Power Allocation in Uplink NOMA-Aided Massive MIMO Systems

A Thesis Submitted
to the College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in the Department of Electrical and Computer Engineering
University of Saskatchewan

by

The Khai Nguyen

Saskatoon, Saskatchewan, Canada

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, it is agreed that the Libraries of this University may make it freely available for inspection. Permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professors who supervised this thesis work or, in their absence, by the Head of the Department of Electrical and Computer Engineering or the Dean of the College of Graduate Studies and Research at the University of Saskatchewan. Any copying, publication, or use of this thesis, or parts thereof, for financial gain without the written permission of the author is strictly prohibited. Proper recognition shall be given to the author and to the University of Saskatchewan in any scholarly use which may be made of any material in this thesis.

Request for permission to copy or to make any other use of material in this thesis in whole or in part should be addressed to:

Head of the Department of Electrical and Computer Engineering
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
S7N 5A9

OR

Dean of College of Graduate and Postdoctoral Studies
116 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
S7N 5C9

Abstract

In the development of the fifth-generation (5G) as well as the vision for the future generations of wireless communications networks, massive multiple-input multiple-output (MIMO) technology has played an increasingly important role as a key enabler to meet the growing demand for very high data throughput. By equipping base stations (BSs) with hundreds to thousands antennas, the massive MIMO technology is capable of simultaneously serving multiple users in the same time-frequency resources with simple linear signal processing in both the downlink (DL) and uplink (UL) transmissions. Thanks to the asymptotically orthogonal property of users' wireless channels, the simple linear signal processing can effectively mitigate inter-user interference and noise while boosting the desired signal's gain, and hence achieves high data throughput. In order to realize this orthogonal property in a practical system, one critical requirement in the massive MIMO technology is to have the instantaneous channel state information (CSI), which is acquired via channel estimation with pilot signaling. Unfortunately, the connection capability of a conventional massive MIMO system is strictly limited by the time resource spent for channel estimation. Attempting to serve more users beyond the limit may result in a phenomenon known as *pilot contamination*, which causes correlated interference, lowers signal gain and hence, severely degrades the system's performance. A natural question is "Is it at all possible to serve more users beyond the limit of a conventional massive MIMO system?". The main contribution of this thesis is to provide a promising solution by integrating the concept of nonorthogonal multiple access (NOMA) into a massive MIMO system.

The key concept of NOMA is based on assigning each unit of orthogonal radio resources, such as frequency carriers, time slots or spreading codes, to more than one user and utilize a non-linear signal processing technique like successive interference cancellation (SIC) or dirty paper coding (DPC) to mitigate inter-user interference. In a massive MIMO system, pilot sequences are also orthogonal resources, which can be allocated with the NOMA approach. By sharing a pilot sequence to more than one user and utilizing the SIC technique, a massive MIMO system can serve more users with a fixed amount of time spent for channel estimation. However, as a consequence of pilot reuse, correlated interference becomes the main challenge that limits the spectral efficiency (SE) of a massive MIMO-NOMA system. To address this issue, this thesis focuses on how to mitigate correlated interference when combining NOMA into a massive MIMO system in order to accommodate a higher number of wireless users.

In the first part, we consider the problem of SIC in a single-cell massive MIMO system in order to serve twice the number of users with the aid of time-offset pilots. With the proposed time-offset pilots, users are divided into two groups and the uplink pilots from one group are transmitted *simultaneously* with the uplink data of the other group, which allows the system to accommodate more users for a given number of pilots. Successive interference cancellation is developed to ease the effect of pilot contamination and enhance data detection.

In the second part, the work is extended to a cell-free network, where there is no cell boundary and a user can be served by multiple base stations. The chapter focuses on the NOMA approach for sharing pilot sequences among users. Unlike the conventional cell-free massive MIMO-NOMA systems in which the UL signals from different access points are equally combined over the backhaul network, we first develop an optimal backhaul combining (OBC) method to maximize the UL signal-to-interference-plus-noise ratio (SINR). It is shown that, by using OBC, the correlated interference can be effectively mitigated if the number of users assigned to each pilot sequence is less than or equal to the number of base stations. As a result, the cell-free massive MIMO-NOMA system with OBC can enjoy unlimited performance when the number of antennas at each BS tends to infinity.

Finally, we investigate the impact of imperfect SIC to a NOMA cell-free massive MIMO system. Unlike the majority of existing research works on performance evaluation of NOMA, which assume perfect channel state information and perfect data detection for SIC, we take into account the effect of practical (hence imperfect) SIC. We show that the received signal at the backhaul network of a cell-free massive MIMO-NOMA system can be effectively treated as a signal received over an additive white Gaussian noised (AWGN) channel. As a result, a discrete joint distribution between the interfering signal and its detected version can be analytically found, from which an adaptive SIC scheme is proposed to improve performance of interference cancellation.

Acknowledgments

I would like to show my appreciation toward people who are always by my side and support me at any time in my life. This thesis is not only for me but also the accomplishment from all of you.

Foremost, I would like to show my greatest gratitude toward Professor Ha Nguyen for his guidance in my research since the very beginning of my graduate studies at the University of Saskatchewan.

I also would like to thank Professor Hoang Duong Tuan and Dr. Hoa Nguyen for their great technical supports. Your constructive cooperation plays a very important role in my research.

I want to extend my special thanks to Professors Eric Salt, Brian Daku, Brian Berscheid, Daniel Teng, and Mr. Rory Gowen for the knowledge that I learned from your teaching. It has made my experience with the University of Saskatchewan one of the best time in my life.

I also would like to show my gratitude towards Professors Chris Zhang, Ebrahim Bedeer Mohamed and Brian Berschied for serving as committee members in my thesis defence. Your constructive comments have significantly improved the quality of my thesis.

My special thanks go to my friends in our research lab: Nghia, Long, Peter, Ali, Shania, Botao, Dr. Gurjar and Dr. Shukla, and my landlord, Mr. Tan Nguyen's family. You have made my time in Saskatoon filled with enjoyable and unforgettable moments.

Finally, I would like to show my deepest love toward my family for loving me unconditionally, and my beloved Thao, for being with me through all the ups and downs.

Table of Contents

Permission to Use **i**

Abstract **ii**

Acknowledgments **iv**

Table of Contents **v**

List of Tables **viii**

List of Figures **ix**

List of Abbreviations **xi**

List of Symbols **xiii**

1 Introduction and Thesis Organization **1**

1.1 Introduction 1

1.2 Organization of the Thesis 4

2 Background **6**

2.1 Statistical Model of a Wireless Channel 6

2.2 Fundamentals of Massive MIMO Systems 9

2.2.1 Channel Estimation 11

2.2.2 UL Training Phase with Pilot Sequences 13

2.2.3 UL Achievable Rate with MRC Combining 15

2.2.4 Challenges with Massive MIMO System Design 17

2.3 Fundamentals of NOMA 18

2.3.1 Channel Capacity with Successive Interference Cancellation 19

2.3.2 Challenges with NOMA 20

2.3.3 Integration of NOMA into Massive MIMO Systems 20

3 Multiuser Massive MIMO Systems with Time-Offset Pilots and Successive Interference Cancellation 25

3.1 Introduction 26

3.1.1 Related Works 27

3.1.2 Contribution 29

3.2 Time-Offset Pilots and Channel Estimation 31

3.3 Uplink Data Transmission 34

3.3.1 Analysis in the training phase 34

3.3.2 Analysis in the Data Phase 37

3.4 Power Control 41

3.4.1 Max-Min QoS Optimization 41

3.4.2 Minimization of Total Power 45

3.5 Simulation Results 47

3.6 Conclusions 56

3.A Appendix I 58

4 Max-Min QoS Power Control in Generalized Cell-Free Massive MIMO-NOMA with Optimal Backhaul Combining 66

4.1 Introduction 68

4.2 System Model 73

4.2.1 Generalized Cell-Free Massive MIMO System 73

4.2.2 Channel Estimation 74

4.2.3 UL Data Transmission 75

4.3 Optimal Backhaul Combining 76

4.4 Asymptotic Analysis 80

4.4.1 Equal-Gain Backhaul Combining and Zero-Forcing Backhaul Combining 80

4.4.2 Asymptotic Analysis 81

4.5 User Grouping 82

4.6 Power Control Optimization 85

4.7 Simulation Results 89

4.8 Conclusions 96

4.A Examination of correlation property 97

4.B Proof of Theorem 1: Optimal weight combining vector 98

4.C Proof of Theorem 1: SINR with OBC 98

4.D Derivation for the SINR of EBC 99

4.E Derivation for the SINR of ZFBC 101

4.F Examination on how many BSs should serve a user 101

4.G Fundamental inequalities for convex approximation 103

5 Adaptive Successive Interference Cancellation in Cell-free Massive MIMO-NOMA 110

5.1 Introduction 111

5.2 System Model 113

5.3 Successive Interference Cancellation 116

5.4 Optimal Power Control with Adaptive SIC 119

5.5 Numerical Results and Discussion 120

5.6 Conclusion 122

6 Conclusion 126

6.1 Conclusion 126

6.2 Future Research Topics 127

List of Tables

3.1 Simulation parameters. 49

3.2 Rate contribution from the training phase. 52

4.1 Simulation parameters. 89

5.1 $P(s_j$ decided $|$ s_i transmitted) 119

5.2 Simulation parameters. 120

List of Figures

2.1 Communication over a wireless channel. 6

2.2 Illustration of a single-cell massive MIMO system. 10

2.3 FDD versus TDD time frame. 12

2.4 MMSE channel estimation with pilots in a massive MIMO system. 13

2.5 Linear combining at the BS over M antennas. 15

3.1 (a) Conventional pilot design, and (b) Time-offset pilot design. 31

3.2 Max-min QoS versus the number of antennas ($N = 30$ users). 50

3.3 Max-min QoS versus coherence interval ($N = 30$ users). 51

3.4 Max-min QoS versus the number of users ($M = 300$ antennas). 52

3.5 Cumulative distribution function of the max-min QoS ($N = 10$ users, $\tau_c = 120$). 53

3.6 Per-user average transmit power versus the number of antennas (QoS=1.5 bit/Hz/s, $N = 30$ users). 54

3.7 Per-user average transmit power versus coherence interval (QoS=1.5 bit/Hz/s, $N = 30$ users). 55

3.8 Per-user average transmit power versus required QoS level ($M = 300$ antennas, $N = 30$ users). 55

3.9 Comparison of the sum SE: $N = 15$ users with orthogonal pilots and $N = 30$ users with time-offset pilots. 56

4.1 System model. 73

4.2 Cumulative distribution function of UL SE with and without power control (3 BSs, each having 300 antennas). 90

4.3 Cumulative distribution functions of UL SE for three different backhaul combining methods (3 BSs, each having 300 antennas). 91

4.4 Cumulative distribution function of UL SE with and without grouping (3 BSs, each having 300 antennas). 92
4.5 Max-min QoS versus the number of antennas (30 users, 3 BSs). 93
4.6 Max-min QoS versus the number of users (3 BSs, each having 300 antennas). . . 94
4.7 Cumulative distribution function of UL SE versus the number of serving BSs (30 users, 300 antennas at each BS). 95
4.8 Cumulative distribution functions of UL SE for a cell-free massive MIMO system: $M \times L = 144$ and 50 users. 95
5.1 Histogram of a user's noise-plus-interference term in the I channel with 4-PAM constellation. 121
5.2 Correlation coefficients on the I and Q channels. 122
5.3 Cumulative distribution functions of the achievable QoS. 122

List of Abbreviations

ADC	Analog to Digital Converter
AP	Access Point
AWGN	Additive White Gaussian Noised
BC	Backhaul Combining
BS	Base Station
CDMA	Code Division Multiple Access
CIR	Channel Impulse Response
CPU	Central Processing Unit
CSI	Channel State Information
DAC	Digital to Analog Converter
dB	decibel
DPC	Dirty Paper Coding
DL	Downlink
EBC	Equal-gain Backhaul Combining
EE	Energy Efficiency
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
GP	Geometric Programming
IDMA	Interleaved Division Multiple Access
IoT	Internet of Things
KKT	Karush-Kuhn-Tucker
LTE	Long-Term Evolution
MIMO	Multiple Input Multiple Output
MMSE	Minimum Mean Square Error
MU-MIMO	Multiuser MIMO

MRC	Maximal-Ratio Combining
NOMA	Nonorthogonal Multiple Access
NP	Non-deterministic Polynomial-time
OBC	Optimal Backhaul Combining
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
PAM	Pulse Amplitude Modulation
QoS	Quality of Service
QAM	Quadrature Amplitude Modulation
SIC	Successive Interference Cancellation
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
UL	Uplink
ZFBC	Zero Forcing Backhaul Combining

List of Symbols

Symbols in Chapters 2

$a_i(t)$	attenuation of the ith path in passband
$\hat{a}_i(t)$	attenuation of the ith path in baseband
C_i	channel capacity of the ith user
$C_{\mathrm{sum}}^{(\mathrm{OMA})}$	sum channel capacity with OMA
$C_{\mathrm{sum}}^{(\mathrm{NOMA})}$	sum channel capacity with NOMA
CU_k	channel gain uncertainty term after MRC combining
DS_k	desired signal term after MRC
$\boldsymbol{e}_q$	channel estimation error vector of the kth user
f_c	carrier frequency
$\boldsymbol{h}_k$	channel vector of the kth to the BS
$\hat{\boldsymbol{h}}_k$	estimation of $\boldsymbol{h}_k$
$h(\tau, t)$	channel's impulse response
$h_{\mathsf{B}}(t, \tau)$	equivalent baseband channel impulse response
$H_{\mathsf{B}}(f, t)$	equivalent baseband channel frequency response
IoU_k	interference from other users after MRC combining
K	number of users
L	number of taps
M	number of antennas
$\boldsymbol{n}$	AWGN noise at BS
N_{p}	number of paths
$\boldsymbol{N}$	noise matrix during pilot training
N_k	noise term after MRC combining
p_k	users' UL data transmit power
$p_k^{(\mathrm{p})}$	users' pilot power

R_k	kth user's UL SE
SINR_k	kth user's SINR
$\boldsymbol{\mathcal{S}}_q$	Set of users sharing the same pilot with the qth user
T_d	coherence time
$\boldsymbol{v}_k$	combining vector of the kth user
W	signal's bandwidth
x_k	UL transmitted data
$x(t)$	transmitted signal in passband
$x_{\mathsf{B}}(t)$	transmitted data in baseband
$x_{\mathsf{B}}[m]$	sampled transmitted data in baseband
$\boldsymbol{y}$	received data signal at BS's antennas
$\boldsymbol{Y}$	received pilot signal at BS's antennas
$y(t)$	received signal in passband
$y_{\mathsf{B}}(t)$	received data in baseband
β_k	large-scale fading coefficient
γ_k	channel estimate's variance
$\delta(t)$	Delta Dirac function
σ^2	noise's variance
$\tau_i(t)$	Delay time of the ith path
τ_p	pilot's length
τ_c	coherence length
$\boldsymbol{\phi_k}$	kth user's pilot sequence

Symbols in Chapters 3

$\text{DS}_{g,k}^{(\text{dp})}$	desired signal after MRC combining in data phase
$\text{DS}_{g,k}^{(\text{tp})}$	desired signal after MRC combining in training phase
$\boldsymbol{e}_{g,k}$	channel estimation error of the gth user in kth group
$h(\cdot)$	differential entropy of a random variables
$\boldsymbol{h}_{g,k}$	channel vector of the kth user of the gth group to the BS
$\hat{\boldsymbol{h}}_{g,k}$	estimation of $\boldsymbol{h}_{g,k}$

$I(\cdot,\cdot)$	mutual information function between two random variables
$\mathrm{IP}_{g,q}^{(\mathrm{tp})}$	interference caused by pilots after MRC combining
$\mathrm{IoG}_{g,q}^{(\mathrm{dp})}$	interference from other group after MRC combining
$\mathrm{IwG}_{g,q}^{(\mathrm{tp})}$	interference from within group in training phase after MRC
$\mathrm{IwG}_{g,q}^{(\mathrm{dp})}$	interference from within group in data phase after MRC
K	number of users in each group
M	number of antennas
$\boldsymbol{n}$	AWGN noise at BS
N	number of users in the system
$\boldsymbol{N}$	noise matrix during pilot training
$\mathrm{N}_{g,k}^{(\mathrm{dp})}$	noise term after MRC combining in data phase
$\mathrm{N}_{g,k}^{(\mathrm{tp})}$	noise term after MRC combining in training phase
$p_{g,k}$	users' UL data transmit power in data phase
$p_{\max}$	maximum UL transmit power
$P_{g,k}^{(\mathrm{total})}$	average transmit power
$P_{\mathrm{new}}^{(\mathrm{total})}$	system's total transmit power from the current loop
$P_{\mathrm{prev}}^{(\mathrm{total})}$	system's total transmit power from the previous loop
r_k	data signal received at BS after MRC combining
$\boldsymbol{r}_k$	pilot signal received at BS after multiplied with kth user's pilot
$R_{g,q}^{(\mathrm{dp})}$	user's UL SE in data phase
$R_{g,q}^{(\mathrm{tp})}$	user's UL SE in training phase
$R_{\mathrm{ini}}^{(\mathrm{tp})}$	optimal max-min QoS in training phase
$R_{\mathrm{req}}^{(\mathrm{dp})}$	required SE in data phase
$R_{\mathrm{req}}^{(\mathrm{tp})}$	required SE in training phase
$R_{g,q}^{(\mathrm{total})}$	user's UL SE in total
$R_{\mathrm{new}}^{(\mathrm{total})}$	achievable total rate from the current loop
$R_{\mathrm{prev}}^{(\mathrm{total})}$	achievable total rate from the previous loop
$R_{\max}$	upper bound on R_{req}
$R_{\min}$	lower bound on R_{req}
$R_{g,q}^{(\mathrm{total})}$	user's total UL SE

