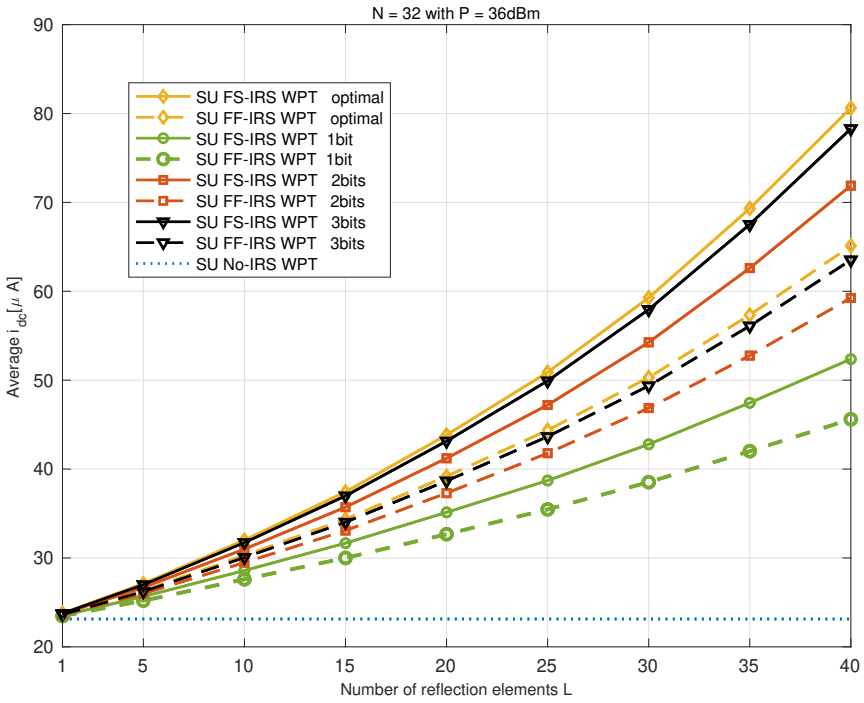



Figure 3.18: Average $i_{\mathrm{dc}}$ versus $L$ for $N=32$ and resolution bits $M=1, 2, 3$.

scenario.

## 3.9 Summary

In this work, we investigated the joint waveform and passive beamforming design for IRS-aided wireless power transfer for both single-user and multi-user deployments. Two different formulations were developed based on FS-IRS and FF-IRS. An optimization framework based on AO and SCA was studied and demonstrated robustness with fast convergence. Numerical results confirmed the inefficiency of linear-based model and highlighted the fact that even FF-IRS-aided WPT can outperform conventional WPT designs with IRS passive beamforming gain. Furthermore, FS-IRS acted as a canonical upper bound for the broadband system while FF-IRS was suitable for

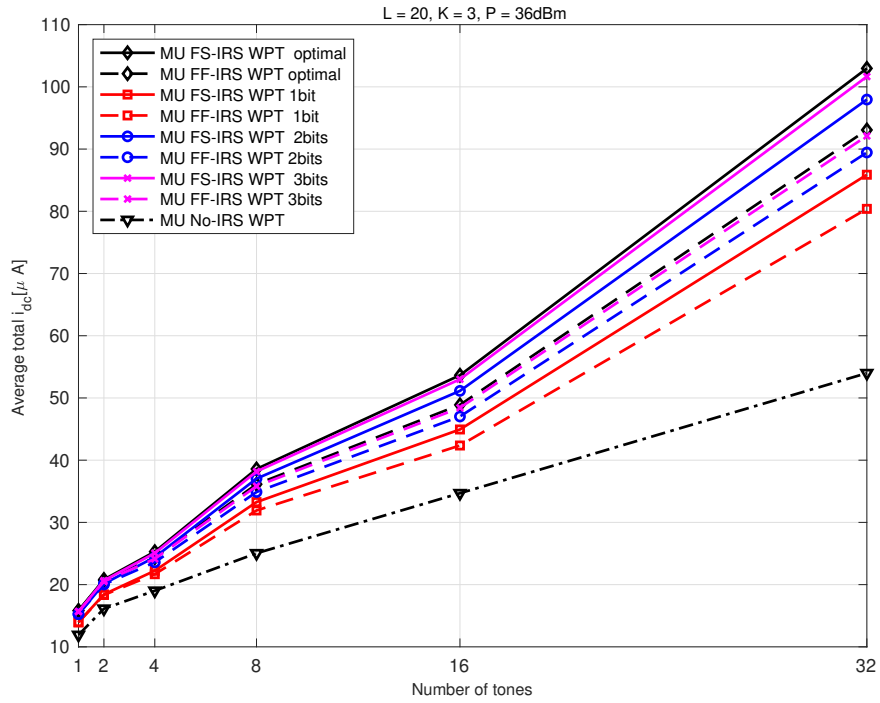



Figure 3.19: Average weighted sum $i_{\mathrm{dc}}$ versus $N$ for $M=1$, 2, 3, $L=20$ and $K=3$.

narrowband transmission. Moreover, for WPT system, the assistance of IRS not only expanded the operation range for SU conditions but also enlarged the output DC region in MU scenarios. In addition, this AO framework can be directly extended to the multi-antenna scenarios to explore the transmit beamforming gain. Last but not least, a near-optimal result can be achieved by leveraging low-resolution discrete-phase IRS for both SU and MU scenarios.

# Chapter 4

# A Novel Multi-agent DQN-based Beamforming Design for Massive MIMO Wireless Communication Networks

This chapter is concerned with the design of a multi-agent DRL framework for massive MIMO with imperfect CSIT, where the transmit and receive beamforming are jointly selected to maximize the average information rate of all users. Leveraging this DRL-based framework, interference management is explored and three DRL-based schemes, namely the DDRL scheme, PDRL, and CDRL scheme, are proposed and analyzed. This work 1) highlights the fact that the DRL-based strategies outperform the random action-chosen strategy and the delay-sensitive strategy named as sample-and-hold (SAH) approach, and achieved over 90% of the information rate of two selected benchmarks with lower complexity: the zero-forcing channel-inversion (ZF-CI) with perfect CSIT and the Greedy Beam Selection strategy, 2) demonstrates the inherent robustness of the proposed designs in the presence of channel aging. 3)

conducts detailed convergence and scalability analysis on the proposed framework.

## 4.1 Introduction

Due to the increasing demand for data and connectivity in 5G [2] 6G [119], multi-antenna technologies have attracted great attention in academia and industry. The research on multi-antenna techniques has promoted the development of MIMO technology. MIMO nowadays plays an indispensable role in the physical layer, media access control (MAC) layer, and network layer in wireless communications and networking [120]. At the physical layer, multi-antenna beamforming strategies have attracted great interest due to their ability to achieve considerable antenna gains, multiplexing gains, and diversity gains [4, 121], and gradually evolved into a massive MIMO system, in which the number of antennas at the BS reaches tens or even hundreds, serving a larger number of users. To enable a high throughput in the massive MIMO system, the BS relies on the huge demand for global and instantaneous CSI based on efficient channel estimation techniques [81, 82]. Nevertheless, the ground/air/space platforms such as high-speed trains/UAV/satellites have a common characteristic of 3D mobility which leads to a stringent time constraint on CSI acquisition and even causes misalignment of narrow beams. Therefore, in future communication systems, how to maintain good connectivity and system capacity without perfect CSIT is regarded as an important problem that yearns for prompt solutions.

Motivated by the benefits of DRL, in this chapter, we study the joint transmit precoder and receive combiner design in massive MIMO downlink transmission with channel aging. The contributions of this work are summarized as follows.

- We construct an efficient multi-agent DRL-based framework for massive MIMO downlink transmission[1], in light of which three DRL-based algorithms were

[1]The terminology massive MIMO in this work implies multiple receive antennas.

derived based on stream-level, user-level, and system-level agent modeling. This is the first work showing that the DRL-based framework can be used to address very high-dimension optimization problems and demonstrates 1) robustness on the degrading effect of channel aging, 2) stringent interference management especially the inter-stream interference and multi-user interference.

- To address the challenge of high-dimensional antenna beamforming problems, by utilizing the DRL-based framework, three DRL-based schemes, namely distributed-learning-distributed-processing DRL-based scheme (DDRL), partial-distributed-learning-distributed-processing DRL-based scheme (PDRL), and central-learning-distributed-processing DRL-based scheme (CDRL), are proposed, analyzed, and evaluated. For DDRL, each stream is modeled as an agent. All the agents save their experiences in a private experience pool for later training. In contrast, in CDRL, the whole system is modeled as a central agent. What's more, to bridge DDRL and CDRL, we demonstrate another algorithm, i.e., PDRL which offers a more flexible design by modeling each user as an agent to balance the performance and complexity. Note that the DDRL and CDRL are different from those in [18, 19] since we are tackling the problem with 1) receive beamforming with multiple receive antennas, 2) transmit beamforming with imperfect CSIT, 3) multiple streams for each user, and 4) a large number of transmit antennas at BS compared with SISO in [18] and 4-antenna MISO in [19, 25], respectively.

- Leveraging the DRL-based framework mentioned above, the precoders at BS and combiners at users are jointly designed by gradually maximizing the average information rate through the observed reward. In particular, the BS decides the transmit precoder and receive combiner for each stream with imperfect CSIT and perfect CSIR. The merits of this design are shown through extensive simulations by benchmarking our schemes against the conventional, sample-and-hold (SAH) approach [99], zero-forcing channel-inversion (ZF-CI) strategy [121], greedy

beam selection and random action-chosen scheme.

- We demonstrate the advantages of DRL-based strategies over the benchmarks above. In particular, the proposed algorithms show 1) fast convergence to efficient beamforming policy, 2) the robustness on tracing the channel dynamic against channel uncertainty due to channel aging, and 3) lower complexity compared with traditional beamforming strategy. All of these properties are essential applications of the proposed techniques in practical wireless networks.

- By numerical results, we show that our proposed DRL-based schemes outperform the SAH approach and random action-chosen scheme. In particular, DDRL can achieve nearly 90% of the performance of the state-of-the-art ZF-CI method with perfect CSI (ZF-CI PCSI) and 95% of the performance of the Greedy Beam Selection method but incurs more hardware complexity and more uplink overhead in an FDD setup. By increasing the resolution of the codebook and hyper-parameter tuning on the reward function, the performance can be further improved.

## 4.2 System Model

Consider the MIMO broadcast channel (BC) with one $M$-antenna BS and $K$ $N$-antenna users indexed by $\mathcal{K}=\{1,...,K\}$ [67] . The BS aims to deliver $M_s$ streams in the time instant of interest. For simplicity, a number of $KN_s$ streams are transmitted simultaneously from the $M$ antennas of the BS. Each group of $N_s$ streams indexed by $n\in\mathcal{N}=\{1,...,N_s\}$ is targeted at one of the $K$ users. Note that we consider a setting where $M \geq KN_s$ to ensure the spatial multiplexing gain. The transmit power $P$ is uniformly allocated to all $KN_s$ streams[2]. We assume that the BS and all users operate in the same time-frequency resource and are synchronized. The transmitted signal,

[2]To the authors' best knowledge, there is no paper addressing the power allocation problem with imperfect channel state information. This problem is not in the scope of this thesis but can be considered in the future research

i.e., the precoded data vector, at time slot $t$ can be written as

$$\mathbf{x}(t)=\sqrt{\frac{P}{KN_s}}\sum_{k=1}^{K}\sum_{n=1}^{N_s}\mathbf{p}_{k,n}(t)s_{k,n}(t) \tag{4.1}$$

where $s_{k,n},\forall k\in\mathcal{K},\forall n\in\mathcal{N}$, is the encoded message from message $W_{k,n}$ with zero mean and $\mathcal{E}(|s_{k,n}|^2)=1$, and precoder $\mathbf{p}_{k,n}(t)\in\mathbb{C}^{M\times 1}$ is subject to $\|\mathbf{p}_{k,n}(t)\|^2=1$. The received signal at user $k$ can be expressed as

$$\mathbf{y}_k(t)=\sqrt{\tfrac{P}{KN_s}}\mathbf{H}_k(t)\textstyle\sum_{n=1}^{N_s}\mathbf{p}_{k,n}(t)s_{k,n}(t)+\sqrt{\tfrac{P}{KN_s}}\mathbf{H}_k(t)\sum_{j\neq k,j=1}^{K}\sum_{i=1}^{N_s}\mathbf{p}_{j,i}(t)s_{j,i}(t)+\mathbf{n}_k(t) \tag{4.2}$$

where the noise vector $\mathbf{n}_k\in\mathbb{C}^{N\times 1}$ is assumed to follow a complex normal distribution, i.e., $\mathbf{n}_k\in\mathcal{CN}(\mathbf{0},\sigma_n^2\mathbf{I}_N)$. At the user side, the combiner vector for each stream is denoted as $\mathbf{w}_{k,n}(t)\in\mathbb{C}^{N\times 1}$, $\|\mathbf{w}_{k,n}(t)\|^2=1,\forall k\in\mathcal{K},n\in\mathcal{N}$. Then, the achievable rate for user $k$ and the average user rate at time slot $t$ can be written as

$$R_k(t)=\sum_{n=1}^{N_s}G_{k,n}(t),\quad \bar{R}(t)=\frac{\sum_{k=1}^{K}R_k(t)}{K}, \tag{4.3}$$

where $G_{k,n}$ is the achievable rate of stream $n$ for user $k$. To indicate the downlink information rate in each stream, by adopting the Shannon capacity equation, $G_{k,n}$ is given as

$$G_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)))=\log(1+\gamma_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))) \tag{4.4}$$

where, for consistency with the notation in the following sections, $\gamma_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$ denotes the Signal-to-Interference-plus-Noise Ratio (SINR) of stream $n$ for user $k$ as

$$\gamma_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))=\frac{\frac{P}{KN_s}|\mathbf{w}_{k,n}^H(t)\mathbf{H}_k(t)\mathbf{p}_{k,n}(t)|^2}{I_{k,n}(t)+I_{c,k}(t)+\|\mathbf{w}_{k,n}(t)\|^2\sigma_n^2(t)} \tag{4.5}$$

where $\mathbf{W}_k(t)=[\mathbf{w}_{k,1}(t),\ldots,\mathbf{w}_{k,N_s}(t)]$ denotes the combining matrix and $\mathbf{P}_k(t)=[\mathbf{p}_{k,1}(t),\ldots,\mathbf{p}_{k,N_s}(t)]$ denotes the precoding matrix. The inter-stream interference for

stream $n$ of user $k$ and the multi-user interference for stream $n$ of user $k$ are shown as

$$I_{k,n}(t) = \sum_{i=1, i\neq n}^{N_s} \frac{P}{KN_s} |\mathbf{w}_{k,n}^H(t)\mathbf{H}_k(t)\mathbf{p}_{k,i}(t)|^2 \tag{4.6}$$

and

$$I_{c,k}(t) = \sum_{j\in\mathcal{K}, j\neq k} \sum_{i=1}^{N_s} \frac{P}{KN_s} |\mathbf{w}_{k,n}^H(t)\mathbf{H}_k(t)\mathbf{p}_{j,i}(t)|^2, \tag{4.7}$$

respectively.

### 4.2.1 Channel Model

We assume an extended Saleh-Valenzuela geometric model [122]. The channel between BS and user $k$ is modeled as a $L$-path channel as is shown below

$$\mathbf{H}_k(t) = \sqrt{\frac{\eta_k MN}{L}} \cdot \sum_{l=1}^{L} \alpha_{k,l}(t) \cdot \mathbf{u}_{k,l}(t)\mathbf{v}_{k,l}^H(t) \tag{4.8}$$

where $\eta_k$ denotes the large-scale fading coefficient and complex gain $\alpha_{k,l} (\forall k \in \mathcal{K}, \forall l \in \{1,2,...,L\})$ is assumed to remain the same at each time slot and varies between adjacent time slots according to the first-order Gaussian-Markov process

$$\alpha_{k,l}(t) = \rho\alpha_{k,l}(t-1) + \sqrt{1-\rho^2} e_{k,l}(t) \tag{4.9}$$

where $e_{k,l}(t) \backsim \mathcal{CN}(0,1)$ and $\rho$ is the time correlation coefficient obeying Jakes' model [123].

$$\rho = J_o(2\pi f_\mathrm{d} \Delta_t \cos\theta) \tag{4.10}$$

where $f_\mathrm{d}$ and $\Delta_t$ denote the Doppler Frequency and the channel instantiation interval, respectively, and $J_0$ denotes the first kind $0^{th}$ Bessel function. Since the users are assumed to move forward to the BS or away, i.e., $\theta = 0$ and maximum Doppler frequency

$f_{\mathrm{d}}^{\max}$ is achieved which is written as

$$\rho = J_o(2\pi f_{\mathrm{d}}^{\max}\Delta_t). \tag{4.11}$$

In the typical case of a ULA where the antennas are deployed at both ends of the transmission, the array steering vectors $\mathbf{u}_{k,l}$ and $\mathbf{v}_{k,l}$ corresponding to the angle of arrival (AoA) $\phi_{\mathrm{A},k,l}$ and the angle of departure (AoD) $\phi_{\mathrm{D},k,l}$ in the azimuth are written as

$$\mathbf{u}_{k,l} = \frac{1}{\sqrt{N}}[1, e^{j2\pi\frac{d}{\lambda}\cos\phi_{\mathrm{A},k,l}}, \ldots, e^{j2\pi\frac{d}{\lambda}(N-1)\cos\phi_{\mathrm{A},k,l}}]^T \tag{4.12}$$

and

$$\mathbf{v}_{k,l} = \frac{1}{\sqrt{M}}[1, e^{j2\pi\frac{d}{\lambda}\cos\phi_{\mathrm{D},k,l}}, \ldots, e^{j2\pi\frac{d}{\lambda}(M-1)\cos\phi_{\mathrm{D},k,l}}]^T, \tag{4.13}$$

respectively, where $\lambda$ is the wavelength of the signal and $d$ denotes the inter-antenna space, which is usually set as $d = \lambda/2$, $\phi_{\mathrm{A},k,l} \backsim \mathcal{U}(\theta_{\mathrm{A},k,l} - \frac{\delta_{\mathrm{A}}}{2}, \theta_{\mathrm{A},k,l} + \frac{\delta_{\mathrm{A}}}{2})$ and $\phi_{\mathrm{D},k,l} \backsim \mathcal{U}(\theta_{\mathrm{D},k,l} - \frac{\delta_{\mathrm{D}}}{2}, \theta_{\mathrm{D},k,l} + \frac{\delta_{\mathrm{D}}}{2})$ with $\{\theta_{\mathrm{A},k,l}, \theta_{\mathrm{D},k,l}\}$ referring to the elevation angles and $\{\delta_{\mathrm{A}}, \delta_{\mathrm{D}}\}$ denoting the angular spread for arrival and departure, respectively [124].

### 4.2.2 Problem Formulation

As described above, the system performance heavily relies on precoding and combining vectors design. However, there is an inevitable feedback delay between the time point when the user estimates the channel and the BS starts transmitting data with the estimated channel fed back by the users. As can be seen in Section 4.2.1, such delay becomes quite problematic in high mobility scenarios since the channel changes fast and the correlation coefficient $\rho$ decreases dramatically. Therefore, it is necessary to develop strategies that are robust to feedback delay and user mobility, which, in this work, is interpreted as maximizing the sum-rate of $K$ users based on the knowledge

of past channels. The problem can be formulated as follows

$$\max_{\mathbf{W}_k(t),\mathbf{P}_k(t)} \quad \sum_{k=1}^{K}\sum_{n=1}^{N_s} G_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)) \tag{4.14a}$$

$$\text{s.t.} \quad \|\mathbf{p}_{k,n}(t)\|^2 = 1, \forall k,n, \tag{4.14b}$$

$$\|\mathbf{w}_{k,n}(t)\|^2 = 1, \forall k,n, \tag{4.14c}$$

$$\mathcal{F}(\mathbf{H}_k(t')), \forall k \text{ until } t' = t-1 \text{ are available}, \tag{4.14d}$$

where $\mathcal{F}(\mathbf{H}_k(t'))$ is a function of $\mathbf{H}_k(t')$ which is listed in Section 4.3. Problem (4.14) aims at optimizing the precoder and combiner to maximize the sum-rate for served users subject to constraints (4.14b)- (4.14d), which is a non-convex problem[3].To solve this problem, one traditional method using ZF-CI is illustrated in Section 2.3.8 and three efficient DRL-based strategies are proposed in Section 4.3.

## 4.3 Multi-agent DRL-based Discrete Beamforming Optimization

To build up the foundation for the proposed DRL-based designs, an overview of DQN is illustrated first, followed by the description of the state, action, reward function, and three multi-agent DRL-based algorithms for the problem (4.14).

### 4.3.1 A Brief Overview of DQN

As illustrated in Fig. 4.1, model-free Q-learning algorithms are demonstrated to continuously improve the policy through interactions with the environment. To be

[3]Note that this problem can be solved with different traditional methods such as ZF, WMMSE, and Fractional programming [4]. This thesis proposed a another method using multi-agent reinforcement learning.