$\mathrm{SINR}_{g,q}^{(\mathrm{dp})}$	user's SINR in data phase
$\mathrm{SINR}_{g,q}^{(\mathrm{tp})}$	user's SINR in training phase
$\boldsymbol{v}_{g,k}$	combining vector of the kth user of the gth group
$x_{1,k}^{(\mathrm{dp})}$	data signal in data phase
$x_{1,k}^{(\mathrm{tp})}$	data signal in training phase
$\boldsymbol{Y}$	received pilot signal matrix at BS
$\boldsymbol{y}^{(\mathrm{dp})}$	received data signal in data phase
$\boldsymbol{y}^{(\mathrm{tp})}$	received data signal in training phase
$\beta_{g,k}$	large-scale fading coefficient
$\gamma_{g,k}$	channel estimate's variance
$\rho_{g,k}^{(\mathrm{d})}$	users' data power in training phase
$\rho_{g,k}^{(\mathrm{p})}$	users' pilot power
σ^2	noise's variance
$\Upsilon_{1,q}^{(\mathrm{IP})}$	limit of $(\mathrm{IP})_{g,q}$ when $M \to \infty$
$\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}$	estimate of $\Upsilon_{1,q}^{(\mathrm{IP})}$
$\Upsilon_{1,q}^{(\mathrm{IoG})}$	limit of $(\mathrm{IoG})_{g,q}$ when $M \to \infty$
$\hat{\Upsilon}_{1,q}^{(\mathrm{IoG})}$	estimate of $\Upsilon_{1,q}^{(\mathrm{IoG})}$
τ_p	pilot's length
τ_c	coherence length
$\boldsymbol{\phi_k}$	the kth pilot sequence

Symbols in Chapters 4 and 5

$\boldsymbol{a}$	eigenvector of $\hat{\mathbf{R}}_{1,q}$
$\boldsymbol{c}_{1,q}$	correlated interference vector
$\hat{\boldsymbol{c}}_{1,q}$	normalized correlated interference vector
$\mathrm{CU}_{l,g,k}$	channel gain uncertainty after MRC at the lth BS
$\mathrm{DS}_{l,g,k}$	desired signal after MRC combining at the lth BS
$\boldsymbol{e}_{l,g,k}$	channel estimation error vector
$\boldsymbol{h}_{l,g,k}$	channel vector of the kth to the BS
$\hat{\boldsymbol{h}}_{l,g,k}$	estimation of $\boldsymbol{h}_{l,g,k}$

$\mathrm{IoG}_{l,g,k}$	interference from other group after MRC combining
$\mathrm{IwG}_{l,g,k}$	interference within group after MRC combining
K	number of groups
L	number of BSs
M	number of each BS's antennas
N	number of users in each group
$\boldsymbol{n}_l$	AWGN noise at the lth BS
$\mathbf{N}_l$	noise matrix at the lth BS
$p_{g,k}$	users' UL data transmit power
$p_{g,k}^{(\mathrm{p})}$	users' UL pilot transmit power
$p_{\max}$	maximum UL transmit power
$\mathbf{R}_{g,q}$	uncorrelated interference plus noise correlation matrix
$\hat{\mathbf{R}}_{g,q}$	normalized uncorrelated interference plus noise correlation matrix
$R_{g,q}^{(\mathrm{OBC})}$	UL SE with OBC before SIC
$R_{g,q}^{(\mathrm{SIC-OBC})}$	UL SE with OBC after SIC
$\mathcal{S}$	set of grouped users
$\mathcal{S}_c$	set of users that are not in any pairs
$\boldsymbol{s}_{1,q}$	desired signal vector
$\hat{\boldsymbol{s}}_{1,q}$	normalized desired signal vector
$\mathrm{SINR}_{1,q}^{(\mathrm{EBC})}$	SINR with EBC
$\mathrm{SINR}_{1,q}^{(\mathrm{OBC})}$	SINR with OBC
$\mathrm{SINR}_{1,q}^{(\mathrm{ZFBC})}$	SINR with ZFBC
$\mathrm{SNR}_{l,g,q}$	SNR of the gth user of the qth group at the lth BS.
$\boldsymbol{u}_{1,q}$	uncorrelated interference plus noise vector
$\hat{\boldsymbol{u}}_{1,q}$	normalized uncorrelated interference plus noise vector
$\boldsymbol{v}_{l,g,k}$	combining vector of the kth user of the gth group in the lth BS
$\boldsymbol{w}_{l,g,k}$	backhaul combining vector
$\boldsymbol{w}_{l,g,k}^{(\mathrm{EBC})}$	equal gain backhaul combining vector
$\boldsymbol{w}_{l,g,k}^{(\mathrm{OBC})}$	optimal backhaul combining vector
$\boldsymbol{w}_{l,g,k}^{(\mathrm{SIC-OBC})}$	optimal backhaul combining vector after SIC

$\boldsymbol{w}_{l,g,k}^{(\mathrm{ZFBC})}$	zero forcing backhaul combining vector
$x_{1,q}$	transmitted data signal
$\hat{x}_{1,q}$	detected data signal
$\boldsymbol{y}_l$	received data signal at the lth BS
$\mathbf{Y}_l$	received pilot signal matrix at the lth BS
$\beta_{l,g,k}$	large-scale fading coefficient
$\gamma_{l,g,k}$	channel estimate's variance
Δ	half distance between two signal points on a PAM constellation
$\boldsymbol{\kappa}_{1,q}$	signal vector before backhaul combining
$\hat{\boldsymbol{\kappa}}_{1,q}$	normalized signal vector before backhaul combining
$\hat{\boldsymbol{\kappa}}_{2,q}^{(\mathrm{aSIC})}$	normalized signal vector after SIC and before backhaul combining
$\lambda_{i,i}$	correlation coefficient between two users' large scale fading
$\xi_{l,1,q}$	correlated interference power without SIC
$\xi_{l,1,q}^{(\mathrm{aSIC})}$	residue power with adaptive SIC
$\xi_{l,1,q}^{(\mathrm{nSIC})}$	residue power with conventional SIC
$\rho_{1,q}^{(\mathrm{I})}$	correlation coefficient between $x_{1,q}$ and $\hat{x}_{1,q}$ in the I axis
$\rho_{1,q}^{(\mathrm{Q})}$	correlation coefficient between $x_{1,q}$ and $\hat{x}_{1,q}$ in the Q axis
σ_{UL}^2	noise's variance at BS
τ_p	pilot length
τ_c	coherence length
$\boldsymbol{\phi_k}$	the kth pilot sequence

1. Introduction and Thesis Organization

1.1 Introduction

Over the last decade, the world has witnessed a breakthrough development of wireless communication technology as well as the growing demand for high-throughput, low-latency and massive-connectivity wireless communication services. This is resulted from the introduction of smartphone, Internet of Things (IoT) applications, auto pilot vehicles, etc. According to Cisco's annual Internet report in 2019, it is predicted that in the next 5 years, the network throughput requirement for a single user may reach to 60 mega bits per second (Mbps) [1]. Moreover, it is expected that there will be billions more communication devices. This presents a huge problem on the connection capability of wireless networks in order to meet such a demand.

To deal with this problem, the author in [2] proposes the concept of massive multiple-input-multiple-output (MIMO), a wireless system in which the base stations (BSs) are equipped with hundreds antennas and can serve multiple users in the same time-frequency resources. Instead of scheduling users to operate on different orthogonal resource units like time slots (time-division multiple access, or TDMA) or frequency bands (frequency-division multiple access, or FDMA), a massive MIMO system enables all users to operate on the common resources simultaneously, which is promising to solve the limited connection problem. Far beyond just a scalable version of a conventional multiuser MIMO (MU-MIMO) system, where the number of antennas at each BS is not too large and inter-user interference is usually a major problem that degrades the system's performance, a massive MIMO system can asymptotically mitigate interference, thanks to the large antenna array effect. In a nutshell, a massive MIMO system creates *favorable propagation* without the need for a sophisticated non-linear interference management method such as successive interference cancellation (SIC) or dirty paper coding (DPC) [3, 4]. Using linear precoding in the downlink (DL), i.e., from the BS to users, and linear combining for the uplink (UL), i.e., from a user to a BS, with the aid of a massive antenna array, a massive MIMO system enjoys a low-complexity solution for the BSs, while only a single antenna is required at an user's equipment [2–4]. As a result, the massive MIMO technology is not only promising in terms of connection capability, but also in terms of providing high spectral efficiency communication without the requirement of extra bandwidth or increasing cell density. The key advantages of massive MIMO systems can be summarized as follows [2–5]:

- **Massive connection capability:** By serving all users in the same time-frequency resources, massive MIMO systems can serve tens to hundreds times more users as compared to existing communication networks with the same radio resources.

- **High spectral efficiency (SE) and energy efficiency (EE):** Thanks to the large antenna arrays, massive MIMO systems can achieve extremely strong, deterministic array gain while mitigating the effect of small-scale fading. Furthermore, inter-user interference can be effectively reduced as a result of asymptotically orthogonal property of users' channels [2–5]. Thus, massive MIMO systems can provide very high SE, high EE and high reliability communication.

- **Low complexity signal processing:** Unlike conventional multiuser MIMO systems, where the number of antennas is not large enough to asymptotically mitigate interference and non-linear interference management methods like SIC or DPC must be employed, a massive MIMO system can have simple linear signal processing at the BS side and does not require any extra signal processing at the users' equipments. Hence, the complexity of the system on both transmission ends can be low [2–5].

With all these advantages, the massive MIMO technology has drawn great attention in both academia and industry, and it is expected to play an important role in the design of next-generation wireless networks.

However, a massive MIMO system also faces several problems. Although time and frequency become common resources for all users, another important resource in massive MIMO systems that needs to be wisely allocated is the pilots, which are known signals used for channel estimation. Unfortunately, the number of pilots is strictly limited by the wireless channel's coherence time [4, 7]. Ideally, users should be allocated with mutually orthogonal pilots [6], which also means that the number of users should not exceed the number of pilots. However, in practice, pilot reuse is inevitable. The problem is that reusing the same pilot sequence for more than one user results in the so-called *pilot contamination* effect, which degrades the quality of channel estimation and causes correlated interference, which cannot be eliminated by the large antenna effect [2, 4, 8]. To deal with this problem, some research works have been done to reduce the effect of pilot contamination, such as strategically reusing pilots with certain reuse factor and pattern [8], creating multiple pilot sets from a basic set [9], or using time-offset pilots [10]. However, all these methods can only reduce the impact of pilot contamination, but cannot completely eliminate it.

Recently, a new approach to deal with the problem of having limited number of pilots in massive MIMO systems that has gained a lot of attention is nonorthogonal multiple access (NOMA) [11–14]. Generally, NOMA is based on the idea of sharing each orthogonal radio resource such as time slot, frequency subcarrier or spreading code to more than just one user [15]. It is shown in many research works that NOMA can outperform orthogonal multiple access (OMA) in terms of both the sum data rate and fairness by smartly allocating more powers to users which have worse channel conditions and performing SIC to mitigate interference [15–17].

In a massive MIMO system, because users operate in the same time-frequency, the concept of NOMA can be employed by means of sharing pilots [11–13, 18]. Unlike sharing other types of orthogonal resources, where the performance gain comes from optimally allocating different levels of power to users to achieve the maximum sum rate (which is shown to be always equal or greater than the sum rate achieved by OMA [15]), sharing pilots in a massive MIMO system can be either advantageous or disadvantageous. On one hand, by sharing pilots, less time is required for channel estimation, which can enhance SE since more time resources can be spent for data transmission. On the other hand, sharing pilots to multiple users results in severe pilot contamination, which reduces the signal-to-interference-plus-noise ratio (SINR) of the system. The performance of this approach has been analyzed in [11], which shows that the number of connections of a NOMA-aided massive MIMO system can be significantly enhanced, with the trade-off being reduced per-user data rate.

The pilot contamination problem has motivated us to carry out research on power control in massive MIMO systems with the aid of NOMA in the form of non-orthogonal pilots and SIC. The objective is to mitigate the effect of pilot contamination resulted from the nonorthogonality among different users' pilot sequences. This shall be achieved by fulfilling two main tasks. First, we exploit the structure of a massive MIMO system to eliminate correlated interference, which is caused by reusing pilots and could severely degrade the system's performance. Second, we formulate and solve power control problems to ensure that all users are equally served with the best quality of service (QoS).

1.2 Organization of the Thesis

This thesis is presented in a manuscript-based style. In Chapter 2, the main concepts of massive MIMO systems and NOMA are introduced. The benefits as well as existing challenges of massive MIMO systems and NOMA are discussed and linked to the main contributions of the thesis. The remaining body of the thesis contains contributions that have been published or accepted for publication.

Chapter 3 includes a manuscript that proposes the use of the SIC technique in a single-cell massive MIMO system with time-offset pilots, in which users are divided into two groups. Every coherence interval is scheduled so that the pilots of one group are transmitted simultaneously with UL data of the other group. In this way, with a fixed number of pilots, the number of users can be served in the system is doubled. The SIC technique is utilized to remove correlated interference caused by the pilots of each group to the other. Furthermore, a power control algorithm which is based on the bisection method is proposed to optimally balance the rate contribution between the training phase and data phase.

In the next two chapters, the work is extended to a NOMA cell-free massive MIMO system, a wireless system with no cell boundary and having multiple BSs serving all users at the same time. In the manuscript included in Chapter 4, a NOMA approach is proposed to share each pilot sequence to more than one user. Exploiting the co-operation of BSs, which are connected via a backhaul network, we show that the correlated interference caused by reusing pilots can be effectively mitigated by optimally combining the signals from all BSs. The max-min QoS power control problem is also formulated and solved to achieve the best QoS value that can be equally served to all users in the network. In addition, to analyze the effect of imperfect SIC to a NOMA cell-free massive MIMO system, we derive a discrete joint distribution model between the interfering signal and its detected version before performing SIC. Based on this statistical model, an adaptive SIC algorithm is proposed to improve performance of interference cancellation. This contribution is presented in the third manuscript included in Chapter 5.

Finally, Chapter 6 summarizes the contributions of the thesis.

2. Background

2.1 Statistical Model of a Wireless Channel

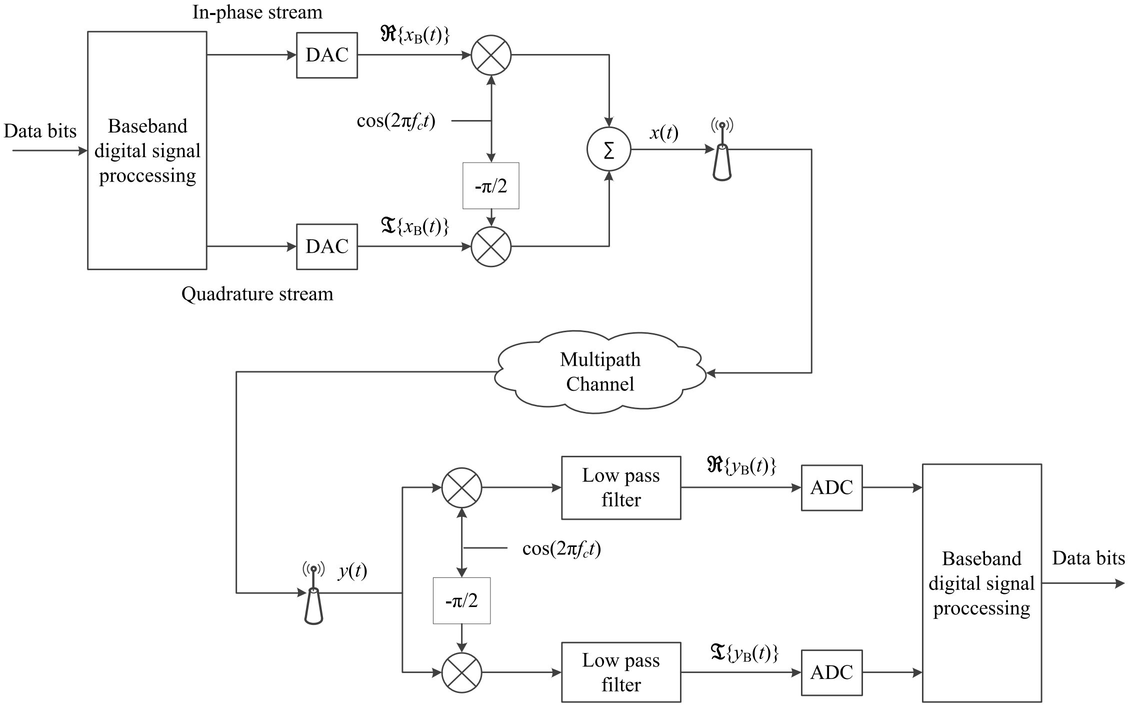


Figure 2.1 Communication over a wireless channel.

In a wireless communication system, in order to transmit a baseband data signal $x_{\mathsf{B}}(t)$, which occupies the frequency band limited to $W/2$ Hz, over a wireless channel, it must be modulated with a sinusoidal carrier at a higher radio frequency (RF) before being transmitted using a transmit antenna. This is illustrated in Fig. 2.1 for the simplest case of having one transmit antenna and one receive antenna.

In the transmitter, the information bits enter a baseband digital signal processing (DSP), whose outputs are the in-phase and quadrature signal samples. These signal samples are converted into the continuous-time baseband signals by two digital-to-analog (D/A) converters, one for the in-phase samples and one for the quadrature samples, respectively. The complex baseband signal $x_{\mathsf{B}}(t) = \Re\{x_{\mathsf{B}}(t)\} + j\Im\{x_{\mathsf{B}}(t)\}$ is up-converted to a passband frequency by multiplying with a sinusoidal carrier of frequency f_c. The resulting passband signal x_t is mathematically expressed by:

$$x(t) = \Re\left\{x_{\mathsf{B}}(t)\exp\left\{j2\pi f_c t\right\}\right\} \tag{2.1}$$

where $\Re\{\cdot\}$ denotes the real part of the enclosed quantity. This passband signal is then transmitted by an antenna to the receiver. At the receiver side, an antenna acquires the signal $y(t)$, which is then down-converted to baseband to obtain the baseband signal $y_{\mathsf{B}}(t)$. The baseband signal $y_{\mathsf{B}}(t)$ is then converted to digital samples (in-phase and quadrature samples) by using a pair of analog-to-digital (A/D) converters. Finally, the digital samples are processed by the DSP block to recover the information bits.

To examine the effect of the wireless channel to the transmitted signal, it is necessary to establish a mathematic relationship between the baseband transmitted and received signals, namely $x_{\mathsf{B}}(t)$ and $y_{\mathsf{B}}(t)$. In practice, when $x(t)$ is transmitted over a wireless channel, at destination, the antenna usually receives multiple replicas of the original signal, which are propagated over different paths. This is because the signal transmitted in different directions can get reflected or diffracted when hitting obstacles, or scattered when traveling over a large number of small objects and reflected in different directions.

This phenomenon results in the so-called *fading* effect, which is further classified into *large-scale fading* and *small-scale fading.* Large-scale fading accounts for the attenuation of the received signal strength due to path loss during propagation and shadowing, which is affected by propagation environment and terrains between the transmitter and receiver. Large-scale fading changes very slowly with respect to the change of distance (over hundreds or thousands of wavelengths). On the other hand, small-scale fading refers to the rapid fluctuation in the signal's strength due to the constructive or destructive effect when different signal copies arrive at the receiver after traversing multiple paths that have different path losses, time delays, and frequency offsets caused by the Doppler effect. Small-scale fading changes rapidly over time and distance, which causes the signal strength to vary significantly.

To represent the characteristics of a multipath wireless channel, a common model is to describe the multipath phenomenon in the form of a filter with time-varying channel impulse response (CIR):

$$h(\tau, t) = \sum_{i=1}^{N_\mathrm{P}} a_i(t)\delta(\tau - \tau_i(t)) \tag{2.2}$$

where $a_i(t)$ and $\tau_i(t)$ denote the attenuation and time delay of the ith path as functions of time and N_P is the number of paths. With this CIR, the passband signal obtained at the receiver's antenna can be expressed as:

$$y(t) = x(t) \otimes h(\tau, t) = \sum_{i=1}^{N_\mathrm{P}} a_i(t)x(t - \tau_i(t)) \tag{2.3}$$

By substituting $x(t) = \Re\left\{x_\mathsf{B}(t)\exp\left\{j2\pi f_c t\right\}\right\}$ and $y(t) = \Re\left\{y_\mathsf{B}(t)\exp\left\{j2\pi f_c t\right\}\right\}$ into the above equation, the following relationship can be achieved between $x_\mathsf{B}(t)$ and $y_\mathsf{B}(t)$:

$$y_\mathsf{B}(t) = \sum_{i=1}^{N_\mathrm{P}} a_i(t)\exp\left\{-j2\pi f_c\tau_i(t)\right\} x_\mathsf{B}(t - \tau_i(t)) \tag{2.4}$$

This also represents a linear time-varying system whose impulse response is

$$h_\mathsf{B}(t, \tau) = \sum_{i=1}^{N_\mathrm{P}} \hat{a}_i(t)\delta(\tau - \tau_i(t)) \tag{2.5}$$

where $\hat{a}_i(t) = a_i(t)\exp\left\{-j2\pi f_c\tau_i(t)\right\}$. The channel frequency response can be therefore calculated by:

$$H_\mathsf{B}(f, t) = \sum_{i=1}^{N_\mathrm{P}} a_i(t)\exp\left\{-j2\pi f_c\tau_i(t)\right\} \tag{2.6}$$

By sampling the received signal $y_\mathsf{B}(t)$ in (2.4) with the sampling rate W, the following discrete-time model can be obtained:

$$y_\mathsf{B}[m] = \sum_{l=1}^{L} h_l[m]x_\mathsf{B}[m - l], \tag{2.7}$$

where $h_l[m]$ is the lth channel filter tap at time m. The value of $h_l[m]$ depends on the strength of signal $\hat{a}_i(t)$ from the ith paths whose time delay $\tau_i(t)$ is close to l/W. Hence, the number of taps to represent a wireless channel depends on the channel bandwidth W and the maximum delay spread $T_d \triangleq \underset{\max}{i, j}\, |\tau_i(t) - \tau_j(t)|$. This is summarized as follows:

- $1/W \gg T_d$: This means that signals from all paths arrive within a symbol period. Hence, only one tap is needed to represent the channel. This is called a one-tap channel model.

  Moreover, when $W \ll \frac{1}{T_d}$, the change in channel's frequency response in (2.5) over the bandwidth of W can be considered negligible, and the channel is typically called *frequency flat*. Flat fading is desirable in communications since it offers relatively equal gain for a signal at all frequency, which avoids non-linear distortion.

- $1/W \ll T_d$: This means that the signals from different paths arrive at the receiver over different symbol periods. As a result, multiple taps are required to represent the channel.

  In this case, when moving within the bandwidth of W, the change in channel's frequency response is significant, which may cause non-linear distortions to the signal. The channel in this case is called *frequency selective*.

In this thesis, we focus on the case of a flat fading channel. With flat fading, the channel can be represented by one channel tap. As a result, the tap's gain $h_l[m]$ is the sum of all path gains $\hat{a}_i(t)$ evaluated as the corresponding sampling time. Assume that the signals from all paths are mutually independent, from the central limit theorem, the tap's gain can be effectively modeled as a complex Gaussian random variable. In such a case, the amplitude of the tap's gain follows a Rayleigh distribution. This fading model is widely known as a Rayleigh fading channel, which shall also be used throughout this thesis.

2.2 Fundamentals of Massive MIMO Systems

A massive MIMO system is illustrated in Fig. 2.2 in which multiple users are served by a BS equipped with a very large number of antennas (could be hundreds or thousands). All users in the system operate in the same time-frequency resources [2–5].

To see how a massive MIMO system works, consider a simple single-cell massive MIMO system with K users transmitting their uplink (UL) signals to a BS which is equipped with a massive antenna array with M antennas. Because all users operate in the same radio

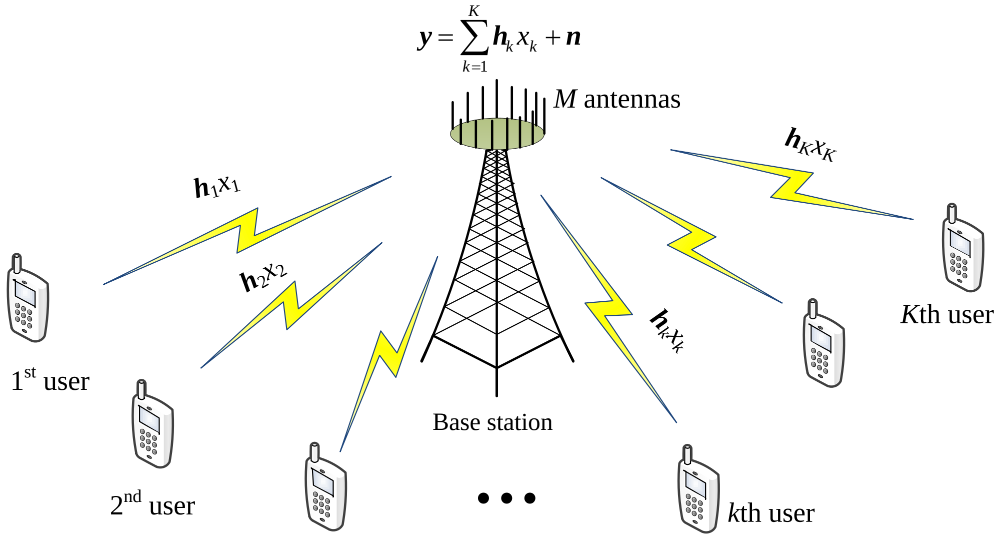


Figure 2.2 Illustration of a single-cell massive MIMO system.

resources, the UL signal received at the BS is a superposition of signals from all users, which can be expressed as:

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

where $\boldsymbol{h}_k=[h_{k,1},\ldots,h_{k,M}]^T\sim\mathcal{CN}(0,\mathbf{I}_M)$ represents the UL Rayleigh fading channel from the kth user to the BS, x_k denotes the UL data symbol which belongs to a unit-power quadrature-amplitude modulation (QAM) constellation, and $\boldsymbol{n}\sim\mathcal{CN}(0,\mathbf{I}_M)$ is AWGN noise at the BS, which is mostly due to thermal vibrations of atoms in conductors.

Assume that the BS has the perfect instantaneous channel state information (CSI) from all the users, one can apply the maximum ratio combining (MRC) to process $\boldsymbol{y}$. For example, in order to detect the data symbol x_1 for the first user, the MRC combining vector is $\boldsymbol{v}_1=\frac{1}{M}\boldsymbol{h}_1$ [7] and:

$$\boldsymbol{s}_1=\boldsymbol{v}_1^H\boldsymbol{y}=\boldsymbol{v}_1^H\boldsymbol{h}_1 x_1+\sum_{k=2}^{K}\boldsymbol{v}_1^H\boldsymbol{h}_k x_k+\boldsymbol{v}_1^H\boldsymbol{n} \tag{2.9}$$

Assume that the channel vectors from different users to the BS are mutually independent, the following properties can be established by the law of large numbers [7]:

$$\boldsymbol{v}_1^H \boldsymbol{h}_1 = \frac{1}{M}\|\boldsymbol{h}_1\|^2 \xrightarrow[M\to\infty]{\text{a.s}} 1 \tag{2.10}$$

$$\boldsymbol{v}_1^H \boldsymbol{h}_k = \frac{1}{M}\boldsymbol{h}_1^H \boldsymbol{h}_k \xrightarrow[M\to\infty]{\text{a.s}} 0, \quad \forall k \neq 1 \tag{2.11}$$

$$\boldsymbol{v}_1^H \boldsymbol{n} = \frac{1}{M}\boldsymbol{h}_1^H \boldsymbol{n} \xrightarrow[M\to\infty]{\text{a.s}} 0. \tag{2.12}$$

where $\xrightarrow[M\to\infty]{\text{a.s}}$ denotes almost sure convergence. This implies that after combining, only the signal from the first user remains, while the interference and noise terms are completely removed. The property in (2.10) is called *channel hardening* since when $M \to \infty$, the desired signal gain gets close to $\mathbb{E}[|h_{1,1}|^2]$, which means the effect of small-scale fading can almost be eliminated and the gain converges to a determined value [4, 7, 8]. This property is very useful in terms of demodulation and power control at the receiver side [4, 7, 8]. On the other hand, the effect in (2.11) is known as *favorable propagation*, which is resulted from the asymptotic orthogonality between different users' channels. This allows multiple users to operate on the same time-frequency resources [4, 7, 8].

2.2.1 Channel Estimation

In the above discussion, we assume that the instantaneous CSI is perfectly known at the BS side. However, in practice, in order to acquire CSI, channel estimation is required. The radio resources are divided into time-frequency resource blocks in which the channels can be considered frequency-flat and time-invariant. The time interval that the channels stay static is called coherence length and assumed to span τ_c symbols. This means that the channels have to be re-estimated after every τ_c symbols and how this can be done depends on the duplexing mode, namely time division duplex (TDD) or frequency division duplex (FDD). This is depicted in Fig.2.3.

FDD mode:

In the FDD mode, UL and DL transmissions occupy separated frequency bands, and the CSI of both channels is required at BS side. To obtain the DL CSI from M antenna of the BS to the users in the system, the BS has to spend at less M symbols in the DL channel to

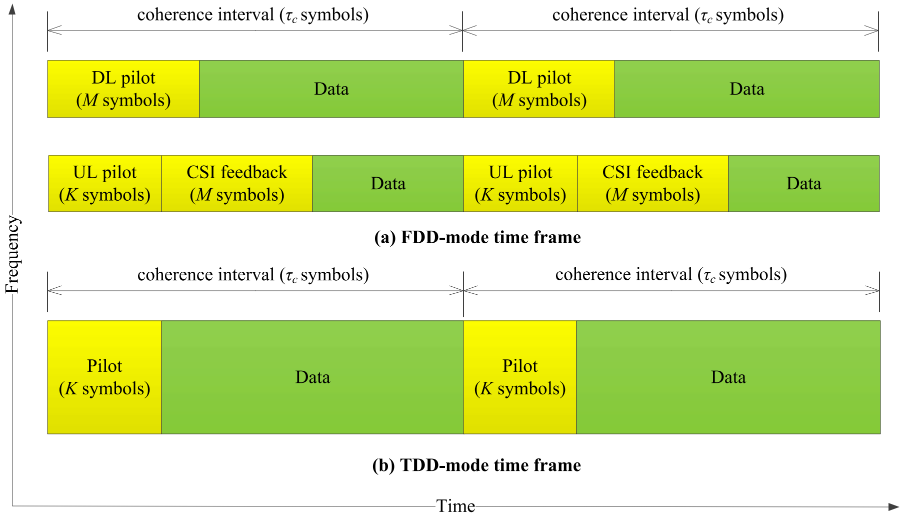


Figure 2.3 FDD versus TDD time frame.

transmit pilots. The users, after receiving pilots and estimating the channels, have to send the estimated CSI back to the BS (called CSI feedback), which requires at least another M symbols in the UL channel. Meanwhile, in order for the BS to estimate UL channels, K users also need to send pilots, which requires K symbols in the UL channel. As a result, in total, channel estimation with the FDD mode requires at least $M + K$ symbols in the UL channel and M symbols in the DL channel. In a situation that the BS is equipped with hundreds to thousands antennas, this leads to a huge overhead for the system.

TDD mode:

In a massive MIMO system operating in the TDD mode, channel reciprocity is exploited in channel estimation. Due to the fact that both the UL and DL transmissions occupy the same frequency band, the system is designed in such a way that the total time resources spent for UL and DL data transmissions fit in one coherence interval. With this design, in every coherence interval, K users only need to send K UL pilots to the BS for the channel estimation purpose, which only takes K symbols. It can be seen that with the TDD mode, the time required for channel estimation is independent of the number of antennas M, which makes the TDD mode more adaptive with the scaling of the antenna array.

2.2.2 UL Training Phase with Pilot Sequences

Given the advantages of the TDD mode over the FDD mode, in this thesis, we choose to investigate massive MIMO systems with the TDD mode. This section shows how channel estimation with pilots in a TDD massive MIMO system can be carried out. Consider a single-cell multi-user massive MIMO system in which one M-antenna BS serves K users, who are randomly distributed over the cell. The channels between the users and the BS are assumed to be Rayleigh fading, frequency flat and approximately constant within a coherence interval of length τ_c symbols. This means that the channel vector of user k can be modeled as $\boldsymbol{h}_k \sim \mathcal{CN}\left(0, \beta_k \boldsymbol{I}_M\right)$, where β_k represents large-scale fading.

It is assumed that the BS does not know the exact channel coefficients but the channel statistics. For the channel estimation purpose, a set of K length-τ_p pilot sequences is used. These pilots are collectively represented by a $\tau_p \times K$ pilot matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K]$ where $\|\boldsymbol{\phi}_k\|^2 = \tau_p$. With the sequence length of τ_p symbols, there are at most τ_p sequences which are mutually orthogonal. Hence, usually, in order to achieve orthogonality among pilot sequences of all users, the pilot length is set at the minimum value $\tau_p = K$. Otherwise, some users have to use pilot sequences which are not orthogonal to other users' sequences.

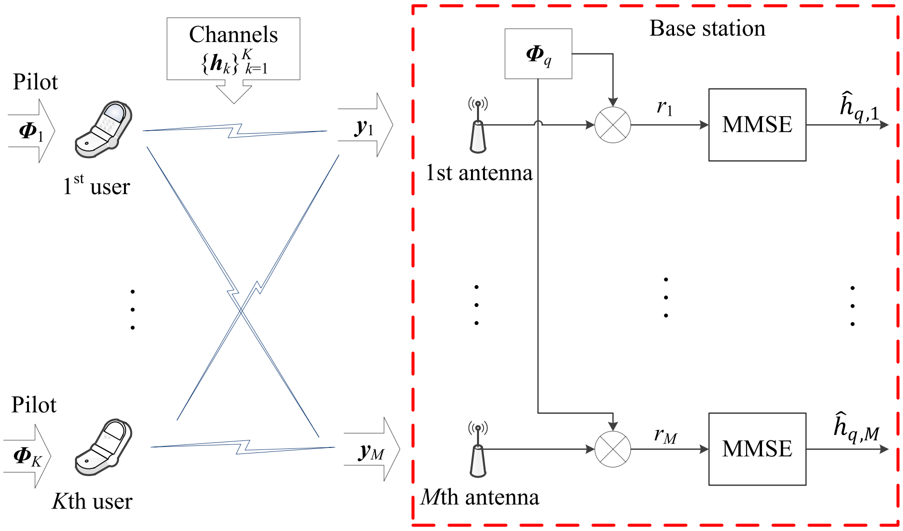


Figure 2.4 MMSE channel estimation with pilots in a massive MIMO system.