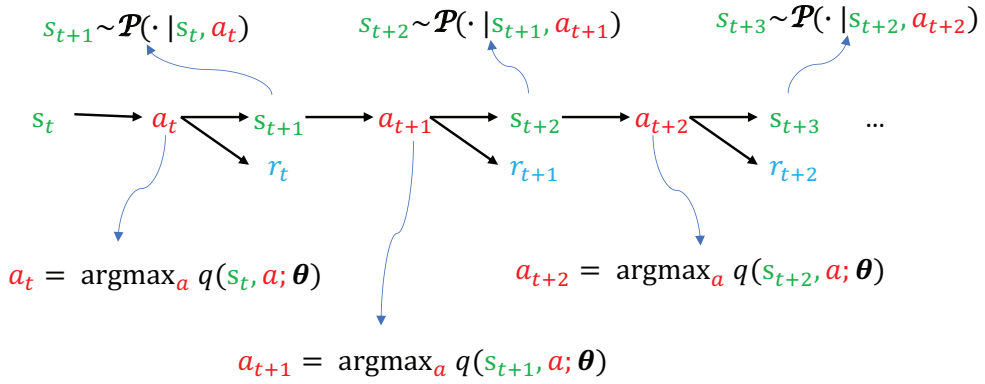



Figure 4.1: Markov decision process of Q-learning.

specific, the state-action value (called Q-value) is denoted as an expected reward of $(s,a)$ by policy $\pi$

$$\mathcal{Q}_\pi(s_t,a_t) = \mathcal{E}_\pi(R_t|s_t = s, a_t = a) \tag{4.15}$$

where the expectation is calculated over all the possible $(s,a)$ pairs given by policy $\pi$, which can be iteratively computed from the Bellman equation

$$\begin{aligned} &\mathcal{Q}_\pi(s_t,a_t) = \mathcal{R}(r_{t+1}|s_t = s, a_t = a) + \gamma \\ &\sum_{s' \in \mathcal{S}} \left( \mathcal{P}(s_{t+1} = s', s_t = s, a_t = a) \max_{a' \in \mathcal{A}} \mathcal{Q}_\pi(s',a') \right) \end{aligned} \tag{4.16}$$

where $\mathcal{P}(s_{t+1} = s', s_t = s, a_t = a)$ denotes the transition probability from state $s$ to $s'$ after taking action $a$. The optimal policy returns the maximum expected cumulative reward at each $s$, i.e., $\pi^* = \text{argmax}_\pi \mathcal{Q}^\pi(s,a)$. Then the Q-value function can be represented as

$$\begin{aligned} &\mathcal{Q}_{\pi^*}(s_t,a_t) = r_{t+1}(s_t = s, a_t = a, \pi = \pi^*) \\ &+ \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s_{t+1} = s', s_t = s, a_t = a) \max_{a' \in \mathcal{A}} \mathcal{Q}_{\pi^*}(s',a'). \end{aligned} \tag{4.17}$$

In classical Q-learning, a Q-value table $q(s,a)$, named as Q-table, is constructed to represent the Q-value function $\mathcal{Q}_\pi(s,a)$. This table consists of a discrete set of $|\mathcal{S}| \times |\mathcal{A}|$ which is randomly initialized. The agent then takes actions according to an $\epsilon$-greedy

policy, receives reward $r=\mathcal{R}(s,a)$ and transfers to the next state $s_{t+1}$ to complete the experience $e_t$. The Q-table is updated as

$$q(s_t,a_t) \longleftarrow (1-\alpha)q(s_t,a_t)+\alpha(r_{t+1}+\gamma \max_{a'} q(s_{t+1},a')) \tag{4.18}$$

where $\alpha \in [0,1)$ is the learning rate. However, it is challenging to directly obtain the optimal $\mathcal{Q}_{\pi^*}(s_t,a_t)$ due to the uncertain variation of the dynamic channel environment, i.e., an unlimited number of states. To address the problems with such an enormous state space, DQN is utilized here to approximate the Q-value function, which can be expressed as $q(s_t,a_t,\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ denoting the weights of DQN. The optimal policy $\pi^*$ can be represented by a group of weights of the DQN. In addition, two techniques are exploited to strengthen the stability of DRL: target network and experience replay. The target network $q(s_t,a_t,\bar{\boldsymbol{\theta}})$ is another network that is initialized with the same set of weights of trained DQN. The target DQN is used to generate the target Q-value which is exploited to formulate the loss function of trained DQN. The weights of target DQN are updated periodically for every fixed number of slots $T_s$ by replicating the weights of trained DQN to stabilize the training of trained DQN. The experience replay is intrinsically a first-input-first-output (FIFO) queue that stores $E_m$ historical experiences in each training slot. During training, $E_b$ experiences are sampled from the experience pool $\mathcal{O}$ to train the trained DQN to minimize the prediction error between the trained DQN and the target DQN. The loss function is defined as

$$L(\boldsymbol{\theta}) = \frac{1}{2E_b} \sum_{\langle s,a,r,s' \rangle \in \mathcal{O}} (r' - q(s,a;\boldsymbol{\theta}))^2 \tag{4.19}$$

where $r' = r+\gamma \max_{a'} q(s',a';\bar{\boldsymbol{\theta}})$, the weights of DQN $\boldsymbol{\theta}$ is updated by adopting a proper optimizer (e.g. RMSprop, Adam, and SGD). The specific gradient update is

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \mathcal{E}_{s,a,r,s' \in \mathcal{O}} \Big[ (r' - q(s,a;\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} q(s,a;\boldsymbol{\theta}) \Big]. \tag{4.20}$$

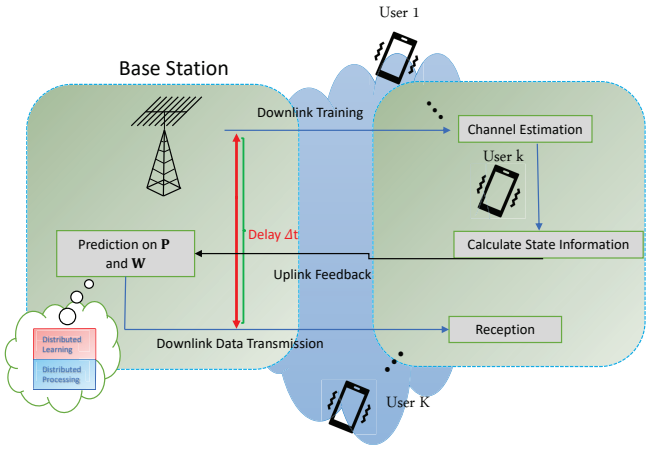



Figure 4.2: The downlink training and uplink feedback of proposed DRL framework. The detail structure of the distributed-learning-distributed-processing framework is shown in Fig. 4.4.

### 4.3.2 The Distributed-learning-distributed-processing DRL-based Algorithm

In this section, we cast the problem (4.14) as a sequential decision-making process and tailor three multi-agent DRL algorithms to solve it. The DRL-based framework is elaborated first, followed by the derived algorithms. To our best knowledge, this is the first work tackling the problem with 1) receive beamforming with multiple receive antennas, 2) transmit beamforming under imperfect CSIT, 3) multiple streams for each user, and 4) multiple users in a single cell compared with SISO in [18] and MISO in [19] with perfect CSIT, respectively. In addition, the PDRL is also firstly demonstrated in this work to bridge DDRL and CDRL to balance the performance and complexity.

#### Downlink Training and Uplink Feedback

As is shown in Fig. 4.2 and Fig. 4.3, at time slot $t-1$, the BS sends downlink pilots to users, based on which the downlink channels are perfectly estimated. User $k$ can estimate the designed state information in Section 4.3.2 and feed it back to

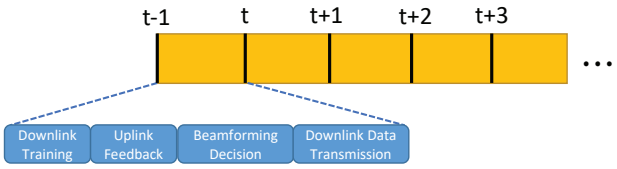



Figure 4.3: Timing of time slot $t-1$.

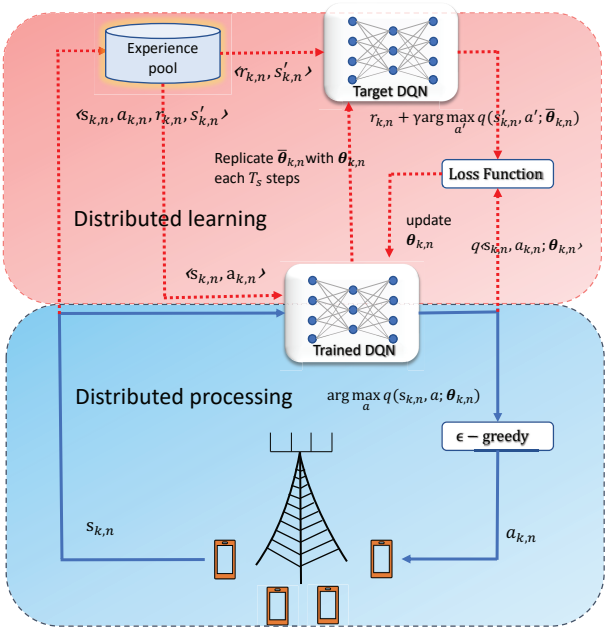



Figure 4.4: The framework of distributed-learning-distributed-processing scheme.

the BS. With feedback from users, the BS can predict the indexes of precoders and combiners for time slot $t$ and start downlink data transmission depending on the selected codewords in (4.22).

### The Proposed DRL-based Algorithm

To bring this insight to fruition, each stream is modeled as an agent, all the agents are independent with each other, so totally $KN_s$ homogeneous agents in our scheme. For example, the CSI delay is the same for all the agents. To be intuitive, we adopt

a distributed-learning-distributed-processing framework as shown in Fig. 4.4[4] and demonstrated in Algorithm 5. At the initialization stage, all the $KN_s$ pairs of DQNs are established at the BS. For instance, one pair of DQNs, namely trained DQN $q(s_{k,n},a_{k,n};\boldsymbol{\theta}_{k,n})$ and target DQN $q(s_{k,n},a_{k,n};\bar{\boldsymbol{\theta}}_{k,n})$ is possessed by agent $(k,n)$. The input and output of trained DQN $q(s_{k,n},a_{k,n};\boldsymbol{\theta}_{k,n})$ are the local state $s_{k,n}$ and action $a_{k,n}$. In terms of the distributed learning procedure for agent $(k,n)$, due to the feedback delay from users, only outdated CSI information is used to formulate the observations $s_{k,n}$ at the beginning of each time slot. Then, the DRL agent adopts an $\epsilon$-greedy to balance exploitation and exploration by choosing actions, i.e, the precoder $\mathbf{p}_{k,n}$, and combiner $\mathbf{w}_{k,n}$ according to $s_{k,n}$, in which the agent executes an action with probability $\epsilon$ randomly, or executes the action $a_{k,n}=\max_a q(s_{k,n},a;\boldsymbol{\theta}_{k,n})$ with probability $1-\epsilon$. Regarding the distributed learning process, the agent accumulates and stores the experience $e_{k,n}=\langle s_{k,n},a_{k,n},r_{k,n},s'_{k,n}\rangle$ into experience pool and the historical experiences can be utilized to train the DQN with local state-action pairs together with the corresponding reward. Each agent has a profound view of the relationship between local state-action pairs and local long-term reward which, in return, leads the whole system to a distributed-learning-distributed-processing manner.

### Actions of the proposed multi-agent DRL approach for massive MIMO scenario

As described in Section 4.2, we aim to optimize the precoder $\mathbf{p}_{k,n}$ and combiner $\mathbf{w}_{k,n},\forall k,n$. Then, the problem can be addressed by building two codebooks, i.e. $\mathcal{S}_{\mathrm{t}},\mathcal{S}_{\mathrm{r}}$, which contain $S_{\mathrm{t}}$ and $S_{\mathrm{r}}$ beamforming vectors. In the decision-making stage, each agent chooses one precoder from $\mathcal{S}_{\mathrm{t}}$ and one combiner from $\mathcal{S}_{\mathrm{r}}$. The action space can

[4]The distributed-learning-distributed-processing framework is preferred over a central-learning-central-processing strategy due to the curse of dimensionality problem since the computation complexity is too high to optimize a neural network with output size $((NK)^{S_tS_r})$.

be represented as

$$\mathcal{A}=\{(\mathbf{c}_{\mathrm{t}},\mathbf{c}_{\mathrm{r}}),\mathbf{c}_{\mathrm{t}}\in\mathcal{S}_{\mathrm{t}},\mathbf{c}_{\mathrm{r}}\in\mathcal{S}_{\mathrm{r}}\} \tag{4.21}$$

where $\mathbf{c}_{\mathrm{t}}$ and $\mathbf{c}_{\mathrm{r}}$ denote the codewords of two codebooks and the cardinal number of action space $\mathcal{A}$ is $S_{\mathrm{t}}\times S_{\mathrm{r}}$. The design of codebooks comes from [125] which is also applied in [19, 25, 26], and introduced here as a quantization of beam directions. To specify each element, we define matrix $\mathbf{C}_{\mathrm{t}}\in\mathbb{C}^{M\times S_{\mathrm{t}}}$ as

$$\mathbf{C}_{\mathrm{t}}[p,q]=\frac{\exp\left(j\frac{2\pi}{T}\lfloor\frac{M\bmod(q+\frac{S_{\mathrm{t}}}{2},S_{\mathrm{t}})}{S_{\mathrm{t}}/T}\rfloor\right)}{\sqrt{M}} \tag{4.22}$$

where $T$ is the number of available phase values and $\mathbf{C}_{\mathrm{r}}\in\mathbb{C}^{N\times S_{\mathrm{r}}}$ can be obtained by substituting the $M$ and $S_{\mathrm{t}}$ with $N$ and $S_{\mathrm{r}}$ accordingly. Each column of $\mathbf{C}_{\mathrm{t}}$ and $\mathbf{C}_{\mathrm{r}}$ corresponds to a specified codeword and the whole matrix forms a beamsteering-based beamformer codebook.

**States of the proposed DRL-based approach for massive MIMO scenarios**

With the mobility scenario, the receiver feedback is delayed at time slot $t$, and the state of agent $(k,n)$ is constructed by the representative feature of observations from the last two successive time slots $t-1$ and $t-2$ without observations from time slot $t$. That is to say, at the beginning of time slot $t-u$, due to the delay of feedback, the BS is unable to instantaneously obtain the power of the received signal, i.e., $|\mathbf{w}_{k,n}^{H}(t)\mathbf{H}_k(t)\mathbf{p}_{k,n}(t)|^2$ and $|\mathbf{w}_{k,n}^{H}(t)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t)|^2$. However, the historical feedback, i.e., $|\mathbf{w}_{k,n}^{H}(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t-1)|^2$ and $|\mathbf{w}_{k,n}^{H}(t-1)\mathbf{H}_k(t-2)\mathbf{p}_{k,n}(t-1)|^2$ are usually available to the BS. Based on this assumption, the state $s_{k,n}(t)$ is designed as follows

- The "desired" information of the agent $(k,n)$ which consists of 5 parameters, i.e., the channel gain $|\mathbf{w}_{k,n}^{H}(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t-1)|^2$, the chosen index of precoder $U_{k,n}(t-1)$, the chosen index of combiner $V_{k,n}(t-1)$, the achievable rate of stream

$n$ for user $k$, i.e., $G_{k,n}(\mathbf{P}_k(t-1),\mathbf{W}_k(t-1))$, and the interference-plus-noise $I_{k,n}(t-1)+I_{c,k}(t-1)+\sigma_k^2$.

- Interference information of the agent $(k,n)$ which is represented by 8 parameters, i.e., $\{\sum_{i=1,i\neq n}^{N_s}|\mathbf{w}_{k,n}^H(t-u)\mathbf{H}_k(t-u)\mathbf{p}_{k,i}(t-u)|^2, \sum_{i=1,i\neq n}^{N_s}|\mathbf{w}_{k,n}^H(t-1-u)\mathbf{H}_k(t-u)\mathbf{p}_{k,i}(t-1-u)|^2, \sum_{j\neq k}\sum_{i=1}^{N_s}|\mathbf{w}_{k,n}^H(t-u)\mathbf{H}_k(t-u)\mathbf{p}_{j,i}(t-u)|^2, \sum_{j\neq k}\sum_{i=1}^{N_s}|\mathbf{w}_{k,n}^H(t-1-u)\mathbf{H}_k(t-u)\mathbf{p}_{j,i}(t-1-u)|^2 | u\in\{1,2\}\}$. It is worth noting here that in such a system, the interference information plays a key role in the maximization of its own information rate (the rate of stream $n$ of user $k$), which, thus, should be included in state space.

- The information of agent $(j,i),(j,i)\neq(k,n),\forall j,i$ consists of of $10(KN-1)$ terms, i.e.,$\{U_{j,i}(t-u),V_{j,i}(t-u),G_{j,i}(\mathbf{M}_j(t-u),\frac{P}{NK}|\mathbf{w}_{j,i}^H(t-u)\mathbf{H}_j(t-u)\mathbf{p}_{j,i}(t-u)|^2,\frac{P}{NK}|\mathbf{w}_{j,i}^H(t-u)\mathbf{H}_j(t-u)\mathbf{p}_{k,n}(t-u)|^2|u\in\{1,2\}\}$. The information of other agents plays an irreplaceable role for agent $k$ to minimize the interference it causes to them, which, thus, should be included in state space.

To sum up, the cardinal number of state space is $10KN_s+3$. Note that the adopted design is not guaranteed to be the optimal one but empirically achieves a good performance as demonstrated with evaluation results in Section 4.3. The output size of the DQN is $S=S_tS_r$ which is equal to the number of available actions.

**The reward of the proposed DRL-based approach for the massive MIMO scenario**

In this massive MIMO scenario, if agent $(k,n)$ only tries to maximize the achievable rate of the stream $(k,n)$ without taking the inter-stream and multi-user interference into consideration, a large interference will be delivered to other agents. Therefore, our proposed reward function $r_{k,n}$ consists of penalty coefficient $\lambda$ and penalty term $P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$ to quantify the adverse impact each agent causes to other agents.

The penalty term $P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$ is given as

$$P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)) = \sum_{j=1,j\neq k}^{K}\sum_{i=1}^{N_s}\Bigg(\log_2(1+\frac{\frac{P}{KN_s}|\mathbf{w}_{j,i}^H(t)\mathbf{H}_j(t)\mathbf{p}_{j,i}(t)|^2}{\sigma^2+\hat{I}_{k,n_1}(t)+\hat{I}_{\mathrm{c}_1,k}(t)}) - G_{j,i}(\mathbf{W}_k(t),\mathbf{P}_k(t))\Bigg) + \sum_{i=1,i\neq n}^{N_s}\Bigg(\log_2(1+\frac{\frac{P}{KN_s}|\mathbf{w}_{k,i}^H(t)\mathbf{H}_k(t)\mathbf{p}_{k,i}(t)|^2}{\sigma^2+\hat{I}_{k,n_2}(t)+\hat{I}_{\mathrm{c}_2,k}(t)}) - G_{k,i}(\mathbf{W}_k(t),\mathbf{P}_k(t))\Bigg) \tag{4.23}$$

where $\hat{I}_{k,n_1}(t)$, $\hat{I}_{k,n_2}(t)$, $\hat{I}_{\mathrm{c}_1,k}(t)$ and $\hat{I}_{\mathrm{c}_2,k}(t)$ are given by

$$\hat{I}_{k,n_1}(t) = \sum_{i=1,i\neq l}^{N_s}\frac{P}{KN_s}|\mathbf{w}_{j,i}^H(t)\mathbf{H}_j(t)\mathbf{p}_{k,i}(t)|^2, \tag{4.24}$$

$$\hat{I}_{\mathrm{c}_1,k}(t) = \sum_{q\in\mathcal{K},q\neq k}\sum_{i=1}^{N_s}\frac{P}{KN_s}|\mathbf{w}_{j,i}^H(t)\mathbf{H}_j(t)\mathbf{p}_{q,i}(t)|^2 - \frac{P}{KN_s}|\mathbf{w}_{j,i}^H(t)\mathbf{H}_j(t)\mathbf{p}_{j,i}(t)|^2, \tag{4.25}$$

$$\hat{I}_{k,n_2}(t) = \sum_{h=1,h\neq i,n}^{N_s}\frac{P}{KN_s}|\mathbf{w}_{k,i}^H(t)\mathbf{H}_k(t)\mathbf{p}_{k,h}(t)|^2, \tag{4.26}$$

and

$$\hat{I}_{\mathrm{c}_2,k}(t) = \sum_{q\in\mathcal{K},q\neq k}\sum_{h=1}^{N_s}\frac{P}{KN_s}|\mathbf{w}_{k,i}^H(t)\mathbf{H}_k(t)\mathbf{p}_{q,h}(t)|^2, \tag{4.27}$$

respectively. Note that $P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$ is always a positive value due to the extraction of the interference from a specified stream. Then, the achievable rate for stream $(k,n)$, i.e., $G_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$, is added into $r_{k,n}$ to highlight the contribution of agent $(k,n)$ to the total information rate. Hence, $r_{k,n}$ at time slot $t$ is given as

$$r_{k,n}(t) = G_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)) - \lambda P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)) \tag{4.28}$$

where penalty coefficient $\lambda$ is used here as a weight parameter to manipulate the amount of negative effect in the reward function. In regard to the reward function, the rationale behind such a design is to maximize the achievable rate of improvement if the interference caused by stream $(k,n)$ is totally eliminated. This design not only maximizes the achievable rate of stream $(k,n)$, i.e., $G_{k,n}$ but also minimizes the negative effect it causes to other streams, i.e., $P_{k,n}$. Similar designs are comprehensively

discussed in [18, 19] which also confirm that a well-formulated reward function should act as a catalyst of the best decisions obtained by multiple agents.