In the following, we consider a general case of τ_p to see how the relationship between τ_p and K can affect the performance of a massive MIMO system in the training phase. With all users sending UL pilots as in Fig. 2.4, the signal matrix $\boldsymbol{Y} \in \mathbb{C}^{M\times\tau_p}$ received at the BS over τ_p time slots (symbols) is given as:

$$\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{h}_k \sqrt{\rho_k^{(\mathrm{p})}} \boldsymbol{\phi}_k^H + \boldsymbol{N}, \tag{2.13}$$

where $\rho_k^{(\mathrm{p})}$ is the pilot power, and $\boldsymbol{N} \in \mathbb{C}^{M\times\tau_p}$ denotes AWGN noise matrix whose entries are complex Gaussian random variables with zero mean and variance of σ^2. To estimate the channel from the qth user, the BS multiplies the received signal with the corresponding pilot, which results in:

$$\boldsymbol{r}_q = \boldsymbol{Y}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \boldsymbol{h}_q\sqrt{\rho_q^{(\mathrm{p})}}\tau_p + \sum_{k\in\mathcal{S}_q, k\neq q}^{K} \boldsymbol{h}_q\sqrt{\rho_q^{(\mathrm{p})}}\tau_p\frac{\boldsymbol{\phi}_k^H\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} + \boldsymbol{N}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}. \tag{2.14}$$

where $\mathcal{S}_q$ is the set containing all pilot sequences $\boldsymbol{\phi}_k$ which are not orthogonal to $\boldsymbol{\phi}_q$. There are two different situations:

- $\tau_p \geq K$: There are enough mutually orthogonal pilot sequences for all K users. As a result, the second term in (2.14) disappears and the observation for channel estimation is corrupted by AWGN noise only.
- $\tau_p < K$: Due to the fact that there are not enough mutually orthogonal sequences for all users, some users (say the qth user) will have to use a pilot sequence which is not orthogonal to at least one of the other users in the system. Consequently, the observation used for estimating the channel of the qth user contains not only AWGN noise, but also the channel information of other users, which degrades the quality of channel estimation. This phenomenon is known as *pilot contamination.*

From the observation in (2.14), the estimate of $\boldsymbol{h}_q$ can be obtained by using the minimum mean-squared error (MMSE) estimator [20] as:

$$\hat{\boldsymbol{h}}_q = \frac{\mathrm{cov}\left\{\boldsymbol{h}_q, \boldsymbol{r}_q\right\}}{\mathrm{var}\left\{\boldsymbol{r}_q\right\}}\boldsymbol{r}_q = \mu_q\boldsymbol{r}_q, \tag{2.15}$$

where

$$\mu_q = \frac{\sqrt{\rho_q^{(\mathrm{p})}\tau_p\beta_q}}{\rho_q^{(\mathrm{p})}\tau_p\beta_q + \rho_k^{(\mathrm{p})}\beta_k\frac{\boldsymbol{\phi}_k^H\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} + \sigma^2}.$$

As a result, the estimated channel is a random vector, which follows the distribution $\hat{\boldsymbol{h}}_q \sim \mathcal{CN}(0, \gamma_q \boldsymbol{I}_M)$, where

$$\gamma_q = \frac{\rho_q^{(\mathrm{p})} \tau_p \beta_q^2}{\rho_q^{(\mathrm{p})} \tau_p \beta_q + \rho_k^{(\mathrm{p})} \beta_k \frac{\boldsymbol{\phi}_k^H \boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} + \sigma^2}. \tag{2.16}$$

Furthermore, the estimation error $\boldsymbol{e}_q = \boldsymbol{h}_q - \hat{\boldsymbol{h}}_q$ is independent of the estimated channel and distributed as $\boldsymbol{e}_q \sim \mathcal{CN}(0, (\beta_q - \gamma_q)\boldsymbol{I}_M)$.

After obtaining the channel estimation, the BS applies a linear processing vector for the detection of the UL data to each user.

2.2.3 UL Achievable Rate with MRC Combining

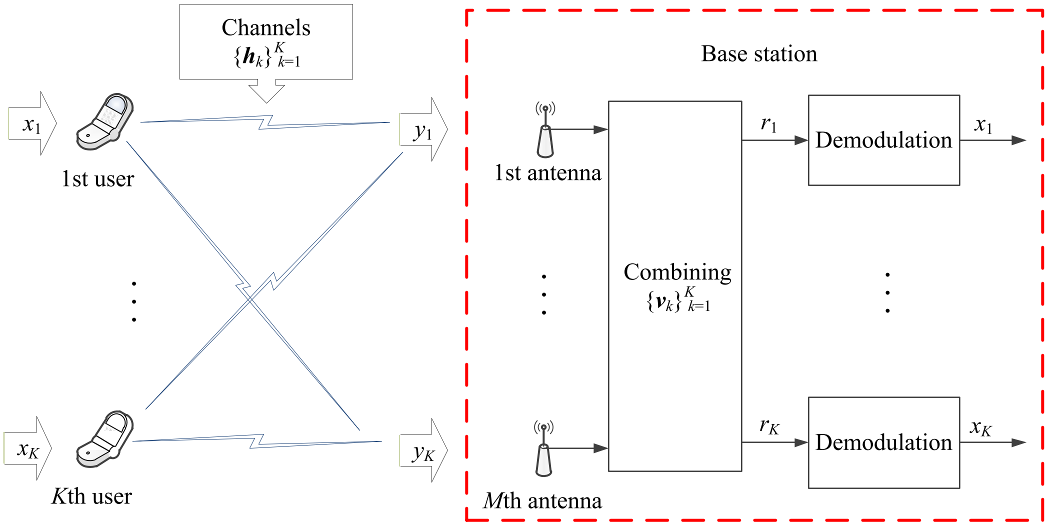


Figure 2.5 Linear combining at the BS over M antennas.

After the CSI has been obtained, the UL data transmission is carried out. With all users in the system simultaneously sending their UL data in the same frequency, the signal received at an arbitrary antenna of the BS is the sum of signals from all K users after propagation through wireless channels. As a result, the signal $\boldsymbol{y} = [y_1, y_2 \dots y_M] \in \mathbb{C}^{M\times 1}$ received at all M antennas of the BS over one symbol period can be written in vector form as:

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{h}_k \sqrt{p_k} x_k + \boldsymbol{n}, \tag{2.17}$$

where, as before, x_k represents the respective data symbol from the kth user. To detect data of an arbitrary user, say the qth user, the BS multiplies the above received signal with the corresponding MRC combining vector $\boldsymbol{v}_q = \hat{\boldsymbol{h}}_q$ before demodulation, as illustrated in Fig. 2.5. This yields

$$r_q = \boldsymbol{v}_q^H \boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{v}_q^H \boldsymbol{h}_k \sqrt{p_k} x_k + \boldsymbol{v}_q^H \boldsymbol{n}. \tag{2.18}$$

To see how the desired data is affected by different components, decompose the signal as:

$$r_q = \underbrace{\mathbb{E}\left\{\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_q\right\} \sqrt{p_q} x_q}_{\text{DS}_q\text{- Desired signal}} + \underbrace{\left(\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_q - \mathbb{E}\left\{\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_q\right\}\right) \sqrt{p_q} x_q}_{\text{CU}_q\text{- Channel gain uncertainty}} + \underbrace{\sum_{k=1, k\neq q}^{K} \boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_k \sqrt{p_k} x_k}_{\text{IoU}_q\text{- Interference}} + \underbrace{\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{n}}_{\text{N}_q\text{- Noise}}, \tag{2.19}$$

The decomposition of the received signal in Eq. (2.19) has an intuitive structure. The first component, DS_q, is the desired signal, which experiences a constant gain. Due to imperfect CSI at the BS, the second term CU_q is the interference originating from the desired signal itself, which is independent from the first term. The last two terms represent interference from other users and thermal noise.

With this signal decomposition, a lower bound on the UL achievable SE can be obtained by the definition of mutual information [21] as:

$$R_q = \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2\left(1 + \text{SINR}_q\right) \quad \text{bits/Hz/s} \tag{2.20}$$

where the effective SINR is defined as:

$$\begin{aligned} \text{SINR}_q &= \frac{\mathbb{E}\left\{\left|\text{DS}_q\right|^2\right\}}{\mathbb{E}\left\{\left|\text{CU}_q\right|^2\right\} + \mathbb{E}\left\{\left|\text{IoU}_q\right|^2\right\} + \mathbb{E}\left\{\left|\text{N}_q\right|^2\right\}} \\ &= \frac{p_q \left|\mathbb{E}\left\{\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_q\right\}\right|^2}{\sum_{k=1}^{K} p_k \mathbb{E}\left\{\left|\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_k\right|^2\right\} - p_q \left|\mathbb{E}\left\{\boldsymbol{v}_q^{\mathsf{H}} \boldsymbol{h}_q\right\}\right|^2 + \sigma_{\text{UL}}^2 \mathbb{E}\left\{\left\|\boldsymbol{v}_q\right\|^2\right\}} \end{aligned} \tag{2.21}$$

By calculating the first and second moments of different terms in (2.19), we can arrive at the following closed-form expression of the effective SINR [2, 21]:

$$\text{SINR}_q^{(\text{MRC})} = \frac{p_q \gamma_q M}{\sum_{k=1}^{K} p_k \beta_k + \sum_{k\in\mathcal{S}_q, k\neq q} p_k \gamma_k \frac{\boldsymbol{\phi}_k^H \boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} M + \sigma_{\text{UL}}^2} \tag{2.22}$$

From (2.22), it can be seen that when all users are assigned with mutually orthogonal pilots (i.e., when $K \leq \tau_p$), the second term of the denominator of (2.22) disappears and the SINR grows proportionally with the number of antennas. However, when $K > \tau_p$, nonorthogonal pilots must be used and pilot contamination exists. As a consequence, the second term of the denominator causes the SINR saturated at a finite value even when the number of antennas tends to infinity. Hence, with nonorthogonal pilots, the system cannot enjoy the array gain from the antennas and the system's performance is saturated as a result of pilot contamination. The second term of the denominator in (2.22) is originated from the so-called *correlated interference*, which is, similar to desired signal, amplified by antenna's array gain and causes the SINR to saturate.

2.2.4 Challenges with Massive MIMO System Design

The effectiveness of a massive MIMO system is based on the key concept of asymptotic orthogonality among users' channels. However, practical realization of this technology is challenged by the following problems, mostly related to the orthogonality property.

Pilot contamination:

From the previous discussion, if all users are assigned with mutually orthogonal pilots, there is no pilot contamination and inter-user interference is effectively mitigated. However, in practice, the number of users tends to be much larger than the number of pilots and consequently, reusing pilots is inevitable [2, 19]. This results in very poor quality of channel estimation. Furthermore, having users share the same pilot also causes correlated interference, which can not be asymptotically mitigated with the large antenna effect. Due to this problem, the system's SE is saturated even when the number of antenna tends to infinity [7, 8].

On the other hand, to provide good SE for all users by assigning them with orthogonal pilots, the number of users which can be served simultaneously on the same radio resources will be strictly limited by the channel's coherence length [7, 22].

Unfavorable channel condition:

As previously discussed, the ability to mitigate inter-user interference of a massive MIMO system comes from the favorable propagation property. This property strongly depends on the correlation among users' channels. In practice, it is not possible to have perfect orthogonality between channel vectors of two users, which results in unfavorable propagation and negatively affects the massive MIMO system's performance [4, 7].

2.3 Fundamentals of NOMA

With the growing demand of higher data throughput and massive connectivity, conventional OMA schemes such as TDMA, FDMA, CDMA and OFDMA are unable to meet the requirements of future wireless networks. The key principle of these conventional OMA methods is to allocate orthogonal radio resources to different users. In TDMA, one time slot is occupied by only one user. In FDMA and OFDMA, a carrier frequency is allocated to only one user. In CDMA, the radio resources are represented as spreading codes, one of which is used for only one user at a time. OMA strictly limits the number of connections available with a fixed amount of resources. This motivates the use of a new multiple access technique that allows sharing/reusing the common radio resources, which is NOMA.

As opposed to OMA, where only one user occupies a resource unit, NOMA allows multiple users to share the common radio resources. This means time slots, carriers or spreading codes can be used by more than just one user. In this way, the number of connections is significantly greater than that in OMA with the same resources. This is promising to meet the massive connectivity requirement of future wireless networks. However, NOMA can result in severe inter-user interference due to reusing resources [15, 17]. Fortunately, this problem can be solved by a key signal processing technique that enable NOMA: *successive interference cancellation.*

2.3.1 Channel Capacity with Successive Interference Cancellation

To illustrate how NOMA works, consider a simple example where two users transmit their data symbols x_i, $(i = 1, 2)$ with the respective powers P_i, $(i = 1, 2)$ to a BS on fixed time-frequency resources. For simplicity, it is assumed that the channels h_i from users to the BS are perfect, i.e., $h_1 = h_2 = 1$. As a result, the signal received at the BS over one symbol period can be expressed follows:

$$y = \sqrt{P_1}h_1x_1 + \sqrt{P_2}h_2x_2 + n = \sqrt{P_1}x_1 + \sqrt{P_2}x_2 + n \tag{2.23}$$

where $n \sim \mathcal{CN}(0, \sigma^2)$ denotes AWGN noise. By treating the second user's signal as noise, the channel capacity of the first user is:

$$C_1 = \log_2\left(1 + \frac{P_1}{P_2 + \sigma^2}\right) \tag{2.24}$$

With the signal from the first user detected, the BS can subtract it from the received signal before detecting data of the second user, which results in the capacity:

$$C_2 = \log_2\left(1 + \frac{P_2}{\sigma^2}\right) \tag{2.25}$$

The sum capacity is therefore calculated by:

$$C_{\text{sum}}^{(\text{NOMA})} = C_1 + C_2 = \log_2\left(1 + \frac{P_1 + P_2}{\sigma^2}\right) \tag{2.26}$$

For a fair comparison with OMA, suppose the radio resources are divided with the weight factors of α_1 and α_2 with $\alpha_1 + \alpha_2 = 1$. The sum capacity of OMA in this case is:

$$\begin{aligned} C_{\text{sum}}^{(\text{OMA})} &= \alpha_1 \log_2\left(1 + \frac{P_1}{\alpha_1\sigma^2}\right) + \alpha_2 \log_2\left(1 + \frac{P_2}{\alpha_2\sigma^2}\right) \\ &\leq \log_2\left(1 + \frac{P_1 + P_2}{\sigma^2}\right) \end{aligned} \tag{2.27}$$

The above inequality comes from the arithmetic-geometric mean inequality. Thus, it can be seen that the sum channel capacity with NOMA is always greater or equal to that of OMA. The equality holds only when $P_1/P_2 = \alpha_1/\alpha_2$. In the case when $P_1 \gg P_2$ (i.e., when the signal from the first user is dominant), in order to achieve the same performance as NOMA, the OMA system has to allocate almost entire radio resources for the dominant user, which is unfair. Meanwhile, with NOMA, the second user, despite of having a weak signal, is still able to achieve the best performance.

2.3.2 Challenges with NOMA

Although NOMA can enhance the system's sum data rate as compared to OMA, there are important problems that need to be considered:

- The user who detects last will have to decode the signals from all previous users, which causes a significant overhead to the hardware as well as a large delay [11, 15].
- One critical problem of NOMA is how well the SIC is implemented. It is obvious that if the signals from users who get detected first are decoded correctly, a large amount of interference can be removed from the signals of the succeeding users. On the other hand, if the decoding step is carried out incorrectly, the SIC may result in more interference, which eventually degrades the system's performance [23].
- From the example discussed in the previous section, the gain of NOMA over OMA in terms of the sum rate is significant only when the channel conditions between the two users are significantly different. Otherwise, the gain is negligible. This leads to the importance of the grouping problem, whose objective is to find which users should be assigned to use the same orthogonal resource unit to achieve the best gain with NOMA over OMA. Of course, this also causes a large overhead to the system [15].

2.3.3 Integration of NOMA into Massive MIMO Systems

In a massive MIMO system, the users have already operated on the same time-frequency resources thanks to the asymptotically orthogonal property of the channels [2, 4, 7]. However, there is still an important resource that can be exploited: the pilot sequences. It has been shown that the number of pilots in a massive MIMO system is strictly limited by the coherence length, and that the more time spent for the UL training phase, the less time left for data transmission. This will limit the number of users admissible in the system. To deal with this problem, a NOMA approach can be applied to share pilots among all users and utilize the SIC technique to reduce interference [11–13].

Another problem when applying NOMA in a massive MIMO system is that it degrades the channel estimation quality due to pilot contamination. Hence, to integrate NOMA with a massive MIMO system, beside SIC, power allocation and user pairing/grouping are critical to reduce the effect of pilot contamination [12, 14]. All these respects of integrating NOMA into massive MIMO systems will be presented in the next chapters of this thesis.

References

[1] Cisco. "Cisco Annual Internet Report " (2018-2023). [Online]. Available: https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3590–3600, Nov. 2010.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. on Commun.*, vol. 61, pp. 1436–1449, Apr. 2013.

[4] E. Bjrnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, pp. 114–123, Feb. 2016.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?," *IEEE J. Sel. Areas in Commun.*, vol. 32, pp. 1065–1082, June 2014.

[6] H. V. Cheng, E. Björnson, and E. G. Larsson, "Uplink pilot and data power control for single cell massive mimo systems with mrc,"in *Proc. 2015 2015 International Symposium on Wireless Communication Systems (ISWCS)*, pp. 396–400, Brussels 2015.

[7] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO.* Cambridge University Press, 2016.

[8] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 1293–1308, Feb. 2016.

[9] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint pilot design and uplink power allocation in multi-cell massive MIMO systems," *IEEE Trans. on Wireless Commun.*, vol. 17, pp. 2000–2015, Mar. 2018.

[10] D. Hu, L. He, and X. Wang, “Semi-blind pilot decontamination for massive MIMO systems,” *IEEE Trans. on Wireless Commun.*, vol. 15, no. 1, pp. 525–536, 2016.

[11] Y. Li and G. A. Aruma Baduge, “NOMA-aided cell-free massive MIMO systems,” *IEEE Wireless Commun. Lett.*, vol. 7, pp. 950–953, Dec. 2018.

[12] D. Kudathanthirige and G. A. A. Baduge, “NOMA-aided multicell downlink massive MIMO,” *IEEE J. Sel. Topics in Signal Process.*, vol. 13, pp. 612–627, June 2019.

[13] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, “NOMA/OMA mode selection-based cell-free massive MIMO,” in *Proc. ICC 2019 - 2019 IEEE Int. Conf. Commun. (ICC)*, pp. 1–6, May 2019.

[14] F. Rezaei, A. R. Heidarpour, C. Tellambura, and A. Tadaion, “Underlaid spectrum sharing for cell-free massive MIMO-NOMA,” *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 907–911, April 2020

[15] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, “A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends,” *IEEE J. Sel. Areas in Commun.*, vol. 35, pp. 2181–2195, Oct. 2017.

[16] Z. Ding and H. V. Poor, “Design of massive-MIMO-NOMA with limited feedback,” *IEEE Signal Process. Lett.*, vol. 23, pp. 629–633, May 2016.