Table 4.1: Comparison Between Strategies

| | FDD & ZF-CI [80] | TDD & ZF-CI [80] | FDD & MA-DRL |
|---|---|---|---|
| CSI Overhead Uplink | $MN+KN$ ($MN$ channel coefficients and $KN$ CSI pilot symbols) | $KN$ sounding pilot symbols | $MN+KN+11$( $MN$ channel coefficients, $KN$ CSI pilot symbols, and $\{\|\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t-1)\|^2$, $G_{k,n}(\mathbf{P}_k(t-1),\mathbf{W}_k(t-1))$, $I_{k,n}(t-1)+I_{c,k}(t-1)+\sigma_k^2$, $\sum_{i=1,i\neq n}^{N_s}\|\mathbf{w}_{k,n}^H(t-v-u)\mathbf{H}_k(t-u)\mathbf{p}_{k,i}(t-v-u)\|^2$, $\sum_{j\neq k}\sum_{i=1}^{N_s}\|\mathbf{w}_{k,n}^H(t-v-u)\mathbf{H}_k(t-u)\mathbf{p}_{j,i}(t-v-u)\|^2$, $u\in\{1,2\},v\in\{0,1\}\}$ |
| CSI Overhead Downlink | $MN$ CSI pilot symbols | 0 | $MN+2N$ ($MN$ channel coefficients and indexes of precoders and combiner $\mathbf{p}_{k,n}$ and $\mathbf{w}_{k,n}$) |
| Computing Complexity of Precoding and Combining Matrices | $\mathcal{O}((MN)^3)$ | $\mathcal{O}((MN)^3)$ | $\mathcal{O}((10KN+3)L_1+L_1L_2+L_2S)$ |

**Discussion on the overhead and complexity of the proposed framework**

As is shown in Table 4.1, if the BS has to tell the users what combiner to use, then it can consume additional overhead on the downlink transmission. Fortunately, this overhead is negligible since only the indexes of combiners are delivered to users. Note that the precoders are also sent to terminals for the calculation of state information listed in the table. This reduces the computation burden on the BS for processing this state information.

In terms of the computational complexity of precoders and combiners in the demonstrated DRL-based approaches, the designed structure of target/trained DQNs includes four fully connected layers. Specifically, the input layer consists of $10KNs+3$ neurons, followed by two hidden layers with $L_1$ and $L_2$ neurons and a specified activation function. The fourth layer serves as the output layer with $S$ neurons. We employ two hidden layers in our design, as a two-layer feedforward neural network is sufficient to approximate any nonlinear continuous function based on the *universal approximation theorem* [126]. The computational complexity of fully connected DNN can be written as $\mathcal{O}((10KN_s+3)L_1+L_1L_2+L_2S)$ for each agent. This is

**Algorithm 5** DDRL Algorithm

**Initialize**: Establish a trained DQN and target DQN with random weights $\boldsymbol{\theta}_{k,n}$ and $\bar{\boldsymbol{\theta}}_{k,n}$, respectively, $\forall k \in \{1,2,...,K\}, \forall n \in \{1,2,...,N_s\}$, update the weights of $\bar{\boldsymbol{\theta}}_{k,n}$ with $\boldsymbol{\theta}_{k,n}$.
In the first $E_s$ time slots, agent $(k,n)$ randomly selects an action from action space $\mathcal{A}$, and stores the corresponding experience $\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n} \rangle$ in its pool, $\forall k,n$.
**for** each time slot $t$ **do**
  **for** each agent $(k,n)$ **do**
    Obtain state $s_{k,n}$ from the observation of agent $(k,n)$.
    Generate a random number $\omega$.
    **If** $\omega < \epsilon$ **then**:
      Randomly select an action in action space $\mathcal{A}$.
    **Else**
      Choose the action $a_{k,n}$ according to the Q-function $q(s_{k,n}.a; \boldsymbol{\theta}_{k,n}), \forall k,n$
    **End if** .
    Agent $(k,n)$ executes the $a_{k,n}$, immediately receives the reward $r_{k,n}$ and steps into next state $s'_{k,n}, \forall k,n$.
    Agent $(k,n)$ puts experience $\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n} \rangle$ into experience pool $\mathcal{O}_{k,n}$, randomly samples a minibatch with size $E_b$. Then, the weights of trained DQN $\boldsymbol{\theta}_{k,n}$ are updated using back propagation approach. The weights of target DQN $\bar{\boldsymbol{\theta}}_{k,n}$ is updated every $T_s$ steps.
  **end for**
**end for**

much smaller than that of ZF-CI scheme due to the fact that ZF-CI involves matrix inversion which limits the scalability to a large number of transmit and receive antennas.

**Remark 7.** *Note that different from [96] where the mobility estimation and channel prediction are needed, our work does not predict the channels sequentially. In this work, we demonstrated a low complexity and efficient DRL-based framework and as this is the first work proposing DRL-based joint transmit and receiver beamforming for massive MIMO downlink transmission, we would like to keep the benchmarks as clear and simple as possible such that researchers can understand the fundamental benefits of the proposed strategies and carry on their studies in more practical scenarios in the future. The comparison with mobility estimation and channel prediction methods (such as VFK and MLP methods in [96]) could be addressed in future research, but not the scope of this work.*

### 4.3.3 The Low-complexity Centralized-learning-distributed-processing DRL-based Algorithm

In this section, we demonstrate an extra algorithm for the problem (4.14) for three reasons. *First*, a lower computation complexity is achieved in the centralized scheme by building and training on an extra pair of DQNs instead of distributedly training with $KN_s$ agents. *Second*, a lower storage space is required with only a central experience pool during the learning process. *Third*, by saving and sampling the experiences from all distributed agents, the central agent can learn the common features from the channels of all users and intelligently guide the decision-making procedure of all distributed agents. Things need to be noted that CDRL is trained more efficiently using parameter sharing, which is based on homogeneous agents. This allows the policy to be trained with the experiences of all agents simultaneously. However, it still allows different actions between agents due to the fact that each agent receives different observations. This algorithm focuses on the decentralized parameter-sharing training scheme since we found it to be scalable if we continue to increase the number of users and streams.

It is worth noting that we don't need to design a new state, action, and reward function in CDRL since all the agents upload their experiences to a central pool for training the central DQN, and the central training DQN broadcasts its weights to all agents for later distributed execution.

The whole process is shown in Algorithm 6. At the initialization stage, only one pair of target and trained DQNs is built for the central agent. For each distributed agent, one trained DQN is established. In the first several time slots, each agent randomly selects an action and saves the experiences into the central experience pool. When the episode begins, the central agent adopts an $\epsilon$-greedy strategy to balance exploitation and exploration so as to find the optimal policy. After learning from the sample experiences, the central agent broadcasts the updated weights of the central trained

**Algorithm 6** CDRL Algorithm

**Initialize**: Establish a central trained DQN and central target DQN with random weights $\boldsymbol{\theta}_{\mathrm{c}}$ and $\bar{\boldsymbol{\theta}}_{\mathrm{c}}$ for the central agent, update the weights of $\bar{\boldsymbol{\theta}}_{\mathrm{c}}$ with $\boldsymbol{\theta}_{\mathrm{c}}$. Establish a trained DQN with random weight $\boldsymbol{\theta}_{k,n}$, $\forall k \in \{1,2,...,K\}, \forall n \in \{1,2,...,N_s\}$ for each distributed agent.
In the first $E_s$ time slots, agent $(k,n)$ randomly selects an action from action space $\mathcal{A}$, and stores the experience $\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n} \rangle, \forall k,n$ in the experience pool of central agent $\mathcal{O}_{\mathrm{c}}$.
**for** each time slot $t$ **do**
  **for** each agent $(k,n)$ **do**
    Obtain state $s_{k,n}$ from the observation of agent $(k,n)$.
    Generate a random number $\omega$.
    **If** $\omega < \epsilon$ **then**:
      Randomly select an action in action space $\mathcal{A}$.
    **Else**
      Choose the action $a_{k,n}$ according to the Q-function $q(s_{k,n}.a; \boldsymbol{\theta}_{k,n}), \forall k,n$
    **End if** .
    Agent $(k,n)$ executes the $a_{k,n}$, immediately receives the reward $r_{k,n}$ and steps into next state $s'_{k,n}, \forall k,n$.
    Agent $(k,n)$ puts experience $\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n} \rangle$ into central experience pool $\mathcal{O}_{\mathrm{c}}$.
  **end for**
  Central agent randomly samples a minibatch with size $E_b$. Then, the weights of central trained DQN $\boldsymbol{\theta}_{\mathrm{c}}$ are updated using the back propagation approach. The weights of target DQN $\bar{\boldsymbol{\theta}}_{\mathrm{c}}$ is updated every $T_s$ steps. Then, central agent broadcasts the weights $\boldsymbol{\theta}_{\mathrm{c}}$ to all the distributed agents, i.e., $\boldsymbol{\theta}_{k,n} = \boldsymbol{\theta}_{\mathrm{c}}, \forall k,n$.
**end for**

DQN to all other distributed agents for decision-making purposes.

### 4.3.4 Bridging the DDRL and CDRL: Partial-distributed-learning-distributed-processing Scheme

In contrast with DDRL and CDRL, the partial-distributed-learning-distributed-processing DRL-based scheme (PDRL) offers a more flexible solution to the problem (4.14) by modeling each user as an agent. In the extreme case of $N_s = 1, K > 1$, PDRL boils down to DDRL by simply treating each stream as an agent. In the other extreme case of $K = 1, N_s > 1$, PDRL boils down to CDRL by forcing one

central agent to do the training work. Compared with CDRL, PDRL demonstrates better performance-complexity balance by learning the representative features of the propagation environment for a specified user which is demonstrated in Fig. 4.12. The whole algorithm is illustrated in Algorithm 7.

**Algorithm 7** PDRL Algorithm

**Initialize**: Establish $K$ pairs of trained/target DQNs with random weights $\boldsymbol{\theta}_k$ and $\bar{\boldsymbol{\theta}}_k, \forall k \in \{1,2,3,...,K\}$ as user-specified agents, update the weights of $\bar{\boldsymbol{\theta}}_k$ with random $\boldsymbol{\theta}_k$. Build experience pool $\mathcal{O}_k, \forall k$. Establish a trained DQN with weight $\boldsymbol{\theta}_{k,n}$, $\forall k \in \{1,2,...,K\}, \forall n \in \{1,2,...,N\}$ for each distributed agent.
In the first $E_s$ time slots, agent $(k,n)$ randomly selects an action from action space $\mathcal{A}$, and stores the experience $\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n} \rangle, \forall k,n$ in the experience pool of corresponding user-specified agent $\mathcal{O}_k$.
**for** each time slot $t$ **do**
  **for** each agent $(k,n)$ **do**
    Obtain state $s_{k,n}$ from the observation of agent $(k,n)$.
    Generate a random number $\omega$.
    **If** $\omega < \epsilon$ **then**:
      Randomly select an action in action space $\mathcal{A}$.
    **Else**
      Choose the action $a_{k,n}$ according to the Q-function $q(s_{k,n}.a; \boldsymbol{\theta}_{k,n}), \forall k,n$
    **End if** .
    Agent $(k,n)$ executes the $a_{k,n}$, immediately receives the reward $r_{k,n}$ and steps into next state $s'_{k,n}, \forall k,n$.
    Agent $(k,n)$ puts experience $\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n} \rangle$ into central experience pool $\mathcal{O}_k$.
  **end for**
  User-specified agent $k$ randomly samples a minibatch with size $E_b$. Then, the weights of its trained DQN $\boldsymbol{\theta}_k$ are updated using back propagation approach. The weights of its target DQN $\bar{\boldsymbol{\theta}}_k$ is updated every $T_s$ steps. Then, the user-specified agent broadcasts the weights $\boldsymbol{\theta}_k$ to the corresponding distributed agents, i.e., $\boldsymbol{\theta}_{k,n} = \boldsymbol{\theta}_k, \forall n$.
**end for**

Table 4.2: Comparison Between Algorithms

| | DDRL | CDRL | PDRL |
|---|---|---|---|
| Computation Complexity Each Channel Realization | High | Middle | Low |
| Number of Agents | $KN_s$ | 1 | $K$ |
| Convergence of Training | Fast | Slow | Slow |

## 4.4 Simulation Result and Analysis

This section demonstrates the performance of our proposed multi-agent DRL-based algorithm to maximize the average throughput of all the users. We first illustrate the simulation setup, followed by the simulation results in different scenarios.

### 4.4.1 Simulation Setup

We consider a downlink transmission from one BS to multiple users. The BS serves $K=4$ users in a single cell. The maximum transmit power $P$ is fixed to 20 dBm and noise variance $\sigma^2$ at users is fixed to -114 dBm. The BS is equipped with $M=32$ transmit antennas and the users are equipped with $N=4$ receive antennas unless otherwise stated. Without loss of generality, the uniform linear array (ULA) is equipped in both transmitter and receiver sides with half-wavelength inter-antenna spacing. The large-scale channel fading is characterized by the log-distance path-loss model expressed below

$$\eta = L(d_0) + 10\omega \log_{10} \frac{d}{d_0}. \tag{4.29}$$

where $d=10$ m is the BS-user distance. According to Table III of [127], the value of $L(d_0)$ for $d_0=1$ m is 68 dB and fading coefficient $\omega$ is 1.7. In terms of the shadowing model, the log-normal shadowing standard deviation $\beta_k$ is set to 1.8 dB. The small-scale fading channel is generated according to the channel model introduced in Section 4.2. Regarding the parameters of Jake's model with user speed 3.55 km/h, the maximum Doppler frequency $f_{\mathrm{d}}^{\max}$ and channel instantiation interval $T_i$ are set as 800 Hz and $1\times 10^{-3}$ s, respectively [127]. The corresponding correlation coefficient $\rho$ is $0.6514\approx 0.65$.

As is illustrated in Fig. 4.4, the whole framework can be divided into 2 phases, the learning phase, and the processing phase. Before the learning phase, we randomly generate channels obeying Jake's model, randomly choose actions, observe the reward, and accu-

mulate and store the corresponding experiences into the experience pool with size 1000 for the first 200 time slots, i.e., $E_m = 1000, E_s = 200$. In addition, the mini-batch size $E_b$ is set as 32. Stepping into the learning stage, for the DNN, the number of neurons in two hidden layers, i.e, $L_1, L_2$, are both set as 256, followed by the $ReLu$ activation function. The initial learning rate $\alpha(0)$ is $5e^{-3}$ and the decaying rate $d_c$ is $10^{-4}$ such that the learning rate continues to decay with the number of time slots following $\alpha(t) = \alpha(t-1) * \frac{1}{1+d_c t}$. In terms of optimization, the adaptive moment estimation ($Adam$) is utilized to prevent the diminishing learning rate problem. To minimize the prediction error between trained DQN and target DQN, the weights of trained DQN are substituted into target DQN every 120 time slots, i.e., $T_s = 120$ with discount factor $\gamma$ and penalty coefficient $\lambda$ set as 0.1 and 1, respectively. During the processing phase, for $\epsilon$-greedy strategy, we set the initial exploration coefficient $\epsilon$ as 0.7 which decays exponentially to 0.001 to strike a balance between exploration and exploitation during the training process. Note that the adopted parameters are not guaranteed to be optimal ones, which experimentally perform well in this setup. In the legend of simulation figures, DDRL and CDRL come from Algorithm 5 and Algorithm 6, respectively. The value of each point is a moving average over the previous 500 time slots unless otherwise stated.

To demonstrate the effectiveness of our DRL-based approaches, four benchmark schemes are evaluated, which are as follows:

- ZF-CI PCSI: Each agent executes the action from the scheme in [67] with instantaneous and perfect CSI, i.e., $\mathbf{H}_k(t), \forall k$.
- SAH: This approach stores the most recent estimated channel, i.e., $\mathbf{H}_k(t-1), \forall k$ and this approach always sends the channel coefficients to the BS, which will be used for calculating the precoders using ZF-CI. This strategy essentially ignores the non-negligible delay between the channel estimation and the time point when the actual DL transmission happens [99]. When $\rho = 1$, SAH is the same as ZF-CI PCSI. SAH only captures delay but assumes perfect knowledge of CSI at $t-1$.

- Random: Each agent randomly chooses actions. The performance serves as a lower bound in the simulation.

- Greedy Beam Selection (GBS): Each agent exhaustively selects an action in a greedy manner, the actions with the highest sum information rate are chosen as the solution for each channel realization. The benchmark serves as the upper bound for DRL-based strategies. Note that the size of the beam selection set increases exponentially with the size of the codebooks ($(NK)^{S_t S_r}$). For instance, when $K=4, N=1, S_{\mathrm{t}}=32, S_{\mathrm{r}}=4$, the total number of action combination is $4^{32}$ which is quite large considering the hardware constraint. Thus, we consider (8,1) in this benchmark.

### 4.4.2 DDRL vs CDRL

Fig. 4.5 depicts the average achievable information rate versus the number of time slots with different numbers of transmit and receive beamformer codebook size ($S_{\mathrm{t}}$,$S_{\mathrm{r}}$). A *first* observation is that the performance gaps between two DRL-based schemes and SAH are gradually increased with the number of $S_{\mathrm{t}}$ and $S_{\mathrm{r}}$ and DDRL roughly observes a gain of 380% over SAH when $S_{\mathrm{t}}=32$ and $S_{\mathrm{r}}=4$. The reason behind such a phenomenon is that, in DRL-based strategies, better interference management can be achieved by higher resolution in the codebook which significantly reduces the quantization error and effectively alleviates the interference from other streams. A *second* observation is that DDRL can achieve nearly 90% of the system capacity of the ZF-CI PCSI and 95% of the Beam Selection strategy, utilizing only a few pieces of information in the designed features from itself and other users. An interpretation is that the lack of instantaneous CSI and imperfect codebook design degrade the system performance and this 10% and 5% gap cannot be fully eliminated. A *third* observation is that DDRL (32,8) only demonstrates slightly better performance than DDRL (32,4) and DDRL (16,4) which shows of robustness on codebook size. A *fourth* observation is that the

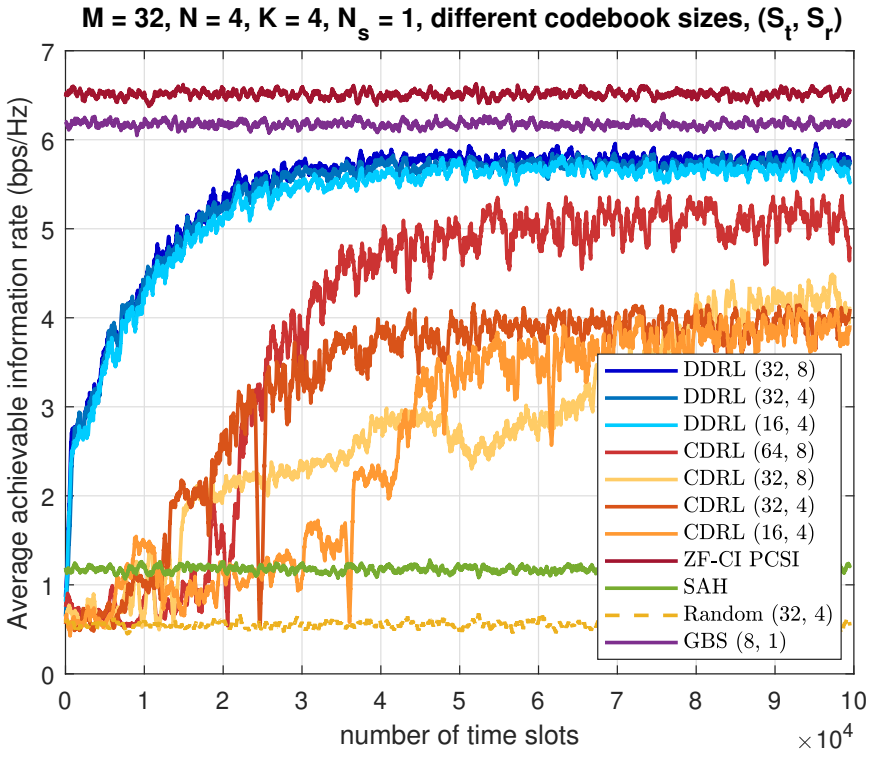



Figure 4.5: Average information rate versus the number of time slots with different codebook sizes ($S_{\mathrm{t}}$,$S_{\mathrm{r}}$).

CDRL always demonstrates instability before convergence. An explanation is that the huge differences between the dynamic environment of different users make it extremely difficult to find the commonality among them. Then, each agent could be misled by the experiences of other agents, which, thus, results in fluctuations before convergence and degradation in system performance. Conversely, in DDRL, each agent selects a specific precoder and combiner for its intended stream which is relevant to its propagation environment and is considerably different among streams. This local adaptability greatly improves the performance of DDRL. After comprehensive considerations among computation complexity, system performance, and convergence speed, (32,4) is chosen as a codebook baseline in the simulations of both DRL-based schemes. Compared with the Greedy Beam Selection method, the DRL-based strategies reveal extremely lower complexity on large action space but achieve roughly 95% of Greedy Beam Selection perfor-

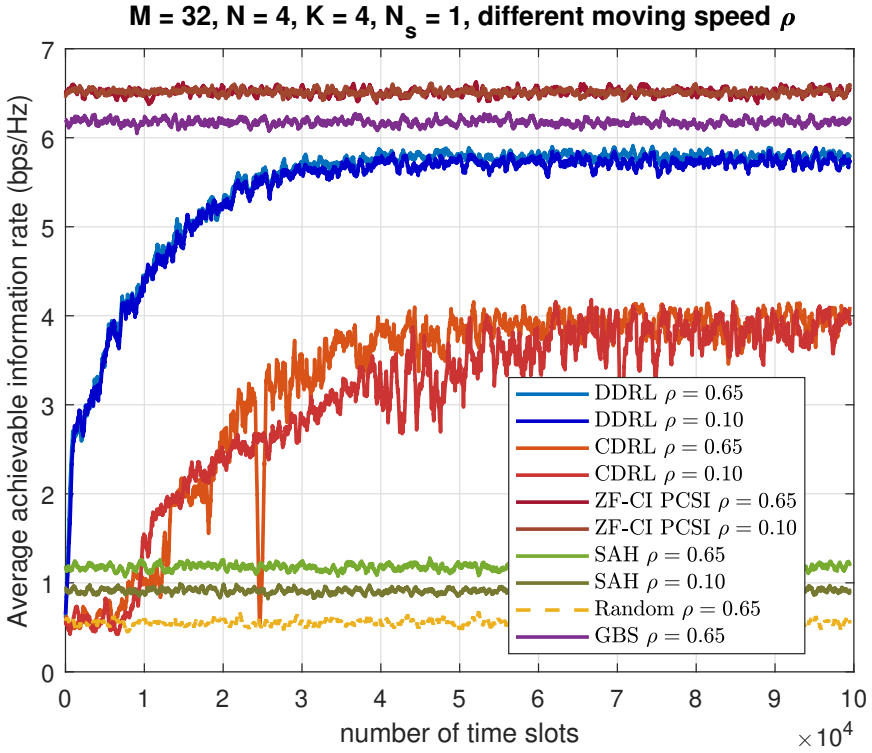



Figure 4.6: Average achievable information rate versus the number of time slots with different correlation coefficients.

mance.4.3. Hence, CDRL is not suitable for practical systems where a large codebook is not available while DDRL demonstrates more robustness regarding codebook size but requires much more amount of memory and computing resources for training. We implemented the demonstrated algorithms with TensorFlow in a general computer, i.e., i7-8700 CPU, 3.20 GHz. The running time for different algorithms is listed in Table. 4.3.