[17] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, “Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends,” *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, 2015.

[18] F. Rezaei, C. Tellambura, A. Tadaion, and A. R. Heidarpour, “Rate analysis of cell-free massive MIMO-NOMA with three linear precoders,” *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3480–3494, June 2020

[19] E. Björnson, E. G. Larsson, and M. Debbah, “Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?,” vol. 15, no. 2, pp. 1293–1308, 2016.

[20] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* PTR Prentice Hall, 1993.

[21] T. V. Chien and E. Björnson, “Massive MIMO communications,” in *5G Mobile Communications*, pp. 77–116, Springer International Publishing, oct 2016.

[22] X. Chen, R. Jia, and D. W. K. Ng, “On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation,” *IEEE Trans. Commun.*, vol. 67, pp. 2539–2551, Mar. 2019.

[23] P. Li, R. C. de Lamare, and R. Fa, “Multiple feedback successive interference cancellation detection for multiuser MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 10, pp. 2434–2439, Aug. 2011.

3. Multiuser Massive MIMO Systems with Time-Offset Pilots and Successive Interference Cancellation

Published as:

The Khai Nguyen, Ha H. Nguyen and Tien Hoa Nguyen, "Multiuser Massive MIMO Systems with Time-Offset Pilots and Successive Interference Cancellation", *IEEE Access*, vol. 7, pp. 132748–132762, 2019.

In Chapter 2, a single-cell massive MIMO system has been considered to explain the channel estimation stage with pilots in two cases: orthogonal and nonorthogonal pilots. This chapter examines a novel approach of arranging pilots, namely *time-offset pilots*. With time-offset pilots, instead of scheduling all users to transmit their pilots synchronously, the uplink training phase is designed such that a group of users transmits their pilots when another user group transmits their uplink data simultaneously. In this way, pilot contamination is not caused by the nonorthogonality between different users' pilot sequences, but between pilots of users in one group and data symbols from the other group. The development of the system model together with detailed analysis in the manuscript show that with this method, the system can server twice the number of users as compared to the conventional case of using orthogonal pilots, while correlated interference caused by pilot contamination can be effectively reduced with successive interference cancellation.

Multiuser Massive MIMO Systems with Time-Offset Pilots and Successive Interference Cancellation

The Khai Nguyen, Ha H. Nguyen, and Tien Hoa Nguyen

Abstract

This paper proposes time-offset pilots for a single-cell multiuser massive multiple-input multiple-output (MIMO) system and studies its performance under the minimum mean-squared error channel estimator and successive interference cancellation. With the proposed time-offset pilots, users are divided into two groups and the uplink pilots from one group are transmitted *simultaneously* with the uplink data of the other group, which allows the system to accommodate more users for a given number of pilots. Successive interference cancellation is developed to ease the effect of pilot contamination and enhance data detection. Closed-form expressions for lower bounds of the uplink spectral efficiencies in both the training and data phases are derived when the maximum-ratio combining receiver is used at the base station. The power control problem is formulated with the objective of either maximizing the quality of service that can be equally provided to all users, or minimizing the total transmit power. Since the original power control problems are NP-hard, we also propose algorithms based on the bisection method to solve the problems separately in training and data phases. Analysis and numerical results show that the effect of pilot contamination can be mitigated by successive interference cancellation and proper power control.

3.1 Introduction

Over the last decade, massive multiple-input multiple-output (MIMO) systems have gained a strong interest as a promising key technology for enabling the next and future generations of wireless communications. With hundreds of antennas equipped at each base station (BS), a massive MIMO system allows multiple users to simultaneously operate in the same time-frequency blocks, while co-channel interference can be effectively mitigated as a result of channel hardening and favorable propagation effects [1–4]. Furthermore, by utilizing proper power control algorithms, massive MIMO systems have the ability to achieve very high spectral efficiency (SE) and energy efficiency (EE) [5–7].

However, performance of a massive MIMO system is limited by the quality of channel estimation [8–10]. As discussed in these papers, in every coherence interval where the wireless channels between BSs and users are approximately constant, the number of symbols spent for channel estimation directly determines the maximum number of pairwise-orthogonal pilot sequences that can be generated for channel estimation. Conventionally and preferably, pilot sequences are designed to be mutually orthogonal and distinct pilot sequences are assigned to different users to avoid pairwise correlation between them. Unfortunately, the number of orthogonal pilot sequences could be limited by the small length of the coherence interval, especially when the propagation environment changes quickly. Therefore, if the number of users served by one BS keeps increasing, pilot sequences must be reused, resulting in the so-called pilot contamination [11–13]. As a consequence, with simple linear receivers such as maximum-ratio combining and zero-forcing, the network's SE becomes saturated, even when the number of antennas goes to infinity [14–17].

3.1.1 Related Works

There have been many research works addressing the pilot contamination problem in multi-cell massive MIMO systems [1, 18–21]. For multi-cell massive MIMO systems, the basic approach to reduce the effect of pilot contamination is reusing pilots [16, 19, 22, 23]. With this approach, an arbitrary pilot sequence can be assigned one time only within a cluster of ϑ cells. This has been investigated in [16] and it was shown that using a higher pilot reuse factor helps to lessen pilot contamination. It should be pointed out that, a larger value of ϑ implies that the cell size, as well as the number of users who can be served within each cell, are reduced. Another method to reduce the effect of pilot contamination in a multi-cell massive MIMO system is to use different pilot sets [18]. Specifically, from a basic mutually-orthogonal pilot set, the authors in [18] construct the so-called dictionary of linear combinations of the original pilot set to exploit the degree of freedom, which is demonstrated to lower the interference level during the training phase. With this method, non-orthogonal pilots are used even within a cell.

All the works discussed above are for multi-cell massive MIMO systems where users are geographically separated into a cellular topology and hence, pilots can be reused across cells with large distance separations [9]. On the other hand, the joint pilot and payload power control problem in a single-cell massive MIMO system is investigated in [24]. In this work, the authors show that the optimal number of pilots should be set equal to the number of users in the system because using orthogonal pilots maximizes the signal-to-interference-plus-noise ratio (SINR). However, when the number of users increases and/or the coherence interval is short (as seen in fast-varying channels), the total throughput inversely decreases with the number of pilots. Another work examining the pilot contamination problem can be found in [6]. In this paper, a cell-free massive MIMO system with multiple access points (APs) is considered. As explained in [6], during the training phase, a set of orthogonal pilots can be assigned to a larger number of users by using a greedy algorithm. This assignment was shown to provide an improvement of approximately 10% in spectral efficiency as compared to a random pilot assignment. However, via the large-antenna analysis, it is shown in [6, 25] that if a pilot sequence is assigned to more than one user, the SINR is still upper-bounded because not only the desire signal power, but also the correlated interference power caused by pilot contamination increases proportionally with the number of antennas.

In all the works discussed above, uplink (UL) pilots are transmitted at the same time for all users. This method is known as aligned pilots in [1] or synchronous pilots in [26]. Another method to deal with pilot contamination is using time-offset (or asynchronous) pilots [1, 21, 26–28]. In particular, the authors in [1, 21] propose to schedule UL pilots so that the pilot signaling of one cell can be carried out while other cells are transmitting downlink (DL) data. Using the large-antenna analysis, these papers show that with such a pilot design, when the number of antennas goes to infinity, the SINRs in both UL and DL increase proportionally with the number of antennas. In addition, the authors also point out that having users in one cell transmitting UL pilots while users in other cells are transmitting UL data is not optimal because the performance is saturated when the number of antennas goes up to infinity.

To address the disadvantage of transmitting UL pilots simultaneously with UL data, the authors in [26, 27] propose a semi-blind pilot decontamination scheme. In such a scheme, under the assumption of *time-invariant* channel, least-square estimation of the channel is obtained by UL pilot sequences and with the aid of UL data extraction. This method is shown to significantly improve the quality of channel estimation when the length of data increases. However, such an improvement is difficult to achieve in the case of *fast-varying* channels as demonstrated in [10]. In particular, the authors in [10] show that, in practice, in order to allow data transmission plus channel estimation, the number of users needs to be well below the coherence length. The authors then propose a blind pilot decontamination method in which the pilot data is not required to find a subspace projection, which is used to improve channel estimation. Other research works on combating pilot contamination with time-offset pilots for multi-cell massive MIMO can be found in [29, 30] which introduce new coherence block structures with extra intervals for BS channel estimation [29] or null transmission [30]. However, if the coherence length is short, spending more symbols for channel estimation may result in an insufficient time interval for data transmission [10].

Another emerging technique to accommodate more users without requiring extra pilots is beam-domain user grouping for massive MIMO [31–33]. In these papers, the authors introduce a beam-domain grouping method that assigns users into different groups based on the direction of arrival (DOA) and then reuse pilots in different groups. The channel vector's elements are assumed to be correlated with an array response vector, which allows a beam-domain presentation of the actual channel. With such a method, it is shown that the training resources can be reduced, whereas inter-group interference and self interference at the BS can be effectively mitigated thanks to the properties of the beam-domain channel.

3.1.2 Contribution

In this paper, we investigate a new approach with time-offset pilots in a single-cell massive MIMO system. For the system considered in this paper, all users are divided into two groups. During the training phase, one group transmits orthogonal pilot signals, while the other group sends data signals. The BS gathers all pilot signals and performs the minimum mean-square error (MMSE) channel estimation. With this method, channel estimation is not contaminated by correlation between pilots, but by the data transmitted by the other group, whose power is typically much lower than the pilot power. In addition, with a fixed number of pairwise orthogonal pilot sequences, this approach allows to double the number of users compared to the orthogonal pilot approach. Different from previous works, in which the pilot power is usually set at the maximal level to maximize the channel estimation quality [6, 34], or assigned based on a long-term average power constraint [24], our work takes into account both pilot power and data power to optimally allocate users' UL power to satisfy a predetermined cost function. Moreover, we also develop a successive interference cancellation method that does not require the perfect channel state information. The method is shown to be able to significantly suppress the interference caused by pilot contamination. Naturally, this advantage comes at the expense of higher implementation complexity. The main contributions of the paper are as follows:

- We derive a closed-form expression of the UL ergodic spectral efficiency for the proposed time-offset pilot method under Rayleigh fading channels and when the maximum ratio combining (MRC) is used at the BS. Many interesting observations concerning the effects of array gain, interference, and additive noise are revealed.

- We develop a successive interference cancellation method for the detection of UL data at the BS to mitigate the impact of pilot contamination in UL transmission. Under the assumption of ideal error-free detection, it is shown that the UL SE is no longer bounded when the number of antennas increases.

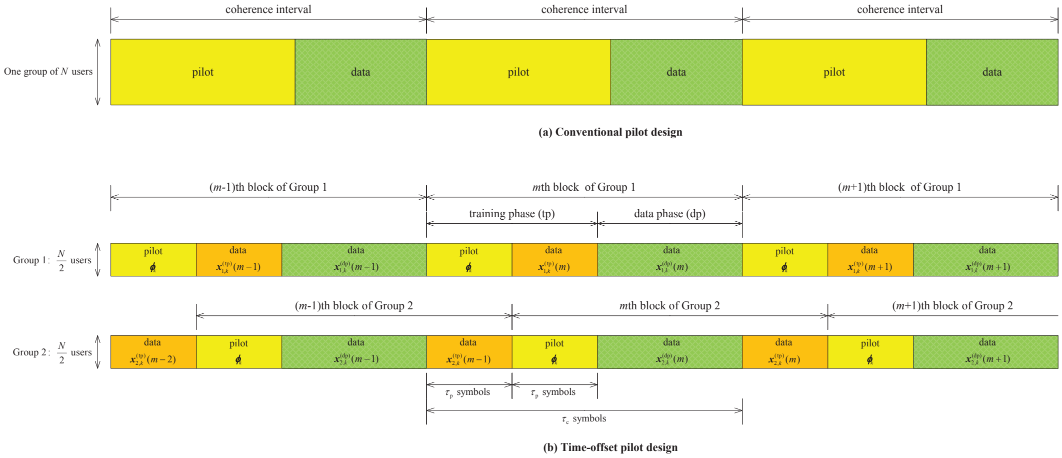
- We formulate and solve the power control problem for two different cost functions: the first problem focuses on maximizing the minimum quality of service (QoS) or max-min QoS, whereas the second problem is on total power minimization. Because of the NP-hardness of the original problems, we propose algorithms based on the bisection method to decompose these NP-hard problems into two subproblems which can be solved in polynomial time.

- A group assignment method is also proposed to mitigate the interference that cannot be removed by the MRC. The proposed group assignment helps to further improve the

UL ergodic spectral efficiency.

The remainder of this paper is organized as follows. Section II presents the model of a single-cell multiuser massive MIMO system with time-offset pilots and channel estimation. Section III analyzes UL spectral efficiencies in both training and data phases. Section IV studies power control problems. Section V provides simulation results and discussion. Section VI concludes the paper.

Notations: Vectors are formatted in bold lower-case, matrices are in bold upper-case. The transpose and conjugate transpose are denoted with superscripts T and H, respectively. The $K \times K$ identity matrix is $\boldsymbol{I}_K$. The operator $\mathbb{E}\{\cdot\}$ denotes the expectation of a random variable. The notation $\|\cdot\|$ stands for the Euclidean norm and $\mathrm{tr}(\cdot)$ represents the trace of a matrix. The notation $\boldsymbol{n} \sim \mathcal{CN}(0, \boldsymbol{C})$ means that $\boldsymbol{n}$ is a zero-mean complex Gaussian vector with covariance matrix $\boldsymbol{C}$.


coherence interval
coherence interval
coherence interval
One group of N users
pilot
data
pilot
data
pilot
data
(m-1)th block of Group 1
mth block of Group 1
(m+1)th block of Group 1
training phase (tp)
data phase (dp)
Group 1: $\frac{N}{2}$ users
pilot ϕ_k
data $\boldsymbol{x}_{1,k}^{(\mathrm{tp})}(m-1)$
data $\boldsymbol{x}_{1,k}^{(\mathrm{dp})}(m-1)$
pilot ϕ_k
data $\boldsymbol{x}_{1,k}^{(\mathrm{tp})}(m)$
data $\boldsymbol{x}_{1,k}^{(\mathrm{dp})}(m)$
pilot ϕ_k
data $\boldsymbol{x}_{1,k}^{(\mathrm{tp})}(m+1)$
data $\boldsymbol{x}_{1,k}^{(\mathrm{dp})}(m+1)$
(m-1)th block of Group 2
mth block of Group 2
(m+1)th block of Group 2
Group 2: $\frac{N}{2}$ users
data $\boldsymbol{x}_{2,k}^{(\mathrm{tp})}(m-2)$
pilot ϕ_k
data $\boldsymbol{x}_{2,k}^{(\mathrm{dp})}(m-1)$
data $\boldsymbol{x}_{2,k}^{(\mathrm{tp})}(m-1)$
pilot ϕ_k
data $\boldsymbol{x}_{2,k}^{(\mathrm{dp})}(m)$
data $\boldsymbol{x}_{2,k}^{(\mathrm{tp})}(m)$
pilot ϕ_k
data $\boldsymbol{x}_{2,k}^{(\mathrm{dp})}(m+1)$
τ_p symbols
τ_p symbols
τ_c symbols

(a) Conventional pilot design

(b) Time-offset pilot design

Figure 3.1 (a) Conventional pilot design, and (b) Time-offset pilot design.

3.2 Time-Offset Pilots and Channel Estimation

Consider a single-cell multi-user massive MIMO system in which one M-antenna BS serves N users, who are randomly distributed over the cell. The channels between the users and the BS are assumed to be frequency flat and approximately constant within a coherence interval of length τ_c symbols. The UL and DL transmissions in the system operate in time-division duplex (TDD) mode. As a result, conventional pilot designs can take advantage of channel reciprocity to estimate both UL and DL channels within a coherence interval. In massive MIMO systems, pilot sequences are usually transmitted synchronously by all users at the same time. This is problematic if the coherence interval is short, since to maintain orthogonal pilots, a smaller number of symbol periods can be used for data transmission. Motivated by the work in [6], we consider time-offset pilot design as illustrated in Fig. 3.1. Here, N users in the system are separated into $G = 2$ groups, each having $K = N/2$ users and taking turn to transmit UL pilots in different time slots. To improve the SE, the transmission of UL pilots by one group happens concurrently with UL data transmission from the other group. An important point to note is that pilot transmission must be carried out at the beginning of every coherence interval.

Dropping the block index for simplicity and without loss of generality, the $M \times 1$ received signal vector at the BS in one symbol time can be generally written as:

$$\boldsymbol{y} = \sum_{k=1}^{K} \left(\boldsymbol{h}_{1,k} \sqrt{p_{1,k}} x_{1,k} + \boldsymbol{h}_{2,k} \sqrt{p_{2,k}} x_{2,k} \right) + \boldsymbol{n}, \tag{3.1}$$

where $x_{g,k}$ $(g = 1, 2)$ is the transmit signal of the kth user in the gth group that is normalized to have unit power, i.e., $\mathbb{E}\left\{|x_{g,k}|^2\right\} = 1$, whereas the actual transmit power is specified by $p_{g,k}$. Note that $x_{g,k}$ represents either the data or the pilot symbol during the training phase (see the illustration in Fig. 3.1). The term $\boldsymbol{n} \sim \mathcal{CN}\left(0, \sigma^2 \boldsymbol{I}_M\right)$ models additive white Gaussian noise (AWGN) at the BS. The channels are assumed to be uncorrelated Rayleigh fading. That is, the channel vector $\boldsymbol{h}_{g,k}$ from the kth user of the gth group to the BS is modeled as having a circularly-symmetric complex Gaussian distribution, $\boldsymbol{h}_{g,k} \sim \mathcal{CN}\left(0, \beta_{g,k} \boldsymbol{I}_M\right)$, where $\beta_{g,k}$ represents large-scale fading.

The BS does not know the exact channel coefficients but the channel statistics. To estimate the channels for each user group, a set of K length-τ_p pilot sequences is used. These pilots are collectively represented by a $\tau_p \times K$ pilot matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K]$ which satisfies $\boldsymbol{\Phi}^H \boldsymbol{\Phi} = \tau_p \boldsymbol{I}_K$. Usually, the pilot length is set at the minimum value $\tau_p = K$ in order to achieve the orthogonality between pilot sequences.

Without loss of generality, suppose that the first group transmits pilots first at the beginning of the training phase, while the other group transmits data. Then the signal matrix $\boldsymbol{Y} \in \mathbb{C}^{M \times \tau_p}$ received at the BS over τ_p time slots (symbols) is given as:

$$\boldsymbol{Y} = \sum_{k=1}^{K} \left(\boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{p})}} \boldsymbol{\phi}_k^H + \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \right) + \boldsymbol{N}, \tag{3.2}$$

where $\rho_{1,k}^{(\mathrm{p})}$ is the pilot power, $\rho_{2,k}^{(\mathrm{d})}$ is the power assigned to the normalized data signal vector $\boldsymbol{x}_{2,k}^{(\mathrm{tp})} \in \mathbb{C}^{1 \times \tau_p}$ of the kth user in the second group during τ_p time slots of the training phase, which satisfies $\boldsymbol{x}_{2,k}^{(\mathrm{tp})} \sim \mathcal{CN}\big(0, \boldsymbol{I}_{\tau_p}\big)$.

To estimate the channel from the qth user in the first group, the BS multiplies the received signal with the corresponding pilot of the qth user. This results in:

$$\boldsymbol{r}_{1,q} = \boldsymbol{Y} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{p})} \tau_p} + \sum_{k=1}^{K} \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} + \boldsymbol{N} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}. \tag{3.3}$$

Then, the estimate of $\boldsymbol{h}_{1,q}$ can be obtained by using the MMSE estimator [35] as:

$$\hat{\boldsymbol{h}}_{1,q} = \frac{\operatorname{cov}\left\{\boldsymbol{h}_{1,q}, \boldsymbol{r}_{1,q}\right\}}{\operatorname{var}\left\{\boldsymbol{r}_{1,q}\right\}} \boldsymbol{r}_{1,q} = \mu_{1,q} \boldsymbol{r}_{1,q}, \tag{3.4}$$

where

$$\mu_{1,q} = \frac{\sqrt{\rho_{1,q}^{(\mathrm{p})} \tau_p} \beta_{1,q}}{\rho_{1,q}^{(\mathrm{p})} \tau_p \beta_{1,q} + \sum_{k=1}^{K} \rho_{2,k}^{(\mathrm{d})} \beta_{2,k} + \sigma^2}.$$

As a result, the estimated channel is a random vector, which follows the distribution $\hat{\boldsymbol{h}}_{1,q} \sim \mathcal{CN}\left(0, \gamma_{1,q} \boldsymbol{I}_M\right)$, where

$$\gamma_{1,q} = \frac{\rho_{1,q}^{(\mathrm{p})} \tau_p \beta_{1,q}^2}{\rho_{1,q}^{(\mathrm{p})} \tau_p \beta_{1,q} + \sum_{k=1}^{K} \rho_{2,k}^{(\mathrm{d})} \beta_{2,k} + \sigma^2}. \tag{3.5}$$

Furthermore, the estimation error $\boldsymbol{e}_{1,q} = \boldsymbol{h}_{1,q} - \hat{\boldsymbol{h}}_{1,q}$ is independent of the estimated channel and distributed as $\boldsymbol{e}_{1,q} \sim \mathcal{CN}\left(0, (\beta_{1,q} - \gamma_{1,q}) \boldsymbol{I}_M\right)$.

After obtaining the channel estimation, the BS applies a linear processing vector for the detection of the UL data belonging to the same user. By employing the maximum ratio combining (MRC), the combining vector is given as:

$$\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q}. \tag{3.6}$$

For the second group, the same channel estimation process applies, but with the roles of the two groups reversed.

3.3 Uplink Data Transmission

3.3.1 Analysis in the training phase

To examine data detection in the training phase, focus on the time slots over which the first group transmits UL data while the second group transmits UL pilots for channel estimation. The signal received at the BS over one symbol can be rewritten from (3.1) as:

$$\boldsymbol{y}^{(\mathrm{tp})} = \sum_{k=1}^{K} \boldsymbol{h}_{1,k}\sqrt{\rho_{1,k}^{(\mathrm{d})}}x_{1,k}^{(\mathrm{tp})} + \sum_{k=1}^{K} \boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{p})}}\phi_k + \boldsymbol{n}, \tag{3.7}$$

where ϕ_k simply denotes one entry of the pilot vector $\boldsymbol{\phi}_k$. To detect data of the qth user of the first group, the BS multiplies the above received signal with the corresponding combining vector $\boldsymbol{v}_{1,q}$ as specified in (3.6). This yields

$$\boldsymbol{v}_{1,q}^{H}\boldsymbol{y}^{(\mathrm{tp})} = \sum_{k=1}^{K} \boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,k}\sqrt{\rho_{1,k}^{(\mathrm{d})}}x_{1,k}^{(\mathrm{tp})} + \sum_{k=1}^{K} \boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{p})}}\phi_k + \boldsymbol{v}_{1,q}^{H}\boldsymbol{n}. \tag{3.8}$$

To see how the desired data is affected by different components, decompose the signal in (3.8) as:

$$\begin{aligned}\boldsymbol{v}_{1,q}^{H}\boldsymbol{y}^{(\mathrm{tp})} = {} & \underbrace{\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{d})}}x_{1,q}^{(\mathrm{tp})}}_{\mathrm{DS}_{1,q}^{(\mathrm{tp})}\text{ – Desired signal}} + \underbrace{\sum_{k=1,k\neq q}^{K} \boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,k}\sqrt{\rho_{1,k}^{(\mathrm{d})}}x_{1,k}^{(\mathrm{tp})}}_{\mathrm{IwG}_{1,q}^{(\mathrm{tp})}\text{ – Interference within group}} \\ & + \underbrace{\sum_{k=1}^{K} \boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{p})}}\phi_k}_{\mathrm{IP}_{1,q}^{(\mathrm{tp})}\text{ – Interference from pilot}} + \underbrace{\boldsymbol{v}_{1,q}^{H}\boldsymbol{n}}_{\mathrm{N}_{1,q}^{(\mathrm{tp})}\text{ – Noise}} .\end{aligned} \tag{3.9}$$

The first component in (3.9) is the desired signal for the detection of data $x_{1,q}^{(\mathrm{tp})}$. The second term accounts for interference from users in the same group. The terms $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ quantifies the interference from pilot transmissions conducted by users in the second group. The last component in (3.9) is filtered additive Gaussian noise.

Next, consider the case that the MRC is used at the BS, i.e., $\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q} = \mu_{1,q}\boldsymbol{Y}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}$. Given the distribution of the channel estimate $\hat{\boldsymbol{h}}_{1,q} \sim \mathcal{CN}\left(0, \gamma_{1,q}\boldsymbol{I}_M\right)$, the following analyzes the behavior of each term in (3.9) when $M \to \infty$.

First, the desired signal component is

$$\begin{aligned}\mathrm{DS}_{1,q}^{(\mathrm{tp})} &= \left(\mu_{1,q}\boldsymbol{Y}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{d})}}x_{1,q}^{(\mathrm{tp})} \\ &= \Bigg[\mu_{1,q}\boldsymbol{h}_{1,q}^H\boldsymbol{h}_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{p})}\rho_{1,q}^{(\mathrm{d})}}\tau_p + \mu_{1,q}\sum_{k=1}^{K}\boldsymbol{h}_{2,k}^H\boldsymbol{h}_{1,q}\sqrt{\rho_{2,k}^{(\mathrm{d})}\rho_{1,q}^{(\mathrm{d})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \\ &+ \mu_{1,q}\left(\boldsymbol{N}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{d})}}\Bigg]x_{1,q}^{(\mathrm{tp})}.\end{aligned} \tag{3.10}$$

Due to the fact that the channels from the BS to all users are mutually independent, by applying the law of large numbers, the second and third components in (3.10) go to zero when M goes to infinity. It follows that

$$\frac{1}{M}\mathrm{DS}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\mathrm{a.s}} \mu_{1,q}\beta_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{p})}\rho_{1,q}^{(\mathrm{d})}}\tau_p x_{1,q}^{(\mathrm{tp})}, \tag{3.11}$$

where the notation $\xrightarrow[M\to\infty]{\mathrm{a.s}}$ means almost sure convergence as $M \to \infty$. On the other hand, due to fact that all the components of $\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q}$ are statistically independent of $\boldsymbol{h}_{1,k}$ for all $k \neq q$, $\mathrm{IwG}_{1,q}^{(\mathrm{tp})}$ and $\mathrm{N}_{1,q}^{(\mathrm{tp})}$ vanish when $M \to \infty$. That is,

$$\frac{1}{M}\mathrm{IwG}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\mathrm{a.s}} 0, \tag{3.12}$$

and

$$\frac{1}{M}\mathrm{N}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\mathrm{a.s}} 0. \tag{3.13}$$

Next, the interference term $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ that originates from the second group which transmits UL pilots is decomposed as:

$$\begin{aligned}&\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{p})}}\phi_k \\ &= \left(\mu_{1,q}\boldsymbol{Y}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{p})}}\phi_k = \mu_{1,q}\boldsymbol{h}_{1,q}^H\boldsymbol{h}_{2,k}\sqrt{\rho_{1,q}^{(\mathrm{p})}\rho_{2,k}^{(\mathrm{p})}}\tau_p\phi_k \\ &+ \mu_{1,q}\sum_{k=1}^{K}\boldsymbol{h}_{2,k}^H\boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{d})}\rho_{2,k}^{(\mathrm{p})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\phi_k + \mu_{1,q}\left(\boldsymbol{N}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{d})}}\phi_k.\end{aligned} \tag{3.14}$$

The first and third terms of (3.14) also vanish when $M \to \infty$. Therefore, the remaining term is:

$$\frac{1}{M}\mathrm{IP}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\mathrm{a.s}} \sum_{k=1}^{K}\mu_{1,q}\beta_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{d})}\rho_{2,k}^{(\mathrm{p})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\phi_k. \tag{3.15}$$

In summary, the above analysis shows that, when the number of antennas at the BS goes to infinity, the received signal for the qth user of the first group consists of the desired signal component as in (3.11) and the interference caused by users of the other group as a result of pilot contamination during its training phase (3.15).

The presence of $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ in (3.15) is due to the correlation between the channel estimation errors of the pilot-transmitting group and the received signals of the data-transmitting group. The impact of this interference can be reduced by applying the following interference cancellation method. At first, it can be seen from (3.14) that the part in $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ that remains when $M \to \infty$ is:

$$\Upsilon_{1,q}^{(\mathrm{IP})} = \mu_{1,q} \sum_{k=1}^{K} \boldsymbol{h}_{2,k}^{H} \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k. \tag{3.16}$$

Since the UL transmit power and UL pilot sequences are known and the UL signal $\boldsymbol{x}_{2,k}^{(\mathrm{tp})}$ was already detected first, the term $\Upsilon_{1,q}^{(\mathrm{IP})}$ can be estimated by replacing $\boldsymbol{h}_{2,k}^{H} \boldsymbol{h}_{2,k}$ with its statistical average. That is,

$$\hat{\Upsilon}_{1,q}^{(\mathrm{IP})} = \mu_{1,q} \sum_{k=1}^{K} \mathbb{E}\left\{\|\boldsymbol{h}_{2,k}\|^2\right\} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k. \tag{3.17}$$

The above estimated value can then be subtracted from the received signal of the qth user in the first group (see (3.9)), which should reduce the interference caused by pilot contamination. The result after performing interference cancellation in (3.9) is:

$$\begin{aligned} s_{1,q} = \boldsymbol{v}_{1,q}^{H} \boldsymbol{y}^{(\mathrm{tp})} - \hat{\Upsilon}_{1,q}^{(\mathrm{IP})} &= \boldsymbol{v}_{1,q}^{H} \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{d})}} x_{1,q}^{(\mathrm{tp})} + \sum_{k=1, k\neq t}^{K} \boldsymbol{v}_{1,q}^{H} \boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{d})}} x_{1,k}^{(\mathrm{tp})} \\ &+ \underbrace{\sum_{k=1}^{K} \left(\boldsymbol{v}_{1,q}^{H} \boldsymbol{h}_{2,k} - \mu_{1,q} \mathbb{E}\left\{\|\boldsymbol{h}_{2,k}\|^2\right\} \frac{\sqrt{\rho_{2,k}^{(\mathrm{d})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \right)}_{\mathrm{IP}_{1,q}^{(\mathrm{tp})} - \hat{\Upsilon}_{1,q}^{(\mathrm{IP})}} \sqrt{\rho_{2,k}^{(\mathrm{p})}} \phi_k + \boldsymbol{v}_{1,q}^{H} \boldsymbol{n}. \end{aligned} \tag{3.18}$$

It can be seen from (3.17) that:

$$\frac{1}{M} \hat{\Upsilon}_{1,q}^{(\mathrm{IP})} \xrightarrow[M \to \infty]{\mathrm{a.s}} \sum_{k=1}^{K} \mu_{1,q} \beta_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k, \tag{3.19}$$

which means that the term $\left(\mathrm{IP}_{1,q}^{(\mathrm{tp})} - \hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right)$ converges to zero when $M \to \infty$, hence pilot contamination can be removed. Thus, as M goes to infinity, only the desired signal component $\mathrm{DS}_{1,q}^{(\mathrm{tp})}$ remains in (3.18).

$$\mathrm{SINR}_{1,q}^{(\mathrm{tp,\ MRC})} = \frac{M\rho_{1,q}^{(\mathrm{d})}\gamma_{1,q}}{\sum_{k=1}^{K}\left(\rho_{1,k}^{(\mathrm{d})}\beta_{1,k} + \rho_{2,k}^{(\mathrm{p})}\beta_{2,k}\right) + \sum_{k=1}^{K}\rho_{2,k}^{(\mathrm{p})}\frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}\tau_p}\gamma_{1,q}\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^2 + \sigma^2}. \qquad (3.21)$$

Next, Theorem 1 gives a closed-form expression for a lower bound of the UL spectral efficiency for the qth user of the first group when MRC is used at the BS. Note that this result is valid for finite M.

Theorem 1: The UL spectral efficiency of the qth user in the first group in the training phase with MRC at the BS and successive interference cancellation is given as:

$$R_{1,q}^{(\mathrm{tp})} \geq \log_2(1 + \mathrm{SINR}_{1,q}^{(\mathrm{tp,\ MRC})}), \qquad (3.20)$$

where $\mathrm{SINR}_{1,q}^{(\mathrm{tp,\ MRC})}$ is given as in (3.21).

Proof: See the Appendix 3.A. It should be pointed out that, the same result applies to users in the second group during its training phase.

From (3.21), one can see that the array gain is proportional to the number of antennas, while the power of interference in the denominator is independent of the number of antennas. In particular, the denominator consists of two components: (i) uncorrelated interference, whose power equals to the signal power of all users received at the BS and noise power, and (ii) correlated interference caused by users in the second group as a result of pilot contamination.

3.3.2 Analysis in the Data Phase

In the data phase, both groups transmit their UL data. The received signal in the data phase is given as in (3.1) by substituting $x_{g,k} = x_{g,k}^{(\mathrm{dp})}$ (for $g = 1, 2$). Similar to the training phase, after applying a combining vector $\boldsymbol{v}_{1,q}$, the received signal of the qth user from the first group is decomposed as:

$$\begin{aligned}\boldsymbol{v}_{1,q}^H\boldsymbol{y}^{(\text{dp})} &= \sum_{k=1}^{K}\left(\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,k}\sqrt{p_{1,k}}x_{1,k}^{(\text{dp})} + \boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\sqrt{p_{2,k}}x_{2,k}^{(\text{dp})}\right) + \boldsymbol{v}_{1,q}^H\boldsymbol{n} \\ &= \underbrace{\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\sqrt{p_{1,q}}x_{1,q}^{(\text{dp})}}_{\text{DS}_{1,q}^{(\text{dp})}\text{ -- Desired signal}} + \underbrace{\sum_{k=1,k\neq q}^{K}\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,k}\sqrt{p_{1,k}}x_{1,k}^{(\text{dp})}}_{\text{IwG}_{1,q}^{(\text{dp})}\text{ -- Interference within group}} \\ &+ \underbrace{\sum_{k=1}^{K}\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\sqrt{p_{2,k}}x_{2,k}^{(\text{dp})}}_{\text{IoG}_{1,q}^{(\text{dp})}\text{ -- Interference from other group}} + \underbrace{\boldsymbol{v}_{1,q}^H\boldsymbol{n}}_{\text{N}_{1,q}^{(\text{dp})}\text{ -- Noise}}.\end{aligned} \tag{3.22}$$

Unlike the training phase, there is no interference caused by pilot transmission of the other group. Instead, there is interference, denoted as $\text{IoG}_{1,q}^{(\text{dp})}$, caused by concurrent data transmission from the other group. Following the same analysis as in the training phase, the terms $\text{IwG}_{1,q}^{(\text{dp})}$ and $\text{N}_{1,q}^{(\text{dp})}$ vanish when M goes to infinity. The only terms remained in (3.22) are the desired signal component,

$$\frac{1}{M}\text{DS}_{1,q}^{(\text{dp})} \xrightarrow[M\to\infty]{\text{a.s}} \mu_{1,q}\beta_{1,q}\sqrt{\rho_{1,q}^{(\text{p})}p_{1,q}\tau_p}x_{1,q}^{(\text{dp})}, \tag{3.23}$$

and interference from users in the other group:

$$\frac{1}{M}\text{IoG}_{1,q}^{(\text{dp})} \xrightarrow[M\to\infty]{a.s} \sum_{k=1}^{K}\mu_{1,q}\beta_{2,k}\sqrt{\rho_{2,k}^{(\text{d})}p_{2,k}}\boldsymbol{x}_{2,k}^{(\text{tp})}\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}x_{2,k}^{(\text{dp})}. \tag{3.24}$$

Different from the training phase, where the correlated interference from the pilot-transmitting group can be subtracted from the received signal of the data-transmitting group, the data signals $x_{1,k}^{(\text{dp})}$ and $x_{2,k}^{(\text{dp})}$ in (3.24) are unknown, and therefore the interference cancellation method cannot be applied in the same way as in the training phase. However, assuming that the signal from the second group, $x_{2,k}^{(\text{dp})}$, is detected first by treating $x_{1,k}^{(\text{dp})}$ as noise, then one can subtract an estimated value of $\text{IoG}_{1,q}^{(\text{dp})}$ from the received signal of the first group. This successive cancellation has been investigated in [36] and [37] and shown to significantly improve the minimal SINR value of users.