Table 4.3: Running Time For Each Network Topology

| | DDRL (32, 4) | ZF PCSI | SAH | Beam Selection (8, 1) | Random Selection |
|---|---|---|---|---|---|
| Time | 0.2s | 0.8s | 0.8s | 10s | 0.1s |

Table 4.4: Total Execution Time For Each Channel Realization With Different $(K,M,N,N_s)$ In DDRL($\epsilon=0$)

| | (1, 32, 4, 1) | (1, 32, 4, 2) | (2, 32, 4, 1) | (4, 32, 4, 1) | (4, 32, 4, 2) |
|---|---|---|---|---|---|
| Time | 0.000992s | 0.001993s | 0.001996s | 0.003786s | 0.007270s |

Fig. 4.6 exhibits the average achievable information rate versus the number of time

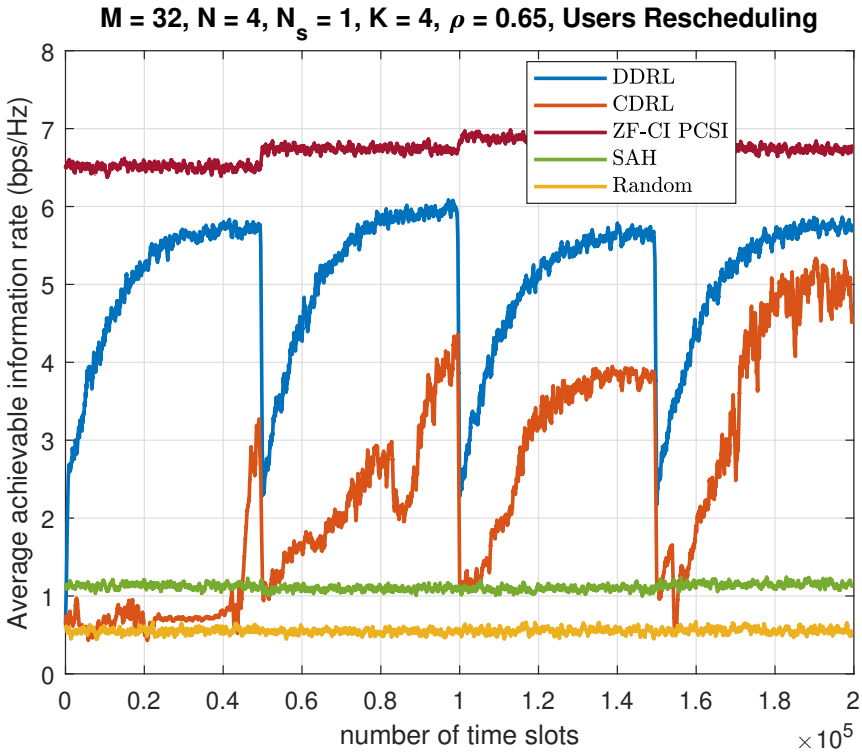



Figure 4.7: Average information rate versus number of time slots with users rescheduling at 5e4th, 1e5th and 15e4th time slot.

slots with different values of correlation coefficient $\rho$. The DDRL scheme with $\rho = 0.65$ and $\rho = 0.1$ can exceed the benchmark SAH with approximately 380% and 500%, respectively. This result greatly embodies the superiority of our DRL-based framework over the traditional massive MIMO optimizing scheme in mobility scenarios since a 20% performance degradation is caused by the fast-changing channels in SAH. In addition, it can be observed that DDRL with $\rho = 0.1$ demonstrates a slightly lower performance than $\rho = 0.65$ which is also shown in CDRL. An explanation is that the DDRL scheme is not sensitive to the dynamic and fast-changing wireless environment but CDRL needs more time steps to learn the representative features of the rapid-changing environment in high mobility scenarios, which results in a lower convergence. Note that the DRL-based methods have certain adaptability to environmental changes in user speed which can be interpreted as robustness on max Doppler frequency. Even though the correla-

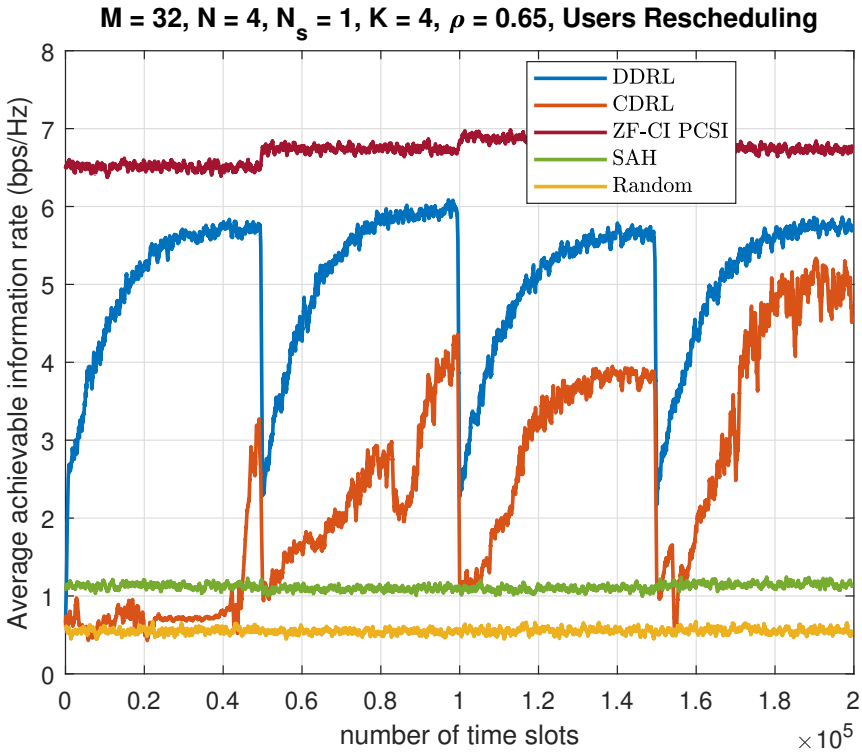



Figure 4.8: Average information rate versus number of time slots with different number of users $K$.

tion between adjacent channels is very small, the DRL-based frameworks still benefit from the exploration-exploitation strategy. Similar results are also observed in [26].

In Fig. 4.7, we assume that the users' rescheduling happens at the 50000th, 100000th, and 150000th-time slots. Instead of re-initializing the weights of all the DQNs in each agent, all of them continue the training process based on the designed information from newly scheduled users. *First*, a much higher start point and a comparable convergence time can be achieved in DDRL without witnessing a great performance collapse compared with the ZF-CI PCSI scheme. This can be interpreted by the fact that each agent tries to find the common features between the first scheduled and reschedule users which naturally makes a better decision based on these features and exhibits the ability for maintaining connectivity against user rescheduling in mobile networks. *Second*, with more rescheduling happening, a higher information rate and a faster convergence

can be observed in CDRL. An interpretation is that the common features learned from the previously scheduled users boost the training in rescheduling. After learning and 'storing' more and more feature information into the weights of DQN, the central agent demonstrates universality to the channel uncertainty of rescheduled users and the neural network weights extracted from the previously trained DQN is a good candidate for the weight initialization of the current trained DQN in both schemes. However, if rescheduling happened frequently, the proposed DRL-based schemes can not converge to a set of good parameters if rescheduling happens before 25000-time slots but a jump-start happens when we implement a group of trained DQN to new users.

Fig. 4.8 investigates the average achievable information rate versus the number of time slots with $K = 4$ and 6, respectively. As opposed to the decrease of average user rate, the total cell throughput improves which suggests that both DRL-based approaches can benefit from multi-user diversity provided by time-varying channels across different users. In addition, although an 11.7% performance degradation can be observed in CDRL which is smaller than 19.8% in DDRL, CDRL achieves a much smaller performance gain over SAH in comparison to DDRL when $K = 6$. This suggests that DDRL is more robust in multi-user scenarios than CDRL.

To sum up in a nutshell, on the one hand, CDRL and DDRL are not equally suitable for massive MIMO systems with channel aging, namely, CDRL is less computationally complex but demonstrates instability and incurs performance loss. In contrast, DDRL offers a promising gain over CDRL by favoring an adaptive decision-making process and facilitating cooperation among all agents to mitigate interference but incurs a higher hardware complexity. Also, DDRL is more robust on rescheduling than CDRL considering the convergence speed and stability. On the other hand, both schemes demonstrate robustness on fading characteristics of the environment and changes on interference conditions.

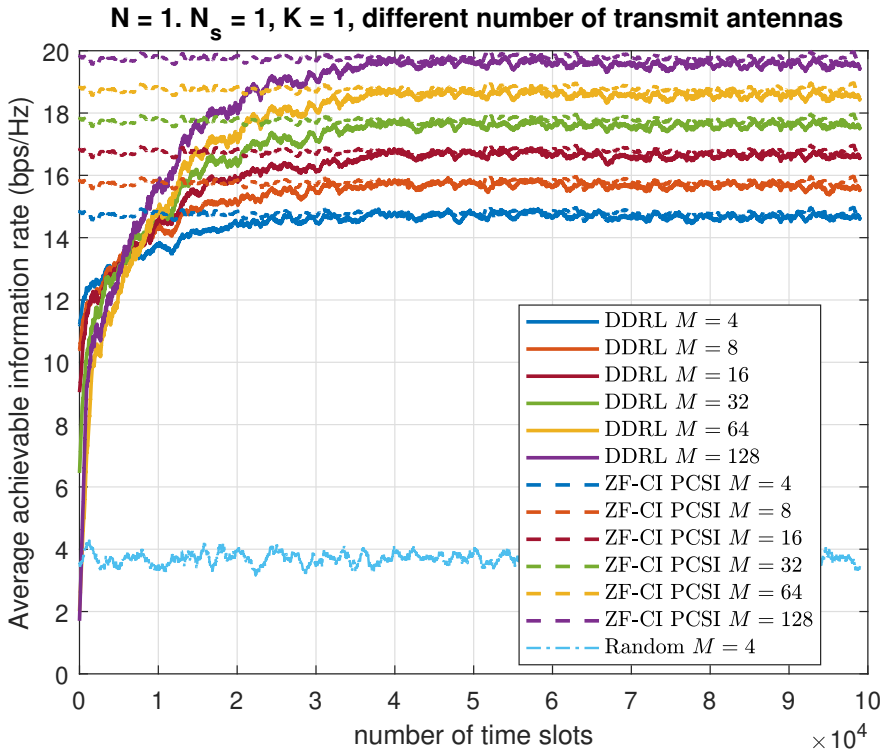



Figure 4.9: Average information rate versus the number of time slots with different number of transmit antennas $M$ when $N=1,N_s=1,K=1$.

### 4.4.3 Multi-antenna and Multi-stream

Fig. 4.9 illustrates the DRL-based scheme with 6 different numbers of transmit antennas $M$ when $N=1,N_s=1,K=1$. Without any penalty, i.e., inter-stream interference and multi-user interference, a near-optimal result can be observed by leveraging the proposed state, action, and reward design in an interference-free scenario with a stable increase of information rate, which thereby validates the effectiveness of the codebook design in Section 4.3. The convergence time is proportional to $M$ which limits its scalability. Intuitively, the reason for this effect on the convergence time is that it takes a longer time to learn the representative feature of high-dimension CSIT in sequential states. In contrast with Fig. 4.9, a serious multi-user and inter-stream interference is managed in Fig. 4.10 when $N=4,N_s=2,K=4$. It can be observed that the transmit diversity and array gain cannot be fully achieved in the proposed DRL-based scheme if the rich inter-

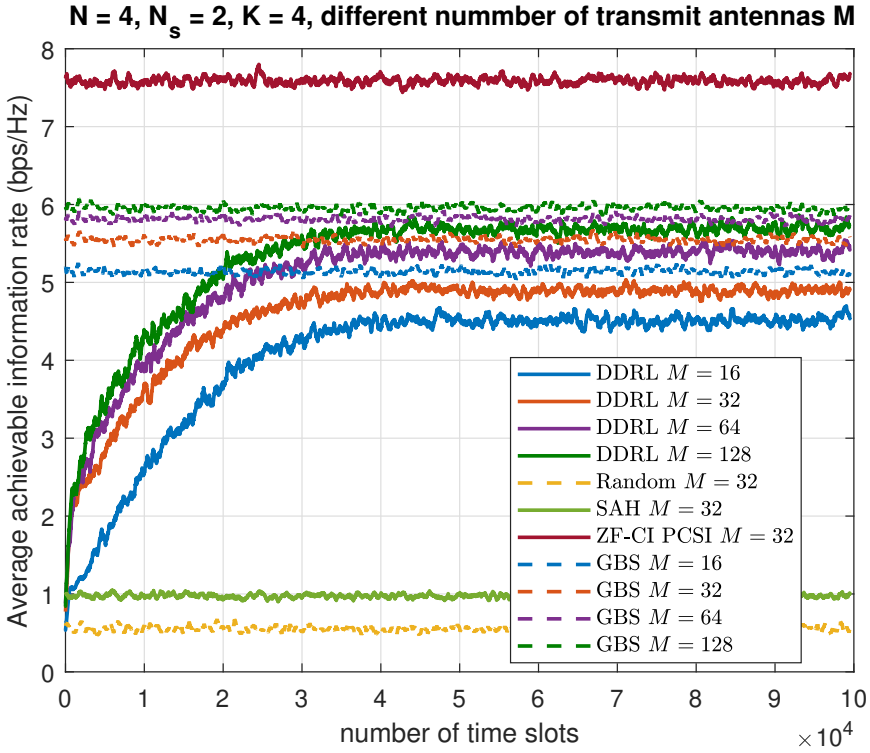



Figure 4.10: Average information rate versus number of time slots with different number of transmit antennas $M$ when $N=4, N_s=2, K=4$.

ference is not properly suppressed due to the constraint of codebook precision and CSIT imperfections. Hence, the drawback of using DRL-based methods is that inter-stream interference can not be sufficiently alleviated if each agent fails to choose an action that causes small interference to all other agents during exploration and exploitation.

Fig. 4.11 characterizes the average achievable information rate versus the number of time slots with different numbers of streams for each user. First, DDRL has significantly higher performance compared to the conventional SAH scheme in different numbers of streams. Second, a 15.7% performance degradation can be observed from the DRL-based scheme between 1-stream and 2-stream scenarios which is smaller than that in SAH (around 28% between black and blue dotted lines). This reveals the privilege of the DDRL in inter-stream interference management.

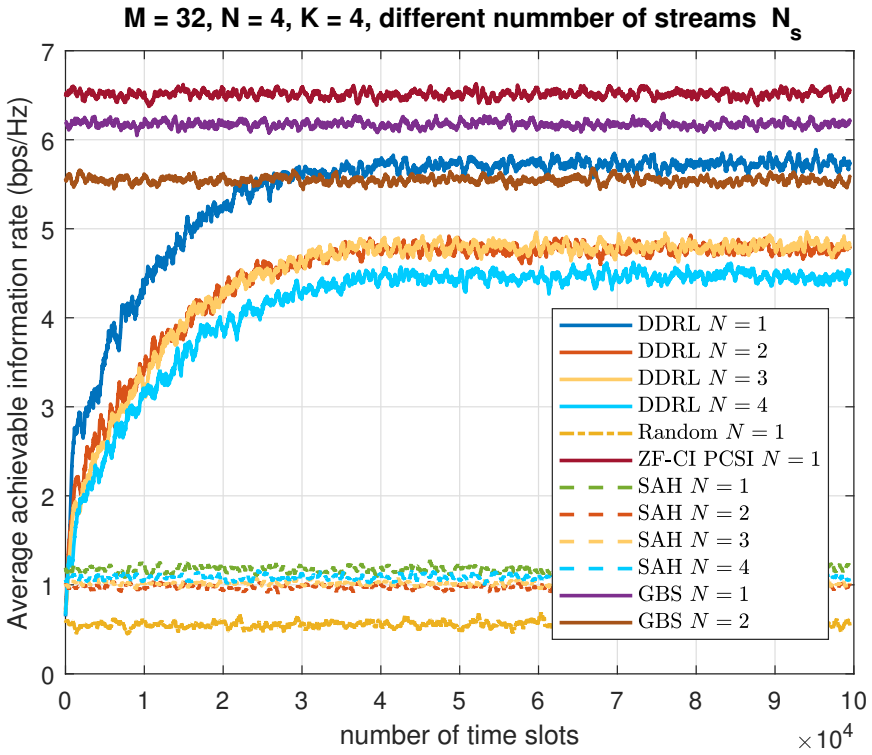



Figure 4.11: Average information rate versus the number of time slots with different number of streams $N$.

An overview of the average information rate versus number of time slots with three DRL-based algorithms is demonstrated in Fig. 4.12. Compared with DDRL and PDRL, a performance collapse is observed in CDRL due to the degrading effect of inter-stream interference. By flexibly modeling each user as an agent, PDRL greatly mitigates the inter-stream interference by learning the local observations from the user's propagation channel.

### 4.4.4 Reward, Penalty Analysis and Statistical Test

To reveal the significance of the neural network size ($L_1$,$L_2$), the learning rate $\alpha$ and the discount factor $\gamma$, Fig. 4.13 shows the sum reward versus the number of time slots with different ($L_1$,$L_2$), $\alpha$ and $\gamma$. The first observation is that a faster convergence is

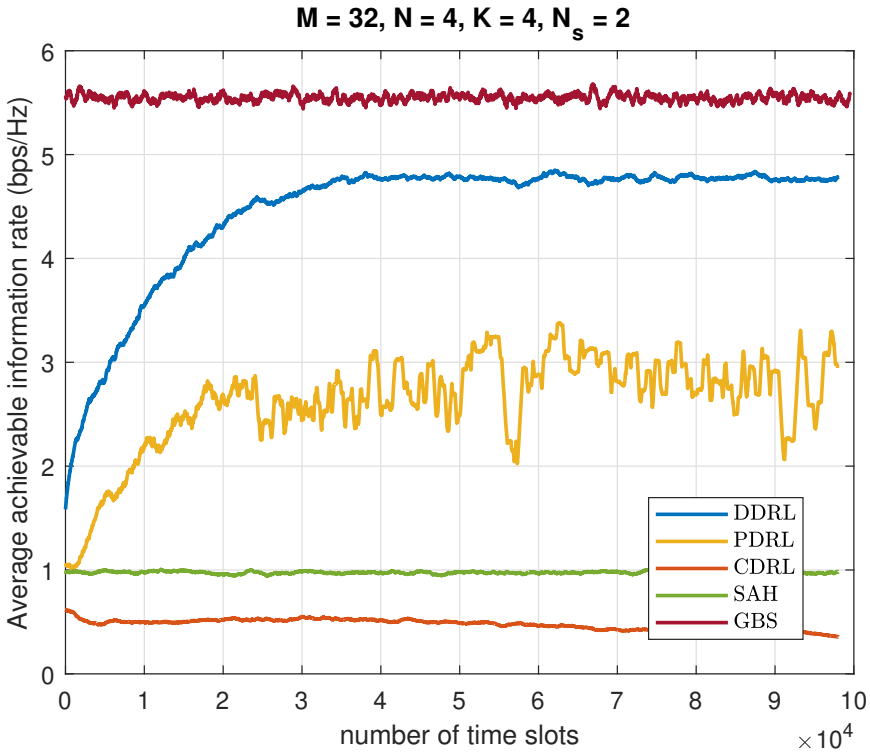



Figure 4.12: Average information rate versus number of time slots with different DRL-based schemes.

observed with larger $\alpha$, this is intuitive since the gradient descent is sped up with a larger value of loss function. The second observation is that, compared with (256, 256), a reward degradation appears with (32, 32), which suggests that increasing the DNN size demonstrates a stronger representation capability of input features and boosts the performance of the DRL-based scheme. Due to the negligible performance improvement in (512, 512), (256,256) is chosen as a baseline to maintain a balance between user connectivity and computational burden.

Fig. 4.14 offers an insight into the impact of different penalty values $\lambda$. This penalty term intrinsically represents an adjustment of reward function for each agent. Different from [18, 19], it is demonstrated in Fig. 4.14 that the system capacity is gradually increased with the penalty value from 0.1 to 5. An interpretation is that each agent causes high interference to other agents while still trying to maximize its information rate. Due

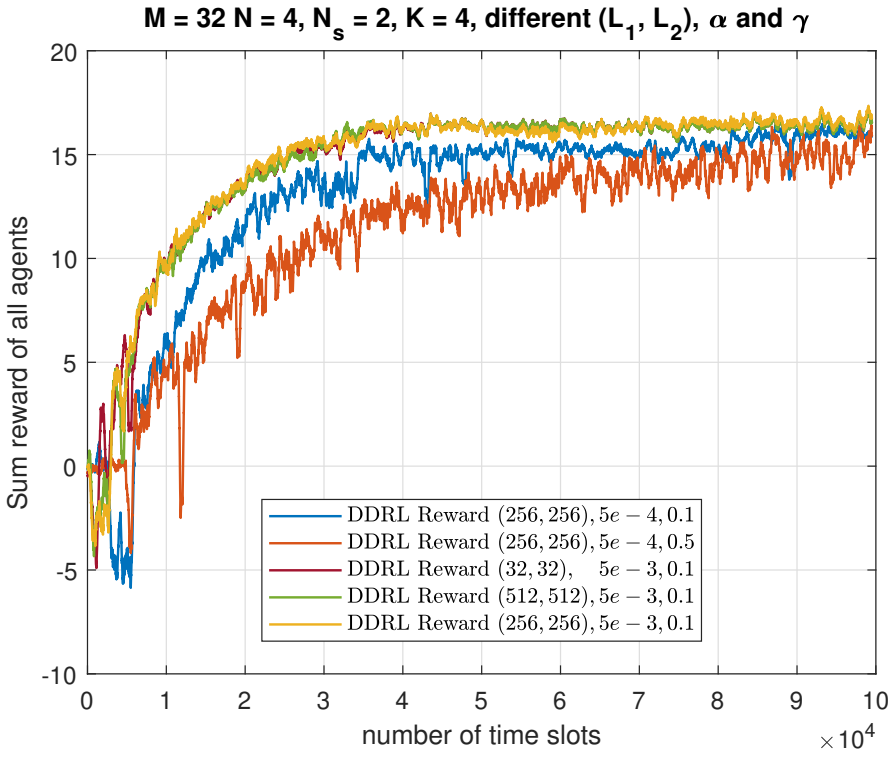



Figure 4.13: Sum reward versus the number of time slots with different number of users $K$.

to the uncertainty of the dynamic environment, the lack of perfect CSI introduces unpredictable interference for all the agents and an increase of penalty value can make a remedy for this by choosing an action that minimizes the interference to other agents instead of maximizing the received power of itself. This result also indicates that the decision-making process of all the agents is robust in unexpected high-interference scenarios.

The cumulative distribution function (CDF) over different DRL-based methods and benchmarks have been plotted in Fig. 4.15. It can be seen that the CDF curves confirm the discussion of the superiority of DDRL over other schemes. Also, the performance of DDRL is significantly limited by the codebook resolution which is confirmed in Fig. 4.16 where we show the boxplot of different methods.

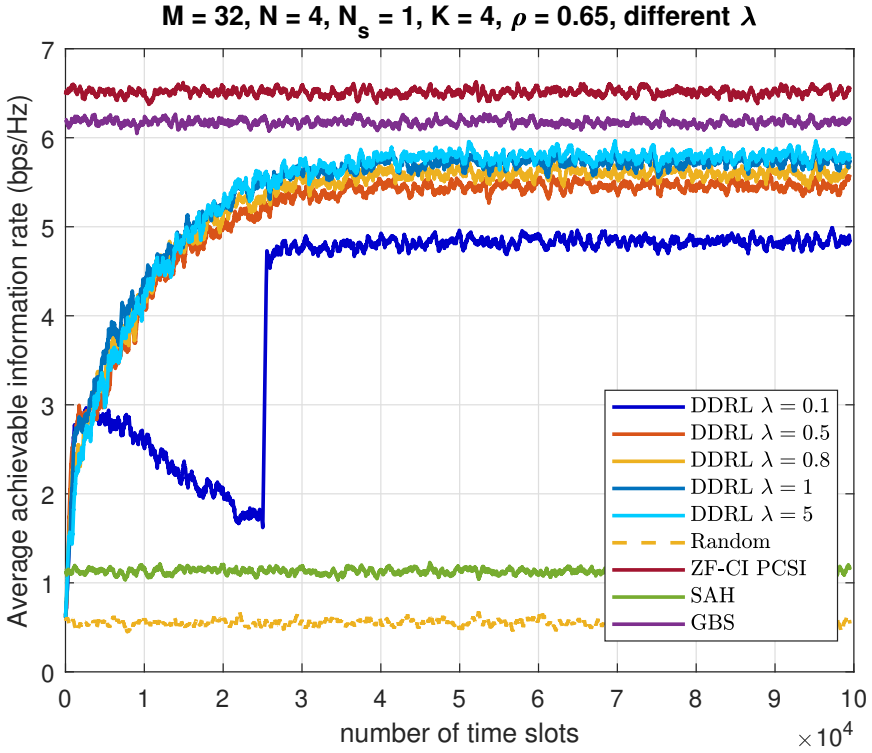



Figure 4.14: Average information rate versus the number of time slots with different penalty $\lambda$.

## 4.5 Summary

In this chapter, we studied the beamforming optimization for massive MIMO downlink transmission with channel aging. An optimization framework in light of DRL was studied and three DRL-based algorithms were derived based on stream-level, user-level, and system-level agent modeling. Specifically, the transmit precoder at BS and receive combiner at user terminals were jointly optimized to maximize the average information rate. Furthermore, we analyzed the performance loss of DRL-based approaches as compared to the ideal case with continuous beamforming with different numbers of codebook sizes, users, antennas, streams, and user speeds. Interestingly, it was shown that even using a designed codebook in DDRL is still able to achieve 95% and 90% as in the case with GBS and ZF-CI, respectively. Simulation results showed that significant

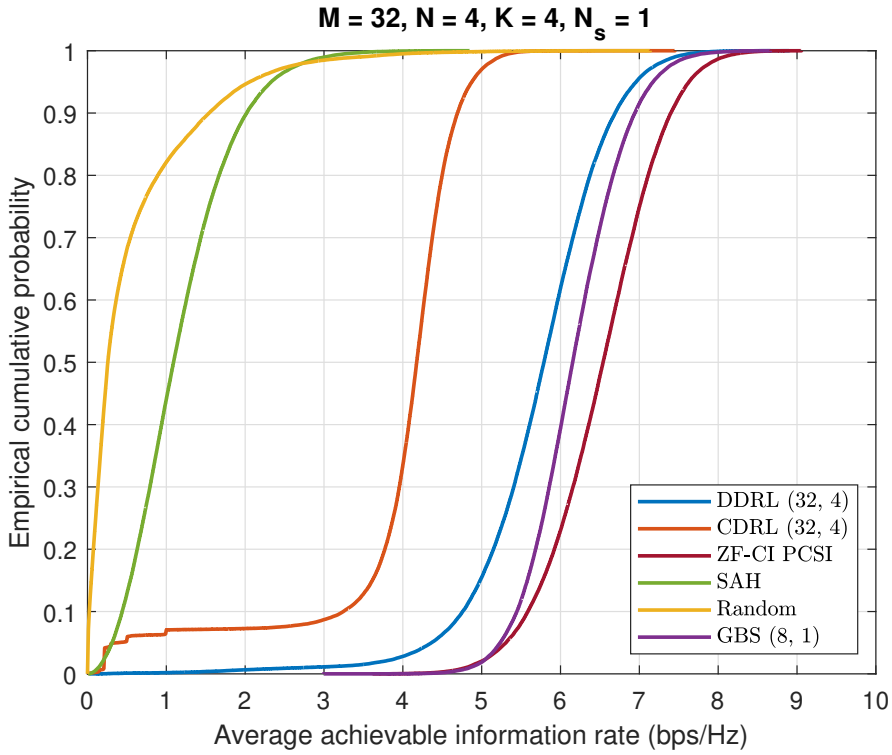



Figure 4.15: CDF of the average information rate over different DRL-based methods and benchmarks.

robustness on user mobility can be achieved by using some received power values of imperfect CSIT at the expense of more uplink overhead. Also, the convergence speed and scalability of the proposed algorithms are discussed. The convergence speed is linearly increased with the number of transmit antennas and performance degradation in the multiuser case is non-negligible due to the severe interference. In addition, CDRL consumes less computation complexity but demonstrates instability and incurs performance loss. In contrast, DDRL offers a promising gain over CDRL by favoring an adaptive decision-making process and facilitating cooperation among all agents to mitigate interference but incurs a higher hardware complexity and non-stationarity. Finally, the reward, penalty analysis, and statistical test confirm the fact that the performance of the proposed algorithms is greatly limited by the resolution of codebooks.

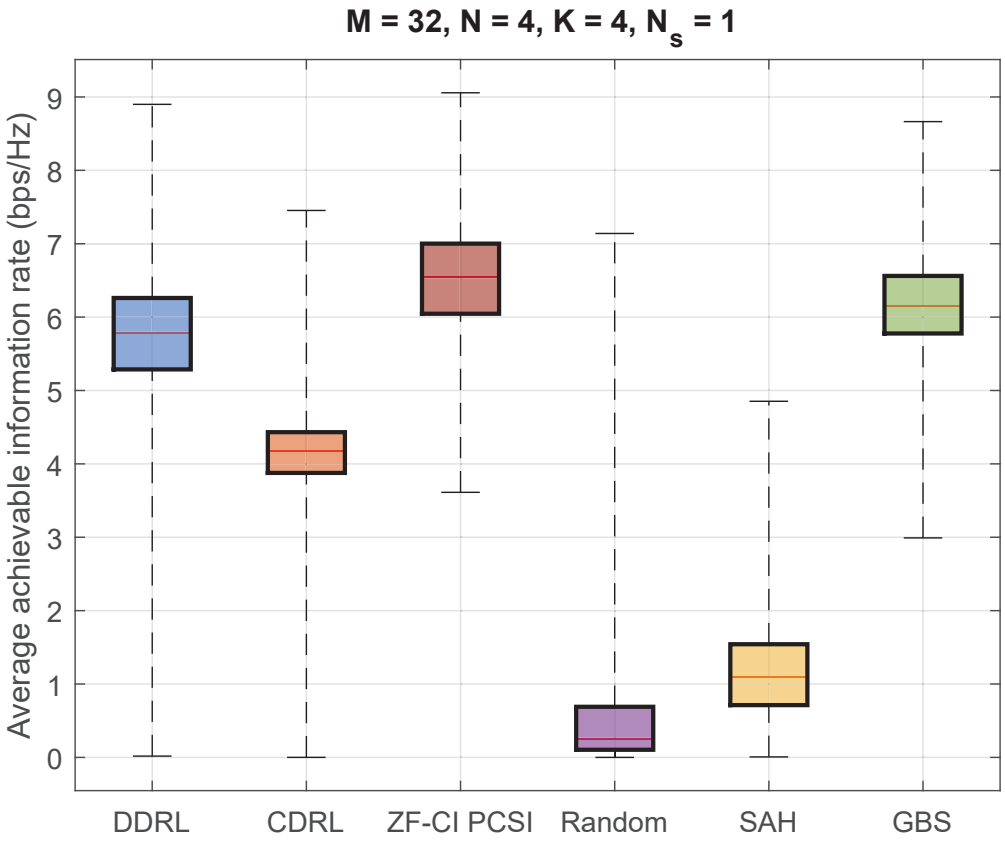



Figure 4.16: Boxplot of the average information rate over different DRL-based methods and benchmarks. The bottom and top of each box are the 25th and 75th percentiles of the information rate values, respectively. The whisker length is set to infinity to ensure there are no outliers.

# Chapter 5

# A Novel Multi-agent DDPG-based Beamforming Design for Massive MIMO Wireless Communication Networks

This chapter introduces a general DDPG-assisted massive MIMO architecture. To design appropriate beamforming, we investigate the DDPG-based beamfoming design which jointly optimize the transmit precoder and receive combiner to maximize the sum rate of all users subject to power constraint. The superiority of DDPG-based framework is verified through simulation results in different scenarios. DDPG is demonstrated to be a powerful reinforcement learning tool for beamforming, and provides a better convergence speed and rate performance compared with that in Chapter 4.

## 5.1 Introduction

As is shown in Fig. 4.10 in Chapter 4, the proposed algorithm is greatly limited by the resolution of the codebooks. An interpretation is that the codebook-based strategies suffer from large quantization error and the rich interference is not properly suppressed due to the constraint of codebook precision and CSIT imperfections. Also, for the DQN method, there are some hurdles for the application. The amount of computation rapidly increases as the number of actions or states increases due to the burden of Q-value calculation. Besides, it can only deal with discrete actions, not with continuous ones. In value-based methods such as DQN, values must be discretized to the handle continuous action, but this method has some limitations. When the action space is discretized, the action space increases exponentially, making learning almost impossible, and since the optimal action can also be removed in the discretization process, it is difficult to find the desired action.

Inspired by Fig. 4.5, the discussion on the codebook size's effect on the optimality of results and convergence rate presents an opportunity to explore potentially more robust solutions, possible through approaches with continuous action spaces. The research on multi-antenna beamforming with continuous action space is still in the early stage [23, 128–130]. On the one hand, [128–130] are based on perfect CSIT which is usually not available. On the other hand, the simulation scenarios are simple, e.g., the number of transmit antennas is 4 in [128, 130] and 8 in [23].

Motivated by the issues above, we study the joint transmit precoder and receive combiner design in massive MIMO downlink transmission with channel aging using DDPG. The contributions of this work are summarized as follows

- We construct an efficient multi-agent DDPG-based framework for massive MIMO downlink transmission with channel aging. Compared with the strategies in Chapter 4, this work shows that the DDPG-based framework demonstrates 1)

better interference management especially multi-user interference for a large number of transmit antennas, 2) faster convergence, 3) lower uplink overhead, and 4) comparable computation complexity due to the simplicity of the state space.

- Leveraging the DDPG-based framework mentioned above, the precoders and combiners are jointly optimized with new state, action, and reward functions. The benefits of this design are shown through simulations by benchmarking the DDRL and ZF-CI PCSI in Chapter 4.

- To have a fair comparison, we study the pros and cons of DDPG-based strategy over the benchmarks above, especially the DDRL benchmark. The proposed algorithm 1) overcomes the hurdles of Q-value calculation in DDRL, 2) gets rid of precoder/combiner selection which greatly increases the rate performance in the scenarios of large number of transmit antennas. 3) suffers from harder parameter tuning and unstable training.

## 5.2 Multi-agent DRL-based Continuous Beamforming Optimization

In this Chapter, in order to have a fair comparison, we are still using the same system models in Chapter 4 to demonstrate the superiority of DDPG-based method over DQN-based method. To build up the foundation for the proposed DRL-based designs, an overview of DDPG is illustrated first, followed by the description of the state, action, reward function and multi-agent DRL-based algorithms for the problem (4.14).

### 5.2.1 A Brief Overview of DDPG

DDPG is a model-free actor-critic learning method [107]. Different from DQN, there are two groups of networks, i.e, the actor network and the critic network. The trained critic network and the trained actor network can be expressed as:

$$Q_c(s_t,a_t) = \mathcal{E}_\pi(R_t|s_t = s,a_t = a), \tag{5.1}$$

$$\mu_a(s) = \mathcal{E}_\pi(R_t|s_t = s), \tag{5.2}$$

where the definition of $R$, $s$ and $a$ can be found in Chapter 2.5.2. In addition, time delayed copy of the trained critic $Q_c(s_t,a_t|\theta_c)$ and the trained actor network $\mu_a(s_t|\theta_a)$ are defined as $Q'_c(s_t,a_t|\theta_{c'})$ and $\mu'_a(s_t|\theta_{a'})$ , respectively. Those networks are also called target critic network and target actor network, respectively.

#### Critic and Actor Networks

The main objective of critic network is to evaluate how good a policy is. The input of the critic network are the current environment state and actions generated by the actor network, and outputs the Q-function based on the DDPG. The activation function is set as *relu* for faster convergence. The update of training critic network is

$$\theta_c^{t+1} = \theta_c^t - \alpha_c \Delta_{\theta_c} L(\theta_c) \tag{5.3}$$

$$L(\theta_c) = (r^t + \beta q(\theta_{c'}|s^{t+1},a') - q(\theta_c|s^t,a^t))^2 \tag{5.4}$$

where $\alpha_c$ and $a'$ denote the learning rate and the action coming from the target actor network respectively. $\Delta_{\theta_c} L(\theta_c)$ represents the gradient of training network. The $\theta_c$ and $\theta'_c$ represent the policy parameters of training and the target critic networks respectively. The $q(\cdot)$ is the approximation of Q-function, usually using deep neural networks.

The function of the actor network is to learn the current environment based on the proposed DDPG algorithm and output the actions to the critic network. The update of the training actor network based on the DDPG policy can be written as

$$\theta_a^{t+1} = \theta_a^t - \alpha_\mu \Delta_a q(\theta_{c'}|s^t,a) \Delta_{\theta_a} \pi(\theta_a|s^t) \tag{5.5}$$

where $\alpha_\mu$ is the learning rate of actor network, and $\pi(\theta_a|s^t)$ denotes the training actor policy at state $s$. The output of the actor network is the action, which is usually a continuous value. Since we do not have e-greedy or other action selection strategies in DDPG, that's why we have the concept of "deterministic" in this algorithm. The gradient of target critic network is denoted by $\Delta_a q(\theta_{c'}|s^t,a)$ with the parameter of $\theta_{c'}$, while $\Delta_{\theta_a} \pi(\theta_a|s^t)$ denotes the gradient of training actor network with the parameter of $\theta_a$. From above the equation, the gradients of target critic network and training actor network both have impact on the updating of training actor network, which also make sure that the actor-critic networks is in right direction to converge for learning the continuous-valued spaces.

Finally, the target actor and critic network can be updated using soft-update manner by the following equations,

$$\theta_{a'} \leftarrow \tau_a \theta_a + (1-\tau_a)\theta_{a'} \tag{5.6}$$

$$\theta_{c'} \leftarrow \tau_c \theta_c + (1-\tau_c)\theta_{c'} \tag{5.7}$$

where the updating rate of the target actor and critic network is denoted by $\tau_a$ and $\tau_c$, respectively. Softly updating the weights can enable a stable and smooth learning process. Note that our proposed DDPG architecture in Fig. 5.1 is different from traditional structure in [107] in several aspects. First, we do not add any noise to the generated action because we find that it is quite hard to control the generated beamformer when the number of transmit antennas becomes massive. Second, we

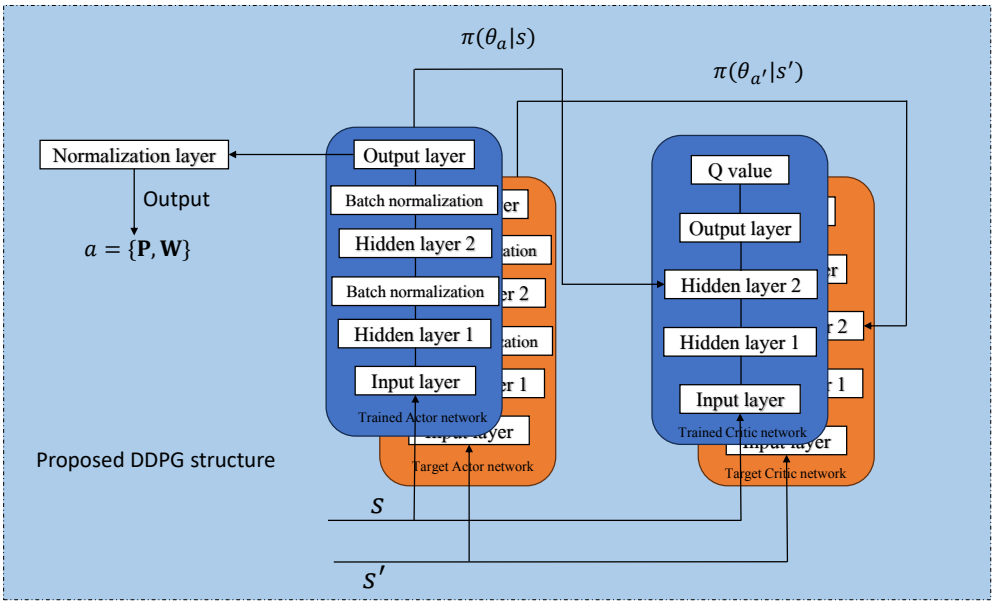



Figure 5.1: Proposed DDPG structure including the critic neural network and the actor neural network.

add a batch normalization layer for better convergence and faster training [131].

### 5.2.2 Proposed DRL Algorithm

In this section, we propose a multi-agent DDPG-based algorithms to solve it. The DDPG-based framework is discussed first, followed by the corresponding algorithm. To our best knowledge, this is the first work using DRL-based algorithm for continuous beamforming for massive MIMO [1] with 1) receive beamforming with multiple receive antennas, 2) multiple streams for each user 3) large number of transmit antennas with imperfect CSIT.