With the knowledge of $x_{2,k}^{(\text{dp})}$, an estimation of $\text{IoG}_{1,q}^{(\text{dp})}$ can be formed as:

$$\hat{\Upsilon}_{1,q}^{((\text{IoG})} = \mu_{1,q}\sum_{k=1}^{K}\mathbb{E}\left\{\|\boldsymbol{h}_{2,k}\|^2\right\}\sqrt{\rho_{2,k}^{(\text{d})}p_{2,k}}\boldsymbol{x}_{2,k}^{(\text{tp})}\frac{\boldsymbol{\phi}_t}{\|\boldsymbol{\phi}_t\|}x_{2,k}^{(\text{dp})}, \tag{3.25}$$

$$\mathrm{SINR}_{1,q}^{(\mathrm{dp,MRC})} = \frac{M p_{1,q}\gamma_{1,q}}{\sum_{k=1}^{K}\left(p_{1,k}\beta_{1,k}+p_{2,k}\beta_{2,k}\right)+\sum_{k=1}^{K} p_{2,k}\frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}\tau_p}\gamma_{1,q}\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^2+\sigma^2}. \tag{3.28}$$

$$\mathrm{SINR}_{2,q}^{(\mathrm{dp,MRC})} = \frac{M p_{2,q}\gamma_{2,q}}{\sum_{k=1}^{K}\left(p_{1,k}\beta_{1,k}+p_{2,k}\beta_{2,k}\right)+\sum_{k=1}^{K} p_{1,k}\frac{\rho_{1,k}^{(\mathrm{d})}}{\rho_{2,q}^{(\mathrm{p})}\tau_p}\gamma_{2,q}\left(\frac{\beta_{1,k}}{\beta_{2,q}}\right)^2(M+1)+\sigma^2}. \tag{3.29}$$

By subtracting (3.25) from (3.22), the received signal corresponding to the qth user in the first group now becomes:

$$\begin{aligned}\boldsymbol{v}_{1,q}^H\boldsymbol{y}^{(\mathrm{dp})}-\hat{\Upsilon}_{1,q}^{(\mathrm{IoG})} &= \boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{d})}}x_{1,q}^{(\mathrm{tp})}+\sum_{k=1,k\neq t}^{K}\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,k}\sqrt{\rho_{1,k}^{(\mathrm{d})}}x_{1,k}^{(\mathrm{tp})}\\ &+\sum_{k=1}^{K}\left(\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}-\mu_{1,q}\mathbb{E}\left\{\|\boldsymbol{h}_{2,k}\|^2\right\}\frac{\sqrt{\rho_{2,k}^{(\mathrm{d})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)\\ &\times\sqrt{p_{2,k}}x_{2,k}^{(\mathrm{dp})}+\boldsymbol{v}_{1,q}^H\boldsymbol{n}.\end{aligned} \tag{3.26}$$

Based on (3.26), a lower bound on the UL SE of the qth user of the first group when the MRC is employed at the BS during the data phase can be obtained as in Theorem 2.

Theorem 2: The UL spectral efficiency of the qth user in the gth group ($g = 1, 2$) when the MRC is employed at the BS in the data phase is given as:

$$R_{g,q}^{(\mathrm{dp})} \geq \log_2(1+\mathrm{SINR}_{g,q}^{(\mathrm{dp,MRC})}), \tag{3.27}$$

where the SINRs of the qth users of the first and second groups can be calculated as in (3.28) and (3.29), respectively.

Proof: The proof follows the same steps as carried out for proving the UL spectral efficiency in the training phase.

From (3.29), it can be seen that the correlated interference originating from pilot contamination in the denominator is proportional to $(M+1)$. As a consequence, for the second group, this component does not vanish when the number of antennas goes to infinity, unless the data power in the training phase is set to zero (equivalently, no data is transmitted in the training phase). Based on this observation, an adaptive power control method is proposed in the next section to optimize the UL data rate.

Before closing this section, it is worth pointing out that how to assign users into two different groups (i.e., group assignment method) can affect the spectral efficiency. In general, it is desired to optimally assign users into two groups such that the highest SE can be obtained. With time-offset pilots, group assignment impacts performance in both training phase and data phase, and not in the same way.

Unfortunately, optimizing group assignment is a combinatorial problem and, therefore, difficult to find the optimal solution. As discussed at the end of Section 3.1.1, a beam-domain group assignment approach was proposed in [31–33] which assigns users to different groups based on DOA. However, this approach is not applicable for the system model considered in this paper in which the channel vector's elements are mutually independent and hence cannot exploit the advantage of the beam-domain channel presentation. In this paper, in order to remove as much correlated interference as possible, we instead consider a group assignment exploiting the large-scale fading conditions of users. In this method, users in the cell are divided into inner and outer regions based on their locations. Since users in the inner region have generally better channel conditions compared to users in the outer region, they are assigned to the second group, whose data is detected first as described in Section III-B. The other users belong to the first group.

Before closing this section, it should be pointed out that the proposed time-offset pilot approach can be extended to more than 2 groups. In such a design, users are also grouped based on large-scale fading by dividing the coverage area into ring regions with different radii. The users in the group experiencing better channels will have their signals detected first, followed by users in the group having the next best channels, and so on. As a result, an arbitrary group can remove the known UL signals of all groups which have been already detected before by using successive interference cancellation (SIC). With more than 2 groups, the training phase needs to be divided into more sub-intervals, each one for one group to transmit its UL pilots. This implies that a more complicated power control method is required. Given the more severe pilot contamination and the higher complexity of interference cancellation if the system is designed with more than 2 groups, considering only 2 groups in the proposed approach appears most attractive and practical.

3.4 Power Control

This section studies power control problems under two cost functions: max-min QoS and total power minimization. The approach to solve these two problems is to decompose the original problem into two subproblems corresponding to two phases (training and data) in one coherence interval.

3.4.1 Max-Min QoS Optimization

Consider the optimization problem in which the cost function is to maximize the QoS value that can be equally provided to all users in the system. In the considered system model with time-offset pilots, data transmission of each user happens in both training and data phases. The training phase lasts for τ_p time slots and has a SE of $R_{g,q}^{(\mathrm{tp})}$ $(g = 1, 2)$. The data phase has a SE of $R_{g,q}^{(\mathrm{dp})}$ and is over $(1 - 2\tau_p)$ time slots. As a result, the average UL SE in one coherence interval of τ_c time slots can be calculated as:

$$R_{g,q}^{(\mathrm{total})} = \frac{\tau_p}{\tau_c} R_{g,q}^{(\mathrm{tp})} + \left(1 - \frac{2\tau_p}{\tau_c}\right) R_{g,q}^{(\mathrm{dp})}. \tag{3.30}$$

With the above UL spectral efficiency, the max-min QoS optimization problem is formulated as:

$$\begin{aligned} \underset{\rho_{g,q}^{(\mathrm{p})}, \rho_{g,q}^{(\mathrm{d})}, p_{g,q}}{\text{maximize}} \ & \min_{q=1,\dots,K} \left\{ R_{1,q}^{(\mathrm{total})}, R_{2,q}^{(\mathrm{total})} \right\} \\ \text{subject to} \quad & 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \forall g, q, \\ & 0 \le p_{g,q} \le p_{\max}, \forall g, q, \end{aligned} \tag{3.31}$$

where the objective is to maximize the minimum QoS and the constraints are to limit the data and pilot powers under a predetermined maximum UL transmit power $p_{\max}$.

The above optimization problem has the same form as the max-sum SE optimization problem studied in [38], which is a signomial programming and proved to be NP-hard [39].

Therefore, an algorithm based on the bisection method is proposed here, which iteratively solves the max-min QoS problem in training phase and data phase, separately.

In Training Phase: The power allocation problem to maximize min-QoS with UL transmit power constraints in the training phase can be formulated as:

$$\begin{aligned} &\underset{\rho_{g,q}^{(\mathrm{p})},\rho_{g,q}^{(\mathrm{d})}}{\text{maximize}} \min_{q=1,\ldots,K} \{R_{1,q}, R_{2,q}\} \\ &\text{subject to} \quad 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \forall g, q, \\ &\qquad\qquad\quad 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \forall g, q. \end{aligned} \tag{3.32}$$

The epigraph form of the above problem is:

$$\begin{aligned} &\underset{\rho_{g,q}^{(\mathrm{p})},\rho_{g,q}^{(\mathrm{d})},\lambda^{(\mathrm{tp})}}{\text{maximize}} \quad \lambda^{(\mathrm{tp})} \\ &\text{subject to} \quad \mathrm{SINR}_{g,q}^{(\mathrm{tp})} \ge \lambda^{(\mathrm{tp})}, \forall g, q, \\ &\qquad\qquad\quad 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \forall g, q, \\ &\qquad\qquad\quad 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \forall g, q. \end{aligned} \tag{3.33}$$

By dividing both the nominator and denominator of $\mathrm{SINR}_{g,q}^{(\mathrm{tp})}$ to $\gamma_{g,q}$, the first constraint of this problem can be converted into a valid constraint of geometric programming (GP) where the left-hand side of the “greater-than” inequality is a monomial and the right-hand side is a posynomial. As a result, this GP can be solved in polynomial time by using GP solvers like MOSEK solver with CVX [39, 40].

In Data Phase: With the power allocation strategy obtained in the training phase, the value of $\gamma_{g,k}$ can be calculated as in (3.5). Similar to the training phase, the max-min QoS power control in the data phase can be formulated in an epigraph form as:

$$\begin{aligned} &\underset{p_{g,q},\lambda^{(\mathrm{dp})}}{\text{maximize}} \quad \lambda^{(\mathrm{dp})} \\ &\text{subject to} \quad \mathrm{SINR}_{g,q}^{(\mathrm{dp})} \ge \lambda^{(\mathrm{dp})}, \forall g, q, \\ &\qquad\qquad\quad 0 \le p_{g,q} \le p_{\max}, \forall n, q. \end{aligned} \tag{3.34}$$

The objective is to maximize $\lambda^{(\mathrm{dp})}$, which is the lower bound of all $\mathrm{SINR}_{g,q}^{(\mathrm{dp})}$ as expressed in the first constraint, whereas the transmit power is limited as in the second constraint.

Similar to the training phase, the max-min QoS optimization problem in the data phase is also a GP and hence can be solved in polynomial time.

Max-Min QoS Power Allocation using the Bisection Method: With the optimization problems formulated above for training and data phases, a joint adaptive max-min QoS power allocation using the bisection method can be performed as follows. In the first stage, the max-min QoS problem in the training phase (3.33) is solved to obtain the maximum value of the achievable QoS (say $R_{\text{ini}}^{(\text{tp})}$) and the corresponding SE $R_{\text{ini}}^{(\text{dp})}$. Intuitively, a higher rate in the training phase causes a lower rate in the data phase because of lower-quality channel estimation. Hence, to find the value of $R_{g,q}^{(\text{tp})}$ that maximizes the total rate $R_{g,q}^{(\text{total})}$, its lower and upper bounds $R_{\min} \leq R_{g,q}^{(\text{tp})} \leq R_{\max}$ are chosen such that $R_{\max} = R_{\text{ini}}^{(\text{tp})}$ is the optimal solution for (3.33) and $R_{\min} = 0$. Applying bisection searching within this interval, in each iteration, the following problem is solved

$$\begin{aligned}
&\underset{\rho_{g,q}^{(\text{p})},\rho_{g,q}^{(\text{d})}}{\text{minimize}} \quad \theta \\
&\text{subject to} \quad \frac{\rho_{g,k}^{(\text{p})}}{\rho_{g',q}^{(\text{d})}} \leq \theta, \forall k, q, g \neq g^{'}, \\
&\qquad\qquad\quad \text{SINR}_{g,q}^{(\text{tp})} \geq \lambda_{\text{req}}^{(\text{tp})}, \forall g, q, \\
&\qquad\qquad\quad 0 \leq \rho_{g,q}^{(\text{p})} \leq p_{\max}, \forall g, q, \\
&\qquad\qquad\quad 0 \leq \rho_{g,q}^{(\text{d})} \leq p_{\max}, \forall g, q,
\end{aligned} \tag{3.35}$$

where $\lambda_{\text{req}}^{(\text{tp})}$ is the value of the SINR corresponding to

$$R_{\text{req}}^{(\text{tp})} = \log_2(1 + \lambda_{\text{req}}^{(\text{tp})}). \tag{3.36}$$

The cost function and the first constraint in (3.35) aim to minimize the interference caused by pilot-transmitting group as in (3.21), while maintaining a required QoS as expressed in the second constraint. After obtaining the power allocation with respect to (3.35), the total achievable UL rate can be calculated by (3.30). This procedure is iterated until $R_{g,q}^{(\text{total})}$ converges. The proposed power allocation method is summarized in Algorithm 1. With the proposed power control algorithm, the max-min QoS with time-offset pilots is not upper bounded by a saturation level as in the case of using non-orthogonal pilots. The reason

Algorithm 1 Bisection-based algorithm for max-min QoS power control

```
Require: The maximum achievable rate in training phase $R_{\text{ini}}^{(\text{tp})}$ and its corresponding SE $R_{\text{ini}}^{(\text{dp})}$
  $\delta = \infty$;
  $R_{\max} = R_{\text{ini}}^{(\text{tp})}$
  $R_{\min} = 0$
  $R_{\text{prev}}^{(\text{total})} = R_{\text{ini}}^{(\text{total})}$ where $R_{\text{ini}}^{(\text{total})}$ is calculated as in (3.30)
  while $\delta > \delta_{\text{threshold}}$ do
    $R_{\text{req}}^{(\text{tp})} = (R_{\min} + R_{\max})/2$
    Solve (3.35) with respect to $\lambda_{\text{req}} = R_{\text{req}}^{(\text{tp})}$
    Recalculate the corresponding $R_{\text{req}}^{(\text{tp})}$ and obtain the new $R_{\text{new}}^{(\text{total})}$ by applying (3.30).
    if $R_{\text{new}}^{(\text{total})} \leq R_{\text{prev}}^{(\text{total})}$ then
      $R_{\max} = R_{\text{req}}^{(\text{tp})}$
    else
      $R_{\min} = R_{\text{req}}^{(\text{tp})}$
      $R_{\text{prev}}^{(\text{total})} = R_{\text{new}}^{(\text{total})}$
    end if
    $\delta = R_{\max} - R_{\min}$;
  end while
  return $R_{\text{new}}^{(\text{total})}$
```

is that the SINR in the training phase grows proportionally with the number of antennas. When the coherence interval is short, this even leads to a larger amount of SE compared to the orthogonal pilot method. This is because the orthogonal pilot method has to spend more time slots for pilot signaling and there will be fewer time slots left for data transmission. On the other hand, when the coherence interval is large, the proposed algorithm can adaptively reduce data power in the training phase to ease the effect of pilot contamination to the data phase. It should also be pointed out that when the data power in the training phase is set to 0, there is no pilot contamination, and the system with time-offset pilots is equivalent to the system using orthogonal pilots with the pilot length of $2\tau_p$.

3.4.2 Minimization of Total Power

This section studies the power control problem in which the objective is to minimize the total transmit power of the system while guaranteeing a predetermined QoS to be equally provided to all users. Because one coherence interval is separated into two phases, the average power is:

$$P_{g,k}^{(\text{total})} = \frac{\tau_p(\rho_{g,k}^{(\text{p})} + \rho_{g,k}^{(\text{d})}) + (\tau_c - 2\tau_p)p_{g,k}}{\tau_c}. \tag{3.37}$$

For a required QoS value of ξ that is equally provided to all users, the optimization problem is:

$$\begin{aligned} & \underset{\rho_{g,k}^{(\text{p})}, \rho_{g,k}^{(\text{d})}, p_{g,k}}{\text{minimize}} && \sum_{g=1}^{2}\sum_{k=1}^{K} P_{g,k}^{(\text{total})} \\ & \text{subject to} && R_{g,k}^{(\text{total})} \geq \xi, \forall g,k, \\ & && 0 \leq \rho_{g,k}^{(\text{p})} \leq p_{\max}, \forall g,k, \\ & && 0 \leq \rho_{g,k}^{(\text{d})} \leq p_{\max}, \forall g,k, \end{aligned} \tag{3.38}$$

where the first constraint is to ensure that the required QoS value of ξ is equally served to all users, whereas the next two constraints limit the transmit power by a maximum value of $p_{\max}$. The left-hand-side of the SE constraint is in the form of a fraction whose denominator and nominator are posynomials, while the right-hand-side is a constant. This means that the above optimization problem is a signomial programming, which is NP-hard [39]. Hence, like in the previous section, the original problem in (3.38) are separated into two subproblems for the training phase and data phase. By iteratively solving this two subproblems until convergence, a suboptimal solution for (3.38) is obtained. The two subproblems are formulated and discussed next.