Each stream is modeled as an agent. At the initialization stage, all $KN_s$ pairs of actor and critic networks are established at the BS. For instance, one pair of actor networks,

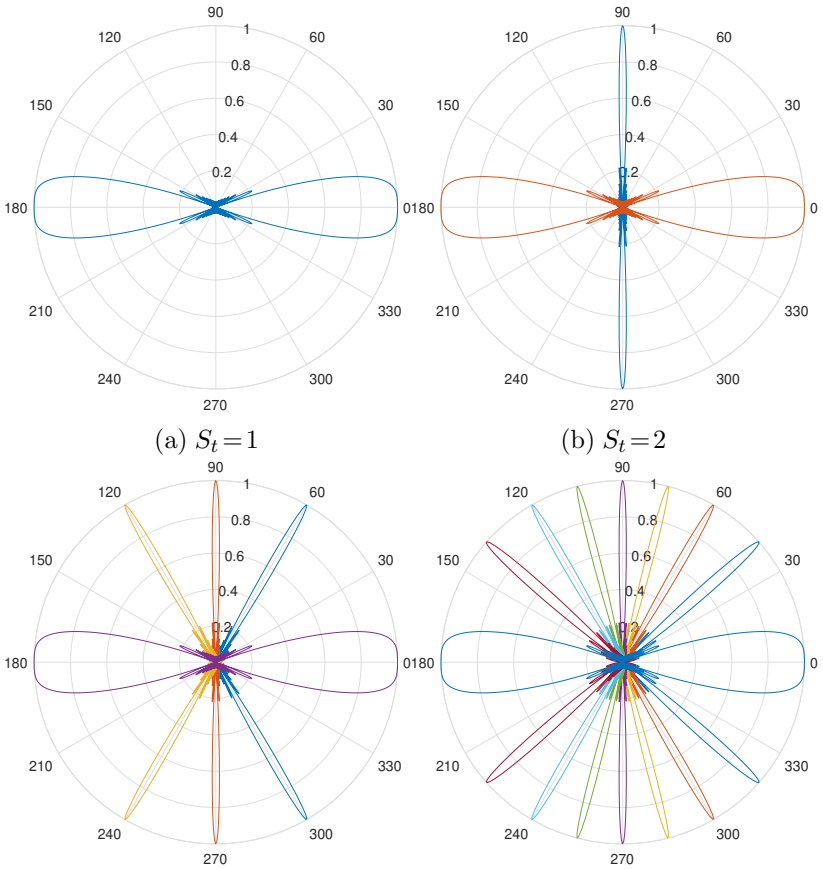
[1]There is no current standard for the least number of transmit antennas for massive MIMO. Our simulation is up to 512 transmit antennas, we believe it is a good number for massive MIMO.

namely trained actor network $q(\theta_{a,k,n}|s_{k,n})$ and target actor network $q(\theta_{a',k,n}|s_{k,n})$, and another pair of critic networks, trained critic network $q(\theta_{c,k,n}|s_{k,n},a_{k,n})$ and target critic network $q(\theta_{c',k,n}|s_{k,n},a_{k,n})$ are possessed by agent $(k,n)$. Note that the input of actor network is only the state but the input of the critic network is the combination of state and action generated from actor networks. $s_{k,n}$ and $a_{k,n}$ are local state and local action, respectively. Similar to that in Chapter 4, only imperfect CSIT information is used to formulate the observations $s_{k,n}$ at the beginning of each time slot. Then, the DDPG agent adopts a deterministic strategy to generate the action $a_{k,n}=\pi(\theta_a|s_{k,n})$ to enhance the exploitation by determining actions, i.e., the precoder $\mathbf{p}_{k,n}$ and combiner $\mathbf{w}_{k,n}$. Regarding the distributed learning process, the agent accumulates and stores the experience $e_{k,n}=\langle s_{k,n},a_{k,n},r_{k,n},s'_{k,n}\rangle$ into experience pool and the historical experiences can be utilized to train the DDPG agent with local state-action pairs together with the corresponding reward. Each agent has a partial view of the whole system which naturally leads to a distributed-learning-distributed-processing manner.

**Actions of the Proposed Multi-agent DDPG approach for massive MIMO scenarios**

Different from the codebook-based method in Chapter 4, we aim to optimize the beam directions. To enhance the understanding of beam direction, the beam patterns when $S_t=2,4,8$ are shown in Figs. 5.2 for $S_t=1,2,4,8$, respectively. As can be seen, the BS can determine its beamformer by selecting one from the codebook with a specified beam direction independently. For instance, when $S_t=4$, we can choose the direction with phases $0,\pi/3,\pi/2,2/3\pi,\pi,4/3\pi,3/2\pi,5/3\pi$. In DDPG-based strategy, the trained actor network generates continuous beam phases as is shown in Fig. 5.3.

However, there are some hurdles for the strategy. The amount of computation rapidly increases as the number of beamformers increases due to the burden of Q-value calculation. Besides, it can only deal with discrete phases, not with continuous phases.




(a) $S_t = 1$ (b) $S_t = 2$

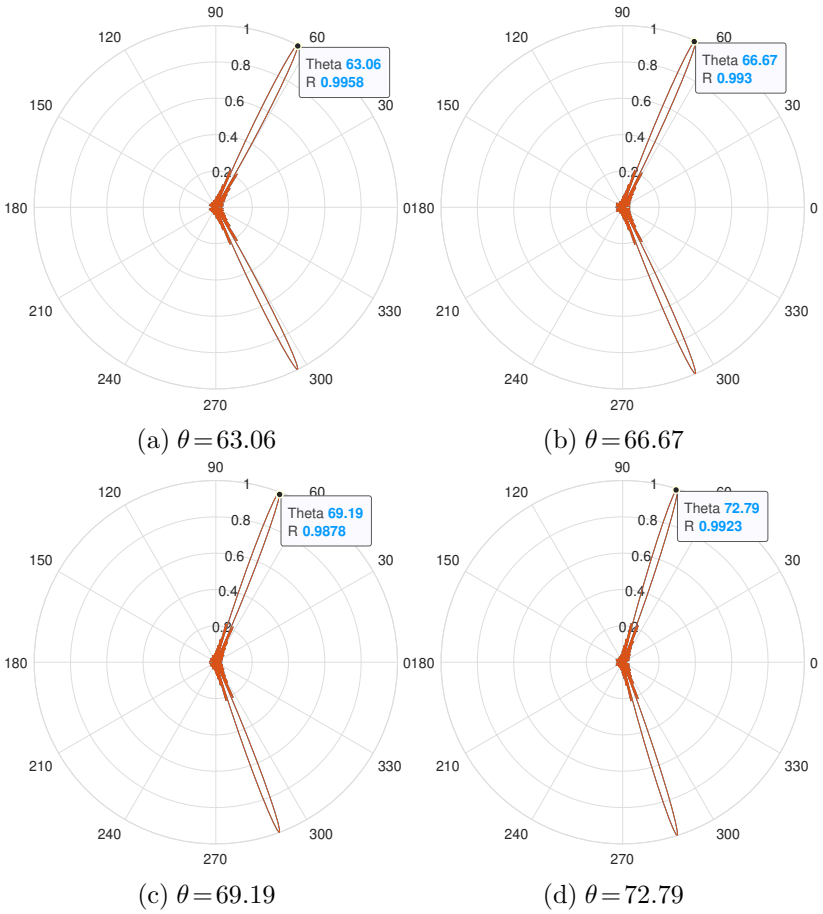
(c) $S_t = 4$ (d) $S_t = 8$

Figure 5.2: Set of beam directions with discrete phases when $S_t = 1,2,4,8$.

Inspired by the observation above, instead of optimizing the index of the beamformer in the codebook, we optimize the phases in a continuous manner. Simulation results confirm that the information rate is greatly improved with continuous phase optimization. The action space can be represented as

$$\mathcal{A} = \{(\theta_t, \theta_r), \theta_t, \theta_r \in [0.2\pi]\} \tag{5.8}$$

where $\theta_t$, and $\theta_r$ denote the phases of the transmit precoder and receive combiner, respectively.




(a) $\theta = 63.06$ (b) $\theta = 66.67$

(c) $\theta = 69.19$ (d) $\theta = 72.79$

Figure 5.3: Set of beam directions with continuous phases.

### States of the Proposed Approach

The state of agent $(k,n)$ is constructed by the representative feature of local observations. Different from that in Chapter 4, we find that the features from time slot $t-2$ don't contribute too much to the current reward function. An interpretation is that in Chapter 4, the features from time slot $t-2$ are used to learn the correlation coefficients $\rho$. However, the proposed DRL-based methods are robust to user mobility and naturally demonstrate robustness to $\rho$. Hence, we simplified the state information which experimentally performs well. The state $s_{k,n}(t)$ is defined as follows

- The local information of the agent $(k,n)$, which has 5 parameters, i.e., the

channel gain $|\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t-1)|^2$, the beam phase of precoder $\theta_{t,k,n}(t-1)$, the beam phase of combiner $\theta_{r,k,n}(t-1)$, the achievable rate of stream $n$ for user $k$, i.e., $G_{k,n}(\mathbf{P}_k(t-1),\mathbf{W}_k(t-1))$, and the interference-plus-noise $I_{k,n}(t-1)+I_{\mathrm{c},k}(t-1)+\sigma_k^2$.

- Interference information of the agent $(k,n)$ which is represented by 2 parameters, i.e., $\{\sum_{i=1,i\neq n}^{N_s}|\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,i}(t-1)|^2, \sum_{j\neq k}\sum_{i=1}^{N_s}|\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{j,i}(t-1)|^2\}$. It is worth noting here that in such a system, the interference information plays a key role in the maximization of its own information rate (the rate of stream $n$ of user $k$), which, thus, should be considered in state space.

- The information of agent $(j,i),(j,i)\neq(k,n),\forall j,i$ consists of of $5(KN-1)$ terms, i.e.,$\{\theta_{t,j,i}(t-1),\theta_{r,j,i}(t-1),G_{j,i}(\mathbf{M}_j(t-1),\frac{P}{NK}|\mathbf{w}_{j,i}^H(t-1)\mathbf{H}_j(t-1)\mathbf{p}_{j,i}(t-1)|^2,\frac{P}{NK}|\mathbf{w}_{j,i}^H(t-1)\mathbf{H}_j(t-1)\mathbf{p}_{k,n}(t-1)|^2\}$. The information of other agents plays an irreplaceable role for agent $k$ to minimize the interference it causes to them, which, thus, should be involved in state space.

To sum up, the cardinal number of state space is $5KN_s+2$ which is greatly reduced compared with that in Chapter 4. Note that the adopted design is not guaranteed to be the optimal one but empirically achieves a good performance.

### The Reward of the Proposed DRL-based Approach

Recall that in Chapter 4, in order to maximize the achievable rate, the instant rewards are affected by two main factors: the contributions to the information rate and the penalty term related to the stream interference. However, during the training in DDPG-based strategy, an observation is that if the reward of one action is always positive but not optimal, this action will have a higher possibility of being generated in the future training process. Then, the whole system converges to a local optimum due to less exploration. Note that due to $\epsilon$-greedy, this is not a problem for DQN-based

approaches in Chapter 4. However, it becomes serious in DDPG due to deterministic action. Hence, for better reward diversity, in this chapter, the reward is set as the gap between the stream rate and penalty achieved via training in the $t$th and $(t-1)$th time slot and is given as

$$\begin{aligned} r_{k,n}(t) = & G_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)) - G_{k,n}(\mathbf{W}_k(t-1),\mathbf{P}_k(t-1)) + \\ & \lambda P_{k,n}(\mathbf{W}_k(t-1),\mathbf{P}_k(t-1)) - \lambda P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t)) \end{aligned} \tag{5.9}$$

where $G_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$ and $\lambda P_{k,n}(\mathbf{W}_k(t),\mathbf{P}_k(t))$ are defined in (4.4) and (4.23), respectively. In this way, if the difference of stream rate and penalty increases after training, the agent will get a positive reward. Otherwise, the reward is negative. Thereby, all agent actions will be adjusted in a direction to achieve a positive reward to achieve the global optimization goal. Simulation results confirm that this form of reward function contributes to the convergence and performance of the DDPG algorithm empirically. The algorithm is shown in Algorithm 8.

Table 5.1: Comparison Between DDPG-based and DQN-based Algorithms

| | FDD & MA-DQN | FDD & MA-DDPG |
|---|---|---|
| CSI Overhead Uplink | $MN+KN+11$( $MN$ channel coefficients, $KN$ CSI pilot symbols, and $\{\lvert\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t-1)\rvert^2$, $G_{k,n}(\mathbf{P}_k(t-1),\mathbf{W}_k(t-1))$, $I_{k,n}(t-1)+I_{c,k}(t-1)+\sigma_k^2$, $\sum_{i=1,i\neq n}^{N_s}\lvert\mathbf{w}_{k,n}^H(t-v-u)\mathbf{H}_k(t-u)\mathbf{p}_{k,i}(t-v-u)\rvert^2$, $\sum_{j\neq k}\sum_{i=1}^{N_s}\lvert\mathbf{w}_{k,n}^H(t-v-u)\mathbf{H}_k(t-u)\mathbf{p}_{j,i}(t-v-u)\rvert^2$, $u\in\{1,2\},v\in\{0,1\}\}$ | $MN+KN+5$( $MN$ channel coefficients, $KN$ CSI pilot symbols, and $\{\lvert\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,n}(t-1)\rvert^2$, $G_{k,n}(\mathbf{P}_k(t-1),\mathbf{W}_k(t-1))$, $I_{k,n}(t-1)+I_{c,k}(t-1)+\sigma_k^2$, $\sum_{i=1,i\neq n}^{N_s}\lvert\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{k,i}(t-1)\rvert^2$, $\sum_{j\neq k}\sum_{i=1}^{N_s}\lvert\mathbf{w}_{k,n}^H(t-1)\mathbf{H}_k(t-1)\mathbf{p}_{j,i}(t-1)\rvert^2$ |
| CSI Overhead Downlink | $MN+2N$ ($MN$ channel coefficients and indexes of precoders and combiner $\mathbf{p}_{k,n}$ and $\mathbf{w}_{k,n}$) | $MN+2N$ ($MN$ channel coefficients and phases of precoders and combiner $\mathbf{p}_{k,n}$ and $\mathbf{w}_{k,n}$) |
| Computing Complexity of Precoding and Combining Matrices | $\mathcal{O}((10KN+3)L_1+L_1L_2+L_2S)$ | $\mathcal{O}((10KN+6)L_1+2L_1L_2+3L_2)$ |

### Discussion on the Overhead, Complexity for the DDPG-based Designs

In this section, we are about to discuss the complexity, and overhead of Algorithm 8 and the corresponding differences with DQN-based algorithms. The complexity

**Algorithm 8** DDPG-Based Beamforming Design Algorithm

**Initialize**: Randomly initialize the value network $Q_c(s_{k,n},a_{k,n}|\theta_{c,k,n})$ and actor network $\mu_a(s_{k,n}|\theta_{a,k,n})\forall k\in\{1,2,...,K\},\forall n\in\{1,2,...,N_s\}$, with weights $\theta_{a,k,n}$ and $\theta_{c,k,n}$. Initialize the target value network $Q_{c'}$ and $\mu_{a'}$ with weights $\theta_{a,k,n}=\theta_{a',k,n}$ and $\theta_{c,k,n}=\theta_{c',k,n}$.
In the first $E_s$ time slots, agent $(k,n)$ randomly selects an action from action space $\mathcal{A}$, and stores the corresponding experience $\langle s_{k,n},a_{k,n},r_{k,n},s'_{k,n}\rangle$ in its pool, $\forall k,n$.
**for** each time slot $t$ **do**
  **for** each agent $(k,n)$ **do**
    Obtain state $s_{k,n}$ from the observation of agent $(k,n)$.
    Choose the action $a_{k,n}=\mu(s_{k,n}|\theta_{a,k,n})$ according to the current policy $\forall k,n$.
    Agent $(k,n)$ executes action $a_{k,n}$ and observe reward $r_{k,n}$ observe new state $s'_{k,n}$.
    Agent $(k,n)$ puts experience $\langle s_{k,n},a_{k,n},r_{k,n},s'_{k,n}\rangle$ into experience pool $\mathcal{O}_{k,n}$.
    Randomly samples a minibatch with size $E_b$.
    Set $y_{k,n}=r_{k,n}+\gamma Q_{c'}(s_{k,n},\mu_{a'}(s'_{k,n}|\theta_{a',k,n})|\theta_{c',k,n})$.
    Update the critic by minimizing the loss:

$$L=\frac{1}{E_b}\sum_{i}^{E_b}(y_{k,n,i}-Q_c(s_{k,n,i},a_{k,n,i}|\theta_{c,k,n}))^2 \tag{5.10}$$

    Update the actor policy using the sampled policy gradient

$$\nabla_{\theta_{a,k,n}}J\approx\frac{1}{E_b}\sum_i\nabla_a Q_c(s_{k,n,i},a_{k,n,i}|\theta_{c,k,n})|_{s=s_{k,n,i},a=\mu_{a,k,n}(s_{k,n,i})}\nabla_{\theta_{a,k,n}}\mu_a(s_{k,n,i}|\theta_{a,k,n})|_{s_{k,n,i}} \tag{5.11}$$

    Update the target networks

$$\begin{aligned}\theta_{c',k,n}&=\tau_c\theta_{c,k,n}+(1-\tau_c)\theta_{c',k,n})\\ \theta_{a',k,n}&=\tau_a\theta_{a,k,n}+(1-\tau_a)\theta_{a',k,n})\end{aligned} \tag{5.12}$$

  **end for**
**end for**

analysis of the DDPG algorithm depends on two aspects: inference floating operations per second (FLOPS) and the convergence rate. The number of FLOPS during the inference is mainly determined by the structure of the policy network and the value network. When we set $S,L_1,L_2$ as the numbers of elements in state, the first hidden layer, the second hidden layer in policy and value network, the inference FLOPS in

the policy network and the value network can be computed as

$$\begin{aligned}
&\text{FLOPS}_{\text{a}} = (5KN_s+2)L_1 + L_1L_2 + 2L_2, \\
&\text{FLOPS}_{\text{c}} = (5KN_s+4)L_1 + L_1L_2 + L_2, \\
&\text{FLOPS}_{\text{total}} = (10KN_s+6)L1 + 2L_1L_2 + 3L_2.
\end{aligned} \tag{5.13}$$

It can be seen that the complexity of the DDPG-based strategy is not strictly higher than that of DQN-based algorithms, which depends on the value of $S, L_1, L_2$. Due to the simplicity on the state space, the uplink overhead of the proposed approach is lower than that in DQN-based strategies. The downlink overhead remains the same.

**Another Reward Function**

Recall that in function (5.9), we set the reward as the gap between the stream rate and penalty achieved via training in the $t$th and $(t-1)$th time slot. If we substitute $r'(t) = G_{k,n}(\mathbf{W}_k(t), \mathbf{P}_k(t)) - G_{k,n}(\mathbf{W}_k(t-1), \mathbf{P}_k(t-1))$. Then, $r_{k,n}(t) = r'(t) - r'(t-1)$. Hence, the final reward we would like to maximize is

$$r_{final} = r'(N) - r'(N-1) + r'(N-1) - r'(N-2) + \cdots + r'(2) - r'(1) \approx r'(N) \tag{5.14}$$

However, the purpose of reinforcement learning is to maximize the accumulative reward, say, $\sum_{t=0} r_{k,n}(t)$ and $r_{final}$ is also maximized to some extend.

Here, we propose a new reward function focusing on accumulative reward maximization. The reward function is shown below

$$\begin{aligned}
r_{k,n}(t) &= 2G_{k,n}(\mathbf{W}_k(t), \mathbf{P}_k(t)) - 2G_{k,n}(\mathbf{W}_k(t-1), \mathbf{P}_k(t-1)) + \\
&\lambda P_{k,n}(\mathbf{W}_k(t-1), \mathbf{P}_k(t-1)) - \lambda P_{k,n}(\mathbf{W}_k(t), \mathbf{P}_k(t)) \\
&= 2r'(t) - r'(t-1)
\end{aligned} \tag{5.15}$$

Then, final reward $r_{final}$ is

$$r_{final} = 2r'(N) - r'(N-1) + 2r'(N-1) - r'(N-2) + \cdots + 2r'(2) - r'(1) \approx \sum_{n=0} r'(n) \tag{5.16}$$

which is inline with the format of reinforcement learning reward function. The performance of this reward function is shown in Section 5.3.

## 5.3 Simulation Result and Analysis

This section demonstrates the performance of our proposed multi-agent DDPG-based algorithm to maximize the average throughput of all the users. In order to have a fair comparison with the algorithms in Chapter 4, the simulation setup remains the same as that in Chapter 4. We will demonstrate the simulation results in different scenarios.

### 5.3.1 Single User

Fig. 5.4 and Fig. 5.5 illustrate the DDPG-based scheme with different numbers of transmit antennas $M$ when $N=4, N_s=1, K=1$ and $N=4, N_s=2, K=1$, respectively. A first observation is that a near-optimal result can be observed by leveraging the proposed state, action, and reward design in an interference-free scenario. A second observation is that DDPG-based approach converges much faster than DDRL. An interpretation is that DDPG uses a deterministic policy, meaning that for a given state, it directly outputs a specific action. This can simplify the learning process. In contrast, DDRL utilizes $\epsilon$-greedy strategy over actions, which can introduce additional complexity and suboptimal beamformer selection. A third observation is that DDPG always outperforms DDRL and the performance gap generally increases with the number of transmit antennas. Intuitively, the reason for this effect is that, in DDRL, the optimal beamformer can be removed in the design of codebook. This impact is

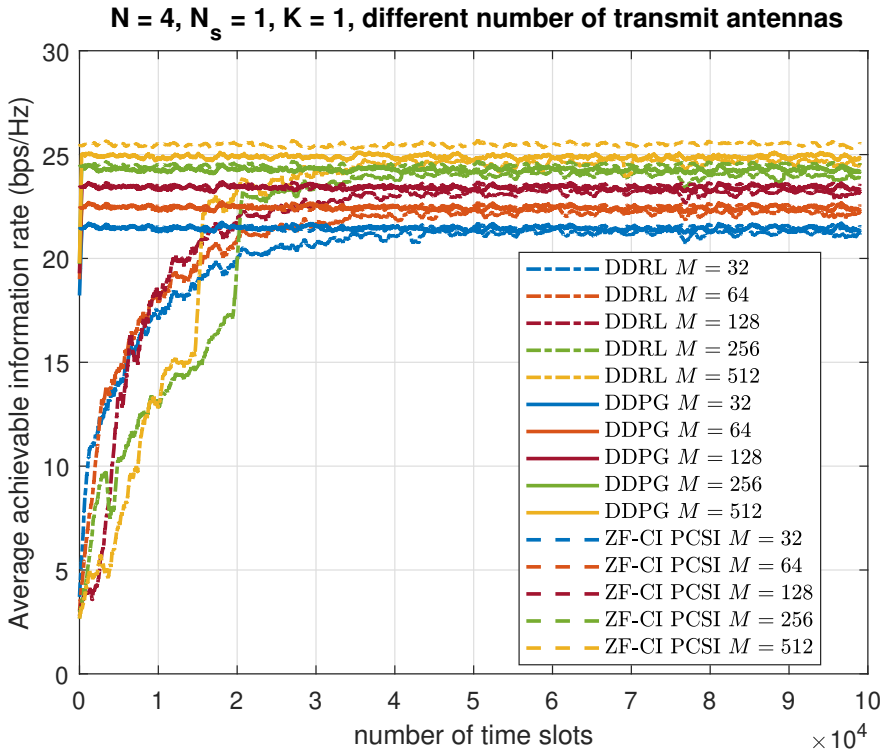



Figure 5.4: Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N=4,N_s=1,K=1$.

greatly reduced due to continuous beam direction optimization in DDPG. Although DDPG-based approach can produce very good results, it has its drawbacks. The performance of DDPG is heavily dependent on searching for the correct hyperparameters for the current task and generates a small gain if the number of antennas is small.

### 5.3.2 Multi User

Fig. 5.6 and Fig. 5.7 offers an insight into the impact of average information rate with different numbers of transmit antennas $M$ when $N=4,N_s=1,K=2$ and its corresponding boxplots. In addition to better performance and convergence, it turns out that one of the biggest drawbacks of DDPG is the high variance in estimating the gradient of $\mathcal{E}[R_t]$. Essentially, each time we perform a gradient update, we are using an estimation

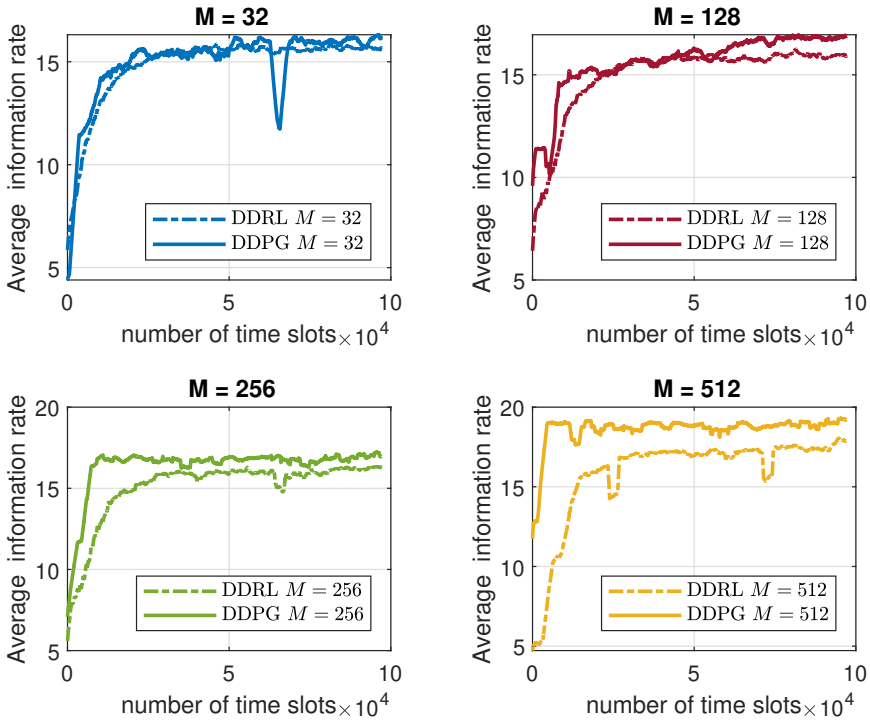



Figure 5.5: Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N=4, N_s=2, K=1$.

of gradient generated by a series of experiences $e_{k,n}=\langle s_{k,n}, a_{k,n}, r_{k,n}, s'_{k,n}\rangle$ accumulated through the whole training process. Hence, the estimation can be very noisy, and bad gradient estimate could adversely impact the stability of the learning algorithm. In contrast, when DQN does work in a task, it usually shows more stable performance since it is just a single model instead of an actor-critic structure in DDPG [107].

### 5.3.3 Different Reward Function

Fig. 5.8 illustrate the DDPG-based scheme with different numbers of transmit antennas $M$ and different reward functions when $N=4, N_s=1, K=2$. The legend $r1$ refers to the reward function in Section 5.2.2.3 and $r2$ refers to that in Section 5.2.2.4. An observation is that the DDPG with two different reward function shows similar

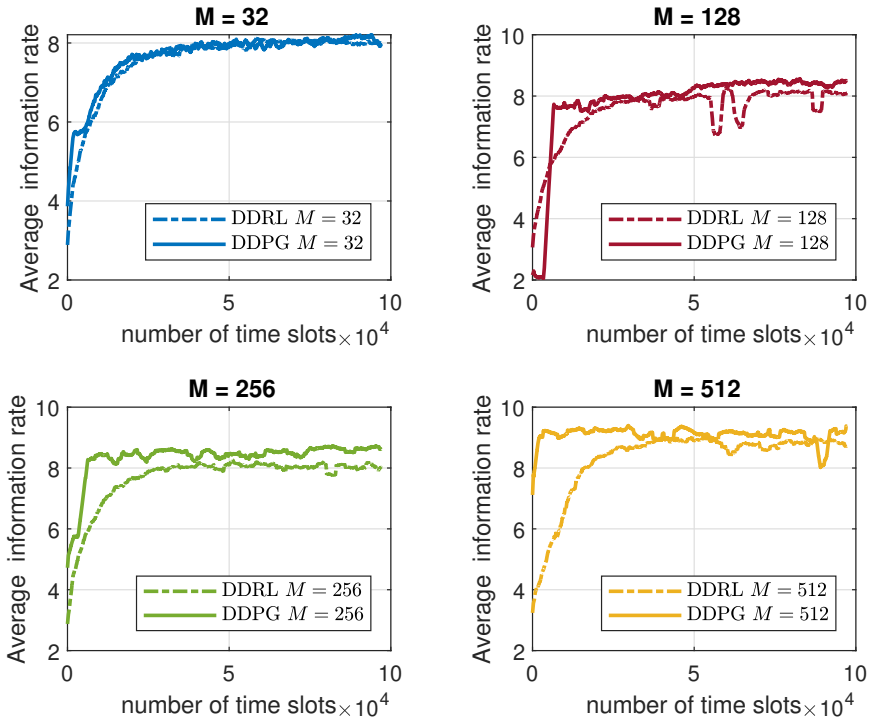



Figure 5.6: Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N=4, N_s=1, K=2$.

performance when the number of transmit antennas is smaller than 256 and the second reward function has worse performance when $M=512$. This suggests that the new reward function is not suitable for large number of transmit antennas scenarios. Another observation is that DDPG-$r2$ always has better performance than DDRL which cross-validates the superiority of DDPG based methods over DQN based methods.

## 5.4 Summary

In this Chapter, we extend the scope of multi-user beamforming design from discrete action space design to continuous action space design. An optimization framework in light of multi-agent DDPG has been investigated. We start from the basics of DDPG and design the corresponding state, action, and reward function. Then, A

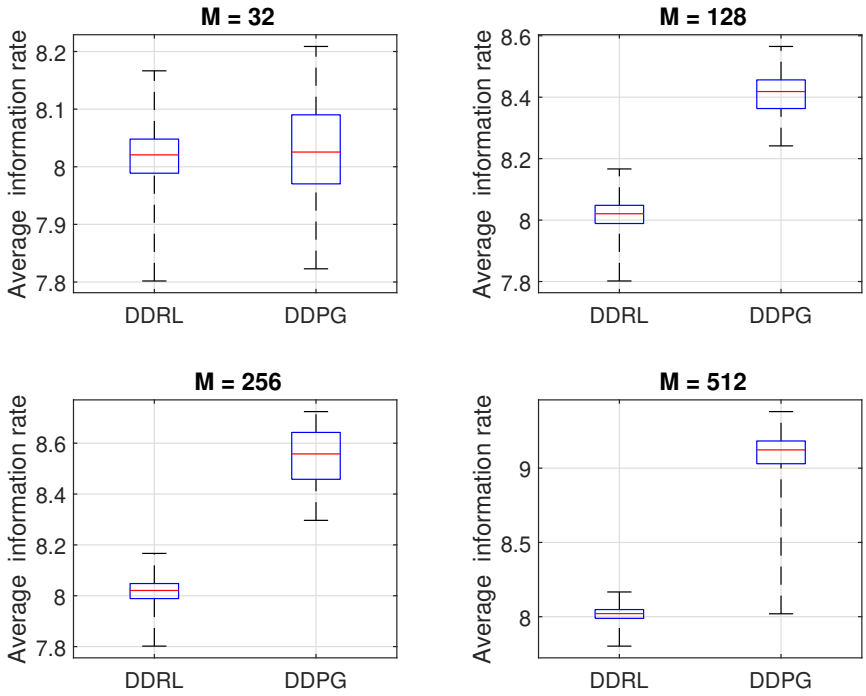



Figure 5.7: Boxplot of the average information rate over DDRL and DDPG with different numbers of transmit antennas $M$ when $N=4, N_s=1, K=2$. The bottom and top of each box are the 25th and 75t percentiles of the information rate values, respectively. The whisker length is set to infinity to ensure there are no outliers.

novel algorithm is elaborated which jointly optimizes the beamformer at BS and user terminals with channel aging. Through simulation results, the DDPG-based strategy is demonstrated to be a very powerful and promising technique for multi-user massive MIMO systems with faster convergence and better system performance.

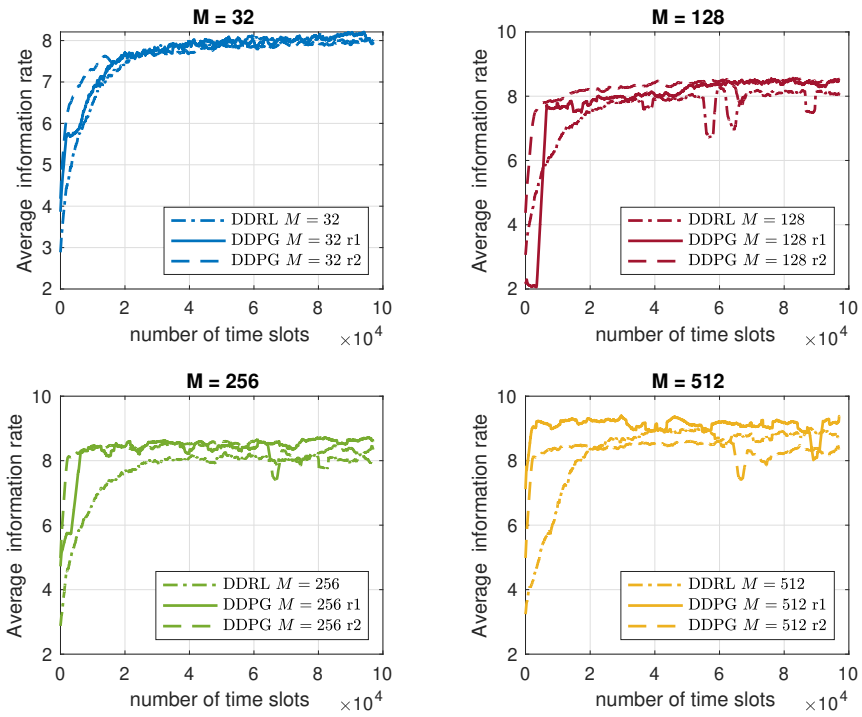



Figure 5.8: Average information rate versus the number of time slots with different numbers of transmit antennas $M$ when $N=4, N_s=1, K=2$ with different reward function.

# Chapter 6

# Conlusion

## 6.1 Summary of Thesis Achievements

In this thesis, we consider the beamforming design problems of multi-antenna wireless networks, and addressed a number of optimization problems and algorithms in various scenarios, namely waveform and beamforming optimization for IRS-aided WPT and DRL-based beamforming design in massive MIMO with channel aging. Simulation results and analysis are presented to evaluate the performance of all the proposed algorithms.

In Chapter 3, we have explored the design of joint waveform and passive beamforming for wireless power transfer assisted by IRS, considering both single-user and multi-user scenarios. Two distinct formulations were created based on FS-IRS and FF-IRS. We investigated an optimization framework employing AO and SCA, demonstrating its robustness and fast convergence. The numerical outcomes underscored the limitations of linear models and emphasized that even FF-IRS-aided WPT outperforms traditional designs with IRS passive beamforming gain. Additionally, FS-IRS served as a benchmark for broadband systems, while FF-IRS proved suitable for narrowband trans-

mission. Furthermore, in WPT systems, IRS not only extended the operational range under single-user conditions but also expanded the output DC region in multi-user scenarios. Moreover, the AO framework can be extended to explore transmit beamforming gains in multi-antenna scenarios. Finally, leveraging low-resolution discrete-phase IRS yielded near-optimal results for both single-user and multi-user scenarios.

In Chapter 4, we have investigated beamforming optimization for massive MIMO downlink transmission in the presence of channel aging. Employing a DRL framework, we developed three DRL-based algorithms that modeled agents at the stream-level, user-level, and system-level, respectivly. Specifically, we have jointly optimized the transmit precoder at the BS and the receive combiner at user terminals to maximize the average information rate. Additionally, we analyzed the performance degradation of DRL-based approaches compared to the ideal continuous beamforming case, considering various factors such as codebook sizes, users, antennas, streams, and user speeds. Notably, even with a very low-resolution codebook, the DRL approach demonstrated a good performance comparable to GBS and ZF-CI, respectively.

In Chapter 5, motivated by the discussion on the codebook size's effect presented in Chapter 4, we further investigated the multi-user beamforming through an approach with continuous action spaces for beamforming in DRL. An algorithm based on a multi-agent DDPG-based approach is elaborated, simulated, and analyzed. Simulation results demonstrated the benefits of DDPG for faster convergence and better system performance compared with the strategies in Chapter 4.

## 6.2 Future Work

During the PhD journey, many problems were investigated and, while some of them could be solved, others still remain open today.

Regarding Chapter 3, we have studied the waveform and beamforming design for IRS-aided WPT, several important issues are not addressed

- *Max-Min Design:* In most of the related literature, IRS-aided WPT problems are modeled as sum power maximization without any fairness among user terminals. Despite max-min design for WPT without IRS has been done in [17], more work has to be done to develop IRS-aided WPT with max-min optimization in order to have a better fairness among users and stronger connectivity between BS and users. Thus, further work is needed to formulate the waveform design and passive beamforming for MU minimum $i_{\mathrm{dc}}$ maximization problem.

- *More Advanced IRS:* Beyond improving the performance, the architecture of IRS also plays a key role in passive beamforming design. Two beyond-diagnal IRS (BD-IRS, also called beyond-diagnal Reconfigurable intelligent surface(BD-RIS) architectures, namely the fully- and group-connected, have been proposed in [132]. More precisely, the scattering matrix $\mathbf{\Theta}$ of the fully-connected architecture is not limited to being diagonal with unit-modulus entries because of the presence of reconfigurable impedances connecting the ports to each other, which constitute an interesting future research direction.

- *Multi-user MIMO Design:* In addition to design efficient rectenna and waveform, using multiple rectennas, also known as multiport rectennas, at the receiver to form a MIMO WPT system can effectively increase the output DC power [51]. There are two strategies for rectenna combing, DC combining in which each receive antenna is connected to a rectifier so that the RF signal received by each antenna is individually rectified, and RF combining in which the received signals at all receive antennas are combined together to be rectified by a single rectenna. Vanilla IRS or more advanced IRS can be involved to better explore the passive beamforming gain in IRS-aided MIMO WPT system.

Regarding Chapter 4 and 5, our initial intention was to study the DRL for beamforming in massive MIMO with channel aging. However, this was posing very difficult challenges in the design of more complex scenarios. To summarize, our work leaves open a number of problems, and we summarize some of them below:

- *Hybrid-Field Beamforming:* The large number of antenna elements in massive MIMO poses major challenges, e.g., a large number of RF chains (one for each antenna element) increases cost, and energy consumption on determining the CSI between each transmit and receive antenna. A promising solution to these problems lies in the concept of hybrid precoding schema is promising to relieve the pressure of excessive power consumption by decomposing the high-dimensional full-digital precoder into the realization of an analog beamformer and digital precoder [133], which formulates an promising future research area.

- *Multi-cell*: In emerging and future wireless networks, inter-cell interference management is one of the key technological challenges as BSs' deployments become denser to meet ever-increasing demand.This thesis considered single-cell multiuser conditions. When it comes to a multi-cell scenario, a BS may increase its transmit power to improve its own data rate, but at the same time it may degrade links it interferes with, due to which the corresponding optimization problem is more challenging to be solved and thus is worthy of further investigation [19].

- *Extremely Large-scale MIMO (XL-MIMO)*: Compared with the conventional massive MIMO, the basic idea of the XL-MIMO is to deploy an extremely large (perhaps infinite) number of antennas in a compact space to obtain a much stronger beamforming gain to compensate for the severe path loss [134]. However, this poses great challenge for agents to learn to allocate physical layer resources and optimize the transmission strategy further. As such, it is worth comparing the proposed massive MIMO with the XL-MIMO in future investigations.

- *IRS-aided massive MIMO:* With the ability to dynamically reconfigure the electromagnetic environment, IRS can improve channel quality and overcome the limitations of the propagation environment. In the future, the evolution of IRS will play a key role for the future 6G wireless networks. Thus, further work is needed to model passive beamforming design for DRL-based massive MIMO systems.

# Appendix A

# Proof of Propositions

## A.1 Proof of Proposition 1

We first prove the convergence of problem (3.15). $f(\mathbf{d}_q)$ and $f(\mathbf{d}_q,\mathbf{d}_q^{(i-1)})$ are demonstrated below

$$f(\mathbf{d}_q) = \mathbf{d}_q^H \mathbf{K}_0 \mathbf{d}_q, \tag{A.1}$$

$$f(\mathbf{d}_q,\mathbf{d}_q^{(i-1)}) = 2\Re\{\mathbf{d}_q^{(i-1)H}\mathbf{K}_0\mathbf{d}_q\} - \mathbf{d}_q^{(i-1)H}\mathbf{K}_0\mathbf{d}_q^{(i-1)}. \tag{A.2}$$

According to [104], it can be checked that $f(\mathbf{d}_q^{(i)},\mathbf{d}_q^{(i)}) = f(\mathbf{d}_q^{(i)})$. Since the solution vector $\mathbf{d}_q^{(i-1)}$ is a feasible solution in iteration $i-1$, by using the inequality in Taylor expansion, we have $f(\mathbf{d}_q^{(i)}) \geq f(\mathbf{d}_q^{(i)},\mathbf{d}_q^{(i-1)})$ which proves that $f(\mathbf{d}_q^{(i)})$ is monotonically increasing. Additionally, problem (3.15) must have a upper bound due to the unit modulus constraint (3.15b). Hence, problem (3.15) is guaranteed to converge.

Now, we prove that the convergent solution is a KKT solution of problem (3.12). (3.15a) and (3.15b) are convex problems with respect to $\mathbf{X}$ and satisfies the slater's condition [101], the dual gap is zero and strong duality hold. The optimal solution $\mathbf{X}^\star$ can be get by

figuring out its dual problem. The corresponding Lagrange function of problem (3.15) is

$$\mathcal{L}(\mathbf{X},\boldsymbol{\tau}) = \mathrm{Tr}\{\mathbf{K}_1\mathbf{X}\} + \sum_{l'=1}^{L+1}\tau_{l'}(\mathbf{X}_{l',l'}-1) + \boldsymbol{\Upsilon}\mathbf{X} \tag{A.3}$$

where $\boldsymbol{\tau} = \{\tau_1,\tau_2,...,\tau_{L+1}\}$ and $\boldsymbol{\Upsilon}$ denote the vector and matrix dual variables of (3.15b) and (3.15c). There must be a $\boldsymbol{\tau}^\star = \{\tau_1^\star,\tau_2^\star,...,\tau_{L+1}^\star\}$ and $\boldsymbol{\Upsilon}^\star$ for guaranteeing that the corresponding KKT conditions are satisfied

$$\nabla_{\mathbf{X}^\star}\mathcal{L}(\mathbf{X},\boldsymbol{\tau})|_{\mathbf{X}=\mathbf{X}^\star} = \mathbf{0} \tag{A.4}$$

$$\tau_{l'}^\star(\mathbf{X}_{l',l'}^\star - 1) = 0, \forall l' \tag{A.5}$$

$$\boldsymbol{\Upsilon}^\star\mathbf{X}^\star = \mathbf{0} \tag{A.6}$$

The KKT conditions of problem (3.15) exactly consist of (A.4), (A.5) and (A.6). Hence, *Proposition* 1 holds.

## A.2 Proof of Proposition 2

Due to the equivalence between problem (3.14) and problem (3.15), (A.4)-(A.6) are also the KKT conditions of problem (15). From (3.13), we have

$$\nabla_{\mathbf{d}_q=\mathbf{d}_q^\star} f(\mathbf{d}_q)|_{\mathbf{d}_q=\mathbf{d}_q^\star} = \nabla_{\mathbf{d}_q=\mathbf{d}_q^\star} f(\mathbf{d}_q,\mathbf{d}_q^{(i-1)})|_{\mathbf{d}_q=\mathbf{d}_q^\star} \tag{A.7}$$

Since problem (3.12b) and problem (3.14b) exactly contain $f(\mathbf{d}_q)$ and $f(\mathbf{d}_q,\mathbf{d}_q^{(i-1)})$ at iteration $i$, respectively, except some constants, upon denoting the (3.12b) as $p(\mathbf{d}_q) = \sum_{q=1}^K \xi_q(-\frac{1}{2}k_2 d_{q,0} - f(\mathbf{d}_q))$ and (3.14b) as $p(\mathbf{d}_q,\mathbf{d}_q^{(i-1)}) = \sum_{q=1}^K \xi_q(-\frac{1}{2}k_2 d_{q,0} - f(\mathbf{d}_q,\mathbf{d}_q^{(i-1)}))$, we arrive at

$$\nabla_{\mathbf{d}_q=\mathbf{d}_q^\star} p(\mathbf{d}_q)|_{\mathbf{d}_q=\mathbf{d}_q^\star} = \nabla_{\mathbf{d}_q=\mathbf{d}_q^\star} p(\mathbf{d}_q,\mathbf{d}_q^{(i-1)})|_{\mathbf{d}_q=\mathbf{d}_q^\star}. \tag{A.8}$$

According to (A.7) and relation between $p(\mathbf{d}_q)$ and $p(\mathbf{d}_q,\mathbf{d}_q^{(i-1)})$, KKT conditions of problem (3.12) exactly consist of (A.4), (A.5) and (A.6) and $\mathbf{X}^\star$ is also a local optimal solution for problem (3.12). Combining the proof above and *Proposition* 4, a local optimal solution for passive beamforming subproblem and a global optimal solution for waveform design subproblem can be obtained. Hence, the AO Algorithm 2 converges to a local optimal solution in original problem (3.8) and *Proposition* 2 holds.

# References


[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *IEEE communications magazine*, vol. 52, no. 2, pp. 74–80, 2014.

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

[3] B. Clerckx, Y. Mao, Z. Yang, M. Chen, A. Alkhateeb, L. Liu, M. Qiu, J. Yuan, V. W. Wong, and J. Montojo, "Multiple access techniques for intelligent and multi-functional 6g: Tutorial, survey, and outlook," *arXiv preprint arXiv:2401.01433*, 2024.

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

[5] R. Zhang and C. K. Ho, "Mimo broadcasting for simultaneous wireless information and power transfer," *IEEE transactions on wireless communications*, vol. 12, no. 5, pp. 1989–2001, 2013.

[6] Z. Xiao, Z. Han, A. Nallanathan, O. A. Dobre, B. Clerckx, J. Choi, C. He, and W. Tong, "Antenna array enabled space/air/ground communications

and networking for 6g," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 10, pp. 2773–2804, 2022.

[7] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[8] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Journal on Selected Areas in Communications*, 2020.

[9] Q. Wu and R. Zhang, "Weighted sum power maximization for intelligent reflecting surface aided SWIPT," *IEEE Wireless Communications Letters*, 2019.

[10] Y. Tang, G. Ma, H. Xie, J. Xu, and X. Han, "Joint transmit and reflective beamforming design for irs-assisted multiuser miso swipt systems," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2020.

[11] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted swipt under qos constraints," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1735–1748, 2020.

[12] B. Clerckx, A. Costanzo, A. Georgiadis, and N. Borges Carvalho, "Toward 1G Mobile Power Networks: RF, Signal, and System Designs to Make Smart Objects Autonomous," *IEEE Microwave Magazine*, vol. 19, no. 6, pp. 69–82, 2018.

[13] Y. Zeng, B. Clerckx, and R. Zhang, "Communications and Signals Design for Wireless Power Transmission," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2264–2290, 2017.

[14] J. Kim, B. Clerckx, and P. D. Mitcheson, "Signal and System Design for Wireless Power Transfer : Prototype, Experiment and Validation," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[15] B. Clerckx, R. Zhang, R. Schober, D. W. K. Ng, D. I. Kim, and H. V. Poor, "Fundamentals of wireless information and power transfer: From RF energy harvester models to signal and system designs," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 1, pp. 4–33, 2018.

[16] B. Clerckx and E. Bayguzina, "Waveform design for wireless power transfer," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6313–6328, 2016.

[17] Y. Huang and B. Clerckx, "Large-scale multiantenna multisine wireless power transfer," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5812–5827, 2017.

[18] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, 2019.

[19] J. Ge, Y.-C. Liang, J. Joung, and S. Sun, "Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6070–6085, 2020.

[20] L. Zhang and Y.-C. Liang, "Deep Reinforcement Learning for Multi-Agent Power Control in Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2551–2564, 2020.

[21] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, 2020.

[22] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, 2020.

[23] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1663–1677, 2021.

[24] W. Li, W. Ni, H. Tian, and M. Hua, "Deep Reinforcement Learning for Energy-Efficient Beamforming Design in Cell-Free Networks," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6, IEEE, 2021.

[25] H. Chen, Z. Zheng, X. Liang, Y. Liu, and Y. Zhao, "Beamforming in Multi-User MISO Cellular Networks with Deep Reinforcement Learning," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pp. 1–5, IEEE, 2021.

[26] R. Zhang, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, "Joint Coordinated Beamforming and Power Splitting Ratio Optimization in MU-MISO SWIPT-Enabled HetNets: A Multi-Agent DDQN-Based Approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 2, pp. 677–693, 2021.

[27] Q. Hu, Y. Liu, Y. Cai, G. Yu, and Z. Ding, "Joint deep reinforcement learning and unfolding: Beam selection and precoding for mmWave multiuser MIMO with lens arrays," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2289–2304, 2021.

[28] H. Ren, C. Pan, L. Wang, W. Liu, Z. Kou, and K. Wang, "Long-Term CSI-based Design for RIS-Aided Multiuser MISO Systems Exploiting Deep Reinforcement Learning," *IEEE Communications Letters*, 2022.

[29] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 338–351, 2013.

[30] T. Ramya and S. Bhashyam, "Using delayed feedback for antenna selection in MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 6059–6067, 2009.

[31] M. Fozi, A. R. Sharafat, and M. Bennis, "Fast MIMO Beamforming via Deep Reinforcement Learning for High Mobility mmWave Connectivity," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 127–142, 2021.

[32] M. S. Trotter and G. D. Durgin, "Survey of range improvement of commercial rfid tags with power optimized waveforms," in *2010 IEEE International Conference on RFID (IEEE RFID 2010)*, pp. 195–202, IEEE, 2010.

[33] M. S. Trotter, J. D. Griffin, and G. D. Durgin, "Power-optimized waveforms for improving the range and reliability of rfid systems," in *2009 IEEE International Conference on RFID*, pp. 80–87, IEEE, 2009.

[34] A. Collado and A. Georgiadis, "Improving wireless power transmission efficiency using chaotic waveforms," in *2012 IEEE/MTT-S International Microwave Symposium Digest*, pp. 1–3, IEEE, 2012.

[35] F.-F. R. P. Harvesting, "Practical rectennas far-field rf power harvesting and transport," 2014.

[36] B. Clerckx, "Wireless information and power transfer: Nonlinearity, waveform design, and rate-energy tradeoff," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 847–862, 2017.

[37] Y. Zeng, B. Clerckx, and R. Zhang, "Communications and signals design for wireless power transmission," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2264–2290, 2017.

[38] A. Boaventura, A. Collado, N. B. Carvalho, and A. Georgiadis, "Optimum behavior: Wireless power transmission system design through behavioral models and efficient synthesis techniques," *IEEE Microwave Magazine*, vol. 14, no. 2, pp. 26–35, 2013.

[39] A. Costanzo and D. Masotti, "Smart solutions in smart spaces: Getting the most from far-field wireless power transfer," *IEEE Microwave Magazine*, vol. 17, no. 5, pp. 30–45, 2016.

[40] A. S. Boaventura and N. B. Carvalho, "Maximizing dc power in energy harvesting circuits using multisine excitation," in *2011 IEEE MTT-S International Microwave Symposium*, pp. 1–4, IEEE, 2011.

[41] H. Shoki, "Issues and initiatives for practical deployment of wireless power transfer technologies in japan," *Proceedings of the IEEE*, vol. 101, no. 6, pp. 1312–1320, 2013.

[42] I. Mayordomo, T. Dräger, P. Spies, J. Bernhard, and A. Pflaum, "An overview of technical challenges and advances of inductive wireless power transmission," *Proceedings of the IEEE*, vol. 101, no. 6, pp. 1302–1311, 2013.

[43] A. Collado and A. Georgiadis, "Optimal waveforms for efficient wireless power transmission," *IEEE Microwave and Wireless Components Letters*, vol. 24, no. 5, pp. 354–356, 2014.

[44] B. Clerckx and E. Bayguzina, "Low-complexity adaptive multisine waveform design for wireless power transfer," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 2207–2210, 2017.

[45] Y. Huang and B. Clerckx, "Waveform design for wireless power transfer with limited feedback," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 415–429, 2017.

[46] S. Shen, J. Kim, and B. Clerckx, "Closed-loop wireless power transfer with adaptive waveform and beamforming: Design, prototype, and experiment," *IEEE Journal of Microwaves*, vol. 3, no. 1, pp. 29–42, 2022.

[47] J. Xu and R. Zhang, "Energy beamforming with one-bit feedback," *IEEE Transactions on Signal Processing*, vol. 62, no. 20, pp. 5370–5381, 2014.

[48] M.-L. Ku, Y. Han, B. Wang, and K. R. Liu, "Joint power waveforming and beamforming for wireless power transfer," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6409–6422, 2017.

[49] D. Masotti, A. Costanzo, M. Del Prete, and V. Rizzoli, "Time-modulation of linear arrays for real-time reconfigurable wireless power transmission," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 2, pp. 331–342, 2016.

[50] S. Shen and B. Clerckx, "Beamforming Optimization for MIMO Wireless Power Transfer with Nonlinear Energy Harvesting: RF Combining versus DC Combining," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[51] S. Shen and B. Clerckx, "Joint waveform and beamforming optimization for mimo wireless power transfer," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5441–5455, 2021.

[52] J. Kim and B. Clerckx, "Range expansion for wireless power transfer using joint beamforming and waveform architecture: An experimental study in indoor environment," *IEEE Wireless Communications Letters*, vol. 10, no. 6, pp. 1237–1241, 2021.

[53] M. Varasteh, B. Rassouli, and B. Clerckx, "Wireless information and power transfer over an AWGN channel: Nonlinearity and asymmetric Gaussian signaling," in *2017 IEEE Information Theory Workshop (ITW)*, pp. 181–185, IEEE, 2017.

[54] M. Varasteh, B. Rassouli, and B. Clerckx, "On capacity-achieving distributions for complex AWGN channels under nonlinear power constraints and their applications to SWIPT," *IEEE Transactions on Information Theory*, 2020.

[55] B. Clerckx and J. Kim, "On the beneficial roles of fading and transmit diversity in wireless power transfer with nonlinear energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7731–7743, 2018.

[56] X. Tan, Z. Sun, J. M. Jornet, and D. Pados, "Increasing indoor spectrum sharing capacity using smart reflect-array," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2016.

[57] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: The potential of data transmission with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2746–2758, 2018.

[58] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[59] S. Lee, Y. Zeng, and R. Zhang, "Retrodirective multi-user wireless power transfer with massive MIMO," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 54–57, 2017.

[60] S. Shen, B. Clerckx, and R. Murch, "Modeling and architecture design of intelligent reflecting surfaces using scattering parameter network analysis," 2020.

[61] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 735–740, IEEE, 2019.

[62] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for intelligent reflecting surface enhanced wireless networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.

[63] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, L. Hanzo, and A. Nallanathan, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Transactions on Wireless Communications*, 2020.

[64] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 518–522, 2019.

[65] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDM: Power allocation and passive array optimization," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.

[66] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[67] H. Sung, S.-R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser mimo systems," *IEEE Transactions on Communications*, vol. 57, no. 11, pp. 3489–3499, 2009.

[68] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "The multicell multiuser mimo uplink with very large antenna arrays and a finite-dimensional channel," *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2350–2361, 2013.

[69] J. Wang, S. Jin, X. Gao, K.-K. Wong, and E. Au, "Statistical eigenmode-based sdma for two-user downlink," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5371–5383, 2012.

[70] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4418–4431, 2012.

[71] C. Hao and B. Clerckx, "Miso broadcast channel with imperfect and (un) matched csit in the frequency domain: Dof region and transmission strategies," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, 2013.

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

[73] A. M. Sayeed and V. Raghavan, "Maximizing mimo capacity in sparse multipath with reconfigurable antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 1, pp. 156–166, 2007.

[74] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE transactions on wireless communications*, vol. 14, no. 11, pp. 6481–6494, 2015.

[75] L. Sanguinetti, E. Björnson, and J. Hoydis, "Toward massive mimo 2.0: Understanding spatial correlation, interference suppression, and pilot contamination," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 232–257, 2019.

[76] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive mimo has unlimited capacity," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 574–590, 2017.

[77] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive mimo is a reality—what is next?: Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, 2019.

[78] T. L. Marzetta, "Massive mimo: an introduction," *Bell Labs Technical Journal*, vol. 20, pp. 11–22, 2015.

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

[80] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive mimo: Ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, 2016.

[81] W. Ding, F. Yang, C. Pan, L. Dai, and J. Song, "Compressive sensing based channel estimation for OFDM systems under long delay channels," *IEEE Transactions on Broadcasting*, vol. 60, no. 2, pp. 313–321, 2014.

[82] J. Meng, W. Yin, Y. Li, N. T. Nguyen, and Z. Han, "Compressive sensing based high-resolution channel estimation for OFDM system," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 1, pp. 15–25, 2011.

[83] A. K. Papazafeiropoulos, "Impact of general channel aging conditions on the downlink performance of massive MIMO," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 2, pp. 1428–1442, 2016.

[84] N. Lee and R. W. Heath, "Space-time interference alignment and degree-of-freedom regions for the MISO broadcast channel with periodic CSI feedback," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 515–528, 2013.

[85] N. Lee, R. Tandon, and R. W. Heath, "Distributed space–time interference alignment with moderately delayed CSIT," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1048–1059, 2014.

[86] H. Yin, H. Wang, Y. Liu, and D. Gesbert, "Addressing the curse of mobility in massive MIMO with Prony-based angular-delay domain channel predictions," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, pp. 2903–2917, 2020.

[87] A. Papazafeiropoulos and T. Ratnarajah, "Linear precoding for downlink massive MIMO with delayed CSIT and channel prediction," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 809–914, IEEE, 2014.

[88] C. Kong, C. Zhong, A. K. Papazafeiropoulos, M. Matthaiou, and Z. Zhang, "Sum-rate and power scaling of massive MIMO systems with channel aging," *IEEE Transactions on communications*, vol. 63, no. 12, pp. 4879–4893, 2015.

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

[90] Y. Anzai, *Pattern recognition and machine learning*. Elsevier, 2012.

[91] M. Nerini, V. Rizzello, M. Joham, W. Utschick, and B. Clerckx, "Machine Learning-Based CSI Feedback With Variable Length in FDD Massive MIMO," *arXiv preprint arXiv:2204.04723*, 2022.

[92] N. Ma, K. Xu, X. Xia, C. Wei, Q. Su, M. Shen, and W. Xie, "Reinforcement learning-based dynamic anti-jamming power control in uav networks: An effective jamming signal strength based approach," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2355–2359, 2022.

[93] J. Kim, H. Lee, and S.-H. Park, "Learning Robust Beamforming for MISO Downlink Systems," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1916–1920, 2021.

[94] T. Lin and Y. Zhu, "Beamforming design for large-scale antenna arrays using deep learning," *IEEE Wireless Communications Letters*, vol. 9, no. 1, pp. 103–107, 2019.

[95] J. Yuan, H. Q. Ngo, and M. Matthaiou, "Machine learning-based channel prediction in massive MIMO with channel aging," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 2960–2973, 2020.

[96] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, and J. Choi, "Massive MIMO channel prediction: Kalman filtering vs. machine learning," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 518–528, 2020.

[97] C. Wu, X. Yi, Y. Zhu, W. Wang, L. You, and X. Gao, "Channel prediction in high-mobility massive MIMO: From spatio-temporal autoregression to deep learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 1915–1930, 2021.

[98] Z. Qin, H. Yin, Y. Cao, W. Li, and D. Gesbert, "A Partial Reciprocity-based Channel Prediction Framework for FDD massive MIMO with High Mobility," *arXiv preprint arXiv:2202.05564*, 2022.

[99] Y. Zhang, A. Alkhateeb, P. Madadi, J. Jeon, J. Cho, and C. Zhang, "Predicting Future CSI Feedback For Highly-Mobile Massive MIMO Systems," *arXiv preprint arXiv:2202.02492*, 2022.

[100] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[101] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[102] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[103] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE transactions on signal processing*, vol. 54, no. 6, pp. 2239–2251, 2006.

[104] O. Mehanna, K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos, "Feasible point pursuit and successive approximation of non-convex QCQPs," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 804–808, 2014.

[105] S. Hwang, H. Lee, J. Park, and I. Lee, "Decentralized computation offloading with cooperative uavs: Multi-agent deep reinforcement learning perspective," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 24–31, 2022.

[106] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[107] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[108] H. Li, W. Cai, Y. Liu, M. Li, Q. Liu, and Q. Wu, "Intelligent reflecting surface enhanced wideband MIMO-OFDM communications: From practical model to reflection optimization," *IEEE Transactions on Communications*, 2021.

[109] T. Adali and S. Haykin, *Adaptive signal processing: next generation solutions*, vol. 55. John Wiley & Sons, 2010.

[110] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.

[111] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[112] W.-K. Ma, P.-C. Ching, and Z. Ding, "Semidefinite relaxation based multiuser detection for M-ary PSK multiuser systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2862–2872, 2004.

[113] W.-K. Ma, T. N. Davidson, K. M. Wong, Z.-Q. Luo, and P.-C. Ching, "Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE transactions on signal processing*, vol. 50, no. 4, pp. 912–922, 2002.

[114] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1838–1851, 2019.

[115] V. Erceg, L. Schumacher, P. Kyritsi, A. Molisch, D. S. Baum, A. Y. Gorokhov, C. Oestges, Q. Li, K. Yu, K. N. Tal, *et al.*, "IEEE P802. 11," 2004.

[116] X. Qian, M. Di Renzo, J. Liu, A. Kammoun, and M.-S. Alouini, "Beamforming through reconfigurable intelligent surfaces in single-user mimo systems: Snr distribution and scaling laws in the presence of channel fading and phase noise," *IEEE wireless communications letters*, vol. 10, no. 1, pp. 77–81, 2020.

[117] J. An, C. Xu, L. Gan, and L. Hanzo, "Low-complexity channel estimation and passive beamforming for ris-assisted mimo systems relying on discrete phase shifts," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 1245–1260, 2021.

[118] M. Nerini, S. Shen, and B. Clerckx, "Discrete-value group and fully connected architectures for beyond diagonal reconfigurable intelligent surfaces," *IEEE Transactions on Vehicular Technology*, 2023.

[119] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6g be?," *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.

[120] Z. Xiao, Z. Han, A. Nallanathan, O. A. Dobre, B. Clerckx, J. Choi, C. He, and W. Tong, "Antenna array enabled space/air/ground communications and networking for 6G," *arXiv preprint arXiv:2110.12610*, 2021.

[121] V. Stankovic and M. Haardt, "Generalized design of multi-user MIMO precoding matrices," *IEEE Transactions on Wireless Communications*, vol. 7, no. 3, pp. 953–961, 2008.

[122] V. Raghavan, J. Cezanne, S. Subramanian, A. Sampath, and O. Koymen, "Beamforming tradeoffs for initial UE discovery in millimeter-wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 543–559, 2016.

[123] T. Kim, D. J. Love, and B. Clerckx, "MIMO systems with limited rate differential feedback in slowly varying channels," *IEEE Transactions on Communications*, vol. 59, no. 4, pp. 1175–1189, 2011.

[124] Y.-C. Liang and F. P. S. Chin, "Downlink channel covariance matrix (DCCM) estimation and its applications in wireless DS-CDMA systems," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 222–232, 2001.

[125] W. Zou, Z. Cui, B. Li, Z. Zhou, and Y. Hu, "Beamforming codebook design and performance evaluation for 60GHz wireless communication," in *2011 11th International Symposium on Communications & Information Technologies (ISCIT)*, pp. 30–35, IEEE, 2011.

[126] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," *Advances in neural information processing systems*, vol. 30, 2017.

[127] P. F. Smulders, "Statistical characterization of 60-GHz indoor radio channels," *IEEE Transactions on Antennas and Propagation*, vol. 57, no. 10, pp. 2820–2829, 2009.

[128] Y. Lee, J.-H. Lee, and Y.-C. Ko, "Beamforming optimization for irs-assisted mmwave v2i communication systems via reinforcement learning," *IEEE Access*, vol. 10, pp. 60521–60533, 2022.

[129] Q. Wang, K. Feng, X. Li, and S. Jin, "Precodernet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1677–1681, 2020.

[130] B. Saglam, D. Gurgunoglu, and S. S. Kozat, "Deep reinforcement learning based joint downlink beamforming and ris configuration in ris-aided mu-miso systems under hardware impairments and imperfect csi," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 66–72, IEEE, 2023.

[131] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, pmlr, 2015.

[132] S. Shen, B. Clerckx, and R. Murch, "Modeling and architecture design of reconfigurable intelligent surfaces using scattering parameter network analysis," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1229–1243, 2021.

[133] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive mimo: A survey," *IEEE Communications magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[134] Z. Liu, J. Zhang, Z. Liu, H. Du, Z. Wang, D. Niyato, M. Guizani, and B. Ai, "Cell-free xl-mimo meets multi-agent reinforcement learning: Architectures, challenges, and future directions," *arXiv preprint arXiv:2307.02827*, 2023.