Power Control in Training Phase: The power minimization problem in the training phase can be written as:

$$\begin{aligned} \underset{\rho^{(\mathrm{p})}_{g,k},\rho^{(\mathrm{d})}_{g,k}}{\text{minimize}} \quad & \sum_{g=1}^{2}\sum_{k=1}^{K}(\rho^{(\mathrm{p})}_{g,k}+\rho^{(\mathrm{d})}_{g,k}) \\ \text{subject to} \quad & \mathrm{SINR}^{(\mathrm{tp})}_{g,k} \geq \lambda^{(\mathrm{tp})}_{\mathrm{req}}, \forall g,k, \\ & 0 \leq \rho^{(\mathrm{p})}_{g,k} \leq p_{\max}, \forall g,k, \\ & 0 \leq \rho^{(\mathrm{d})}_{g,k} \leq p_{\max}, \forall g,k, \end{aligned} \tag{3.39}$$

where $\lambda^{(\mathrm{tp})}_{\mathrm{req}}$ is the required SINR, which is equivalent to a predetermined value of QoS as defined in (3.36). This optimization problem is a GP, and hence can be solved in polynomial time.

Power Control in Data Phase: In the data phase, the power minimization problem is:

$$\begin{aligned} \underset{p_{g,k}}{\text{minimize}} \quad & \sum_{g=1}^{G}\sum_{k=1}^{K} p_{g,k} \\ \text{subject to} \quad & \mathrm{SINR}^{(\mathrm{dp})}_{g,k} \geq \lambda^{(\mathrm{dp})}_{\mathrm{req}}, \forall g,k, \\ & 0 \leq p_{g,k} \leq p_{\max}, \forall g,k, \end{aligned} \tag{3.40}$$

where $\lambda^{(\mathrm{dp})}_{\mathrm{req}}$ is the required SINR, which is equivalent to a predetermined value of QoS $R^{(\mathrm{dp})}_{\mathrm{req}}$:

$$R^{(\mathrm{dp})}_{\mathrm{req}} = \log_2(1+\lambda^{(\mathrm{dp})}_{\mathrm{req}}). \tag{3.41}$$

The above power minimization is convex, hence it is easily solved by existing convex optimization packages such as CVX.

Joint Power Minimization: To minimize the total transmit power during a coherence interval which includes both the training and data phases as in (3.38) is a problem with high complexity. Hence, an iterative method based on the bisection algorithm is performed as follows.

For a QoS requirement of ξ, in the first stage, we solve the max-min QoS problem in the training phase to obtain the maximum value of the achievable QoS in this phase, say $R_{\max}^{(\mathrm{tp})}$. In the next step, we find the optimal UL rate contributed by the data phase, $R_{\mathrm{req}}^{(\mathrm{tp})}$, which minimizes the total UL transmit power. This can be done by bounding $R_{\min} \leq R_{\mathrm{req}}^{(\mathrm{tp})} \leq R_{\max}$ where the upper-bound and lower-bound are initially chosen as $R_{\min} = 0$ and $R_{\max} = R_{\max}^{(\mathrm{tp})}$ and then updated in each iteration until the two bounds converge. With the allocated UL rate in the training phase, $R_{\mathrm{req}}^{(\mathrm{tp})}$, the required QoS in the data phase is:

$$R_{\mathrm{req}}^{(\mathrm{dp})} = \frac{\tau_c \xi - \tau_p R_{\mathrm{req}}^{(\mathrm{tp})}}{\tau_c - 2\tau_p}. \tag{3.42}$$

By using the power profile obtained in the training phase to estimate the channel coefficients, we can solve (3.40) with respect to the required data rate as in (3.42) and acquire the optimal transmit power in the data phase and the total UL transmit power. In the next iteration, the required data rate in the training phase is reduced to lower the effect of pilot contamination, which enhances the data rate in the data phase. The new total UL transmit power is then calculated by solving (3.40) with respect to the new required QoS. If the total UL transmit power in the new iteration is higher than the previous one, it means that the allocated QoS in the training phase has been reduced to much, which causes excessive power in the data phase. In this case, $R_{\min}$ needs to be updated to raise the allocated QoS in the training phase and ease the burden in the data phase. Otherwise, if the total power in the new iteration is lower than the previous one, we can continuously reduce the allocated power in the training phase by updating $R_{\max}$. The iteration process stops when two bounds converges (when $\delta = R_{\max} - R_{\min}$ is lower than a threshold value $\delta_{\mathrm{threshold}}$). The proposed procedure is summarized in Algorithm 2.

3.5 Simulation Results

In this section, numerical results are given to evaluate the performance of the multiuser massive MIMO system with time-offset pilots in terms of achievable QoS and power consumption. The results are also compared to results obtained with orthogonal pilots and non-orthogonal pilots. The performance is observed by changing the number of antennas, coherence interval and the required QoS. The massive MIMO system considered in simula-

Algorithm 2 Bisection algorithm for power minimization

```
Require: The maximum achievable rate in training phase $R_{\max}^{(\mathrm{tp})}$ and the required QoS $\xi$
  $\delta = \infty$;
  $P_{\mathrm{prev}}^{(\mathrm{total})} = \infty$;
  $R_{\max} = R_{\max}^{(\mathrm{tp})}$;
  $R_{\min} = 0$;
  $R_{\mathrm{req}}^{(\mathrm{tp})} = R_{\max}$;
  while $\delta > \delta_{\mathrm{threshold}}$ do
    Solve (3.39) with respect to $\lambda_{\mathrm{req}}^{(\mathrm{tp})}$ calculated in (3.36).
    Calculate the required $R_{\mathrm{req}}^{(\mathrm{dp})}$ as in (3.42) and solve (3.40).
    $P_{\mathrm{new}}^{(\mathrm{total})} = \sum_{g=1}^{G} \sum_{k=1}^{K} P_{g,k}^{(\mathrm{total})}$;
    if $P_{\mathrm{new}}^{(\mathrm{total})} \leq P_{\mathrm{prev}}^{(\mathrm{total})}$ then
      $R_{\max} = R_{\mathrm{req}}^{(\mathrm{tp})}$;
      $P_{\mathrm{prev}}^{(\mathrm{total})} = P_{\mathrm{new}}^{(\mathrm{total})}$;
    else
      $R_{\min} = R_{\mathrm{req}}^{(\mathrm{tp})}$;
    end if
    $R_{\mathrm{req}}^{(\mathrm{tp})} = (R_{\min} + R_{\max})/2$;
    $\delta = R_{\max} - R_{\min}$;
  end while
  return $P_{\mathrm{new}}^{(\mathrm{total})}$
```

Table 3.1 Simulation parameters.

Parameter	Value
Peak UL radio transmit power	23 dBm
Number of users	30
Shadowing standard deviation	10 dB
Penetration loss (indoor users)	20 dB
Noise figure	5 dB
Pathloss	$131 + 42.8\text{log}_{10}d$

tion consists of one multi-antenna BS and 30 randomly-distributed users. In each iteration, the locations of 30 users are randomly generated within the 200 meters radius around the BS. Numerical results are averaged over 200 iterations. The large-scale fading coefficients are modeled according to the 3GPP LTE standard [41]. Specifically, the large scale fading is computed as $\beta_{g,k} = -131 - 42.8\text{log}_{10}d_{g,k} + z_{l,k}$ (dB), where $d_{g,k}$ denotes the distance from the BS to the kth user of the gth group and $z_{g,k}$ is the standard deviation of the shadowing variable. The noise figure of 5dB translates to a noise variance of -96dBm. The simulation parameters are summarized in Table 3.1. In all simulation scenarios, the number of pilots for time-offset and non-orthogonal pilot methods is τ_p, whereas, in order to serve the same number of users, the orthogonal pilot method needs twice the number of pilots, i.e., $2\tau_p$.

Fig. 3.2 plots the maximum QoS that all users can be equally served by the BS. It can be seen that using time-offset pilots yields a far better performance compared to using non-orthogonal pilots. Moreover the performance gap between this two methods increases with the number of antennas, from about 0.5 bits/sec/Hz at $M = 100$ to almost 1 bit/sec/Hz at $M = 500$ for the case of $\tau_c = 100$ symbols. The reason is that, when M increases, the denominator in the SINR expression increases proportionally with M for non-orthogonal pilots [19, 24], while it is not the case with time-offset pilots, thanks to the power control algorithm represented in Section IV-B. Another remarkable observation is that the performance curves with time-offset pilots are just slightly below that with orthogonal pilots for the case $\tau_c = 100$, while the performance curves with time-offset pilots are better when $\tau_c = 50$.

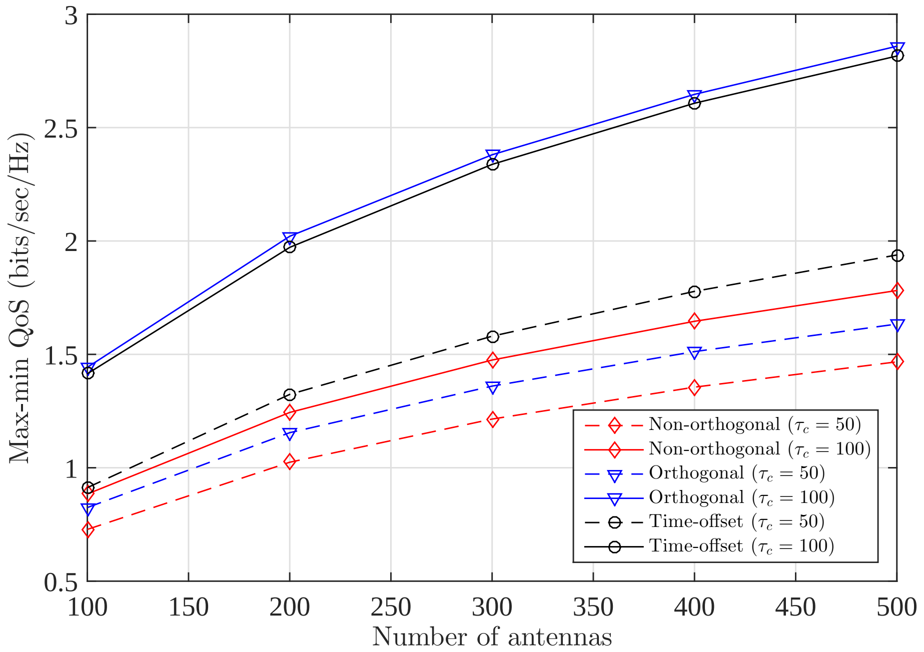


Figure 3.2 Max-min QoS versus the number of antennas ($N = 30$ users).

This is because when the coherence interval is short, the orthogonal pilot method has to spend a larger portion of the coherence interval for channel estimation, while the time-offset pilot method has a much longer duration for data transmission.

The achievable rates that the BS can equally serve all users for different coherence intervals are illustrated in Fig. 3.3. Obviously, when $\tau_c \to K \times G$, there are no time slots available for data transmission in the orthogonal pilot case and the data rate goes to zero. On the other hand, non-orthogonal and time-offset pilot methods can still provide SEs of up to 1 and 1.5 bits/sec/Hz, respectively (when $M = 400$). When τ_c increases, the SE achieved with the non-orthogonal pilot method tends to asymptotically approach 1.7 bits/sec/Hz for 400 antennas and 1.3 bits/sec/Hz for 200 antennas due to pilot contamination. In contrast, the SEs achieved with time-offset and orthogonal pilot methods sharply increase with τ_c and reach up to 2.6 bit/sec/Hz when $\tau_c = 120$. It can also be seen that when the coherence interval is shorter than about 70 symbols, using time-offset pilots yields a better performance than using orthogonal pilots. The intersection value increases when the number of antennas goes up (at $\tau_c = 66$ for 200 antennas and $\tau_c = 70$ for 400 antennas).

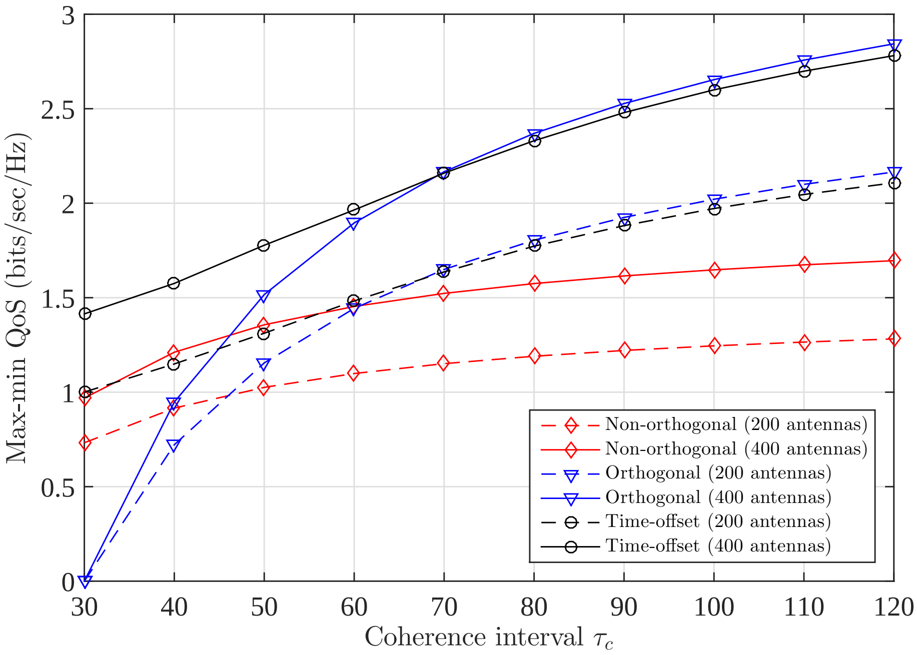


Figure 3.3 Max-min QoS versus coherence interval ($N = 30$ users).

The max-min QoS values versus the number of users for different coherence lengths are shown in Fig 3.4. The number of pilots is set as half of the number of users for time-offset and non-orthogonal pilots. The max-min QoS value decreases when the number of users increases because there are more interference sources. However, the time-offset pilot method still outperforms the non-orthogonal pilot method. This is because interference cancellation can be applied for data detection and better UL channel estimation can be obtained with time-offset pilots compared to non-orthogonal pilots. Remarkably, when $\tau_c = 60$ the time-offset pilot method eventually shows a better performance compared to orthogonal pilots when $N \geq 30$ users. Again, the reason is that with orthogonal pilots, the system has to spend a larger portion of time slots of pilots, which leaves a smaller number of time slots for data transmission. Furthermore, the contribution from the training phase to the total UL SE is presented in Table 3.2 for the case $N = 30$. When the coherence interval $\tau_c = 30$, it is obvious that the training phase contributes 100% of the total UL SE. The contribution in total uplink SE of the training phase decreases when τ_c increases. When $\tau_c = 65$ symbols, the SE contribution from the training phase approximately approaches zero, which means that no data is transmitted in the training phase. In such a case, the system is equivalent

Table 3.2 Rate contribution from the training phase.

Coherence length (τ_c)	30	35	40	45	50	55	60	65	70	75
Percentage (%)	100	79.66	63.74	40.62	32.71	18.24	6.60	1.29	1.13	1.00

to the one that uses orthogonal pilots with the length of $2\tau_p$.

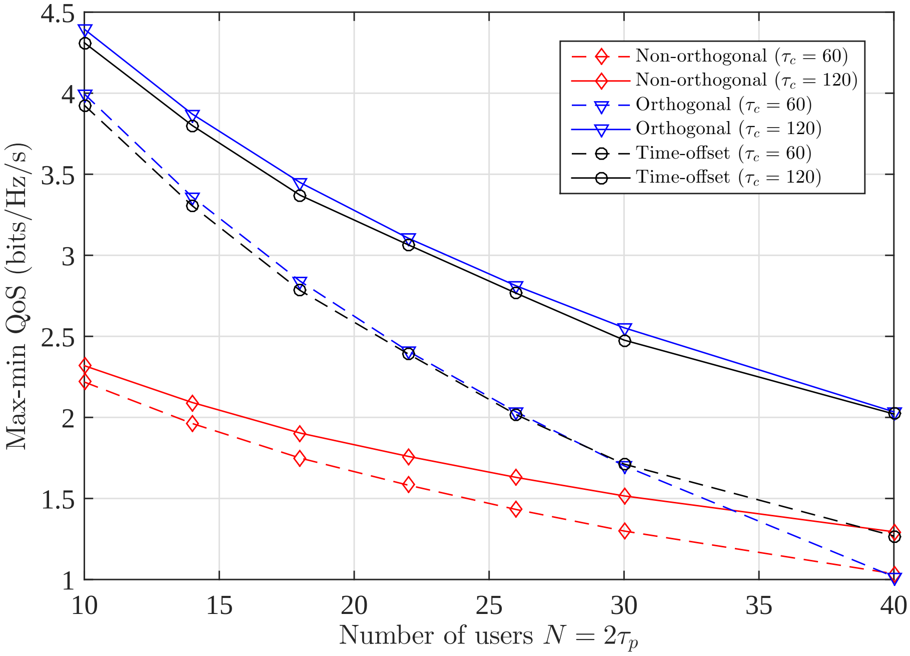


Figure 3.4 Max-min QoS versus the number of users ($M = 300$ antennas).

Fig. 3.5 plots the cumulative distribution function (CDF) of the max-min QoS of all three pilot methods, where the number of antennas is 500 for the non-orthogonal pilot method, and 300 for the other two methods. For the time-offset pilot method, proposed and random group assignments are considered. As can be seen, by performing group assignment according to large-scale fading information as described at the end of Section III, the time-offset pilot method can provide a max-min QoS of more than 4.2 bits/Hz/s, which is far better than when group assignment is performed randomly. In addition, the figure also shows that the time-offset pilot method outperforms the non-orthogonal pilot method when employing the same number of pilots.

Fig. 3.6 illustrates the optimal values of per-user average transmit power at the required QoS of 1.5 bits/sec/Hz with $\tau_c = 80$ and 120 symbols. As can be seen, the non-orthogonal pilot method enjoys a significant transmit power reduction when the number of antennas increases. However, the power minimization problem is not always feasible with non-orthogonal pilots. Specifically, when $\tau_c = 80$, the problem is feasible for the number of antennas larger than 400, whereas for $\tau_c = 120$ the minimum number of antennas required to have a feasible problem is 300. Similarly, the power consumption in the case of orthogonal pilots decreases when the number of antennas increases, but much slower. The same trend can be observed for time-offset pilots, where the power consumption drops noticeably when the number of antennas increases from 200 to 300.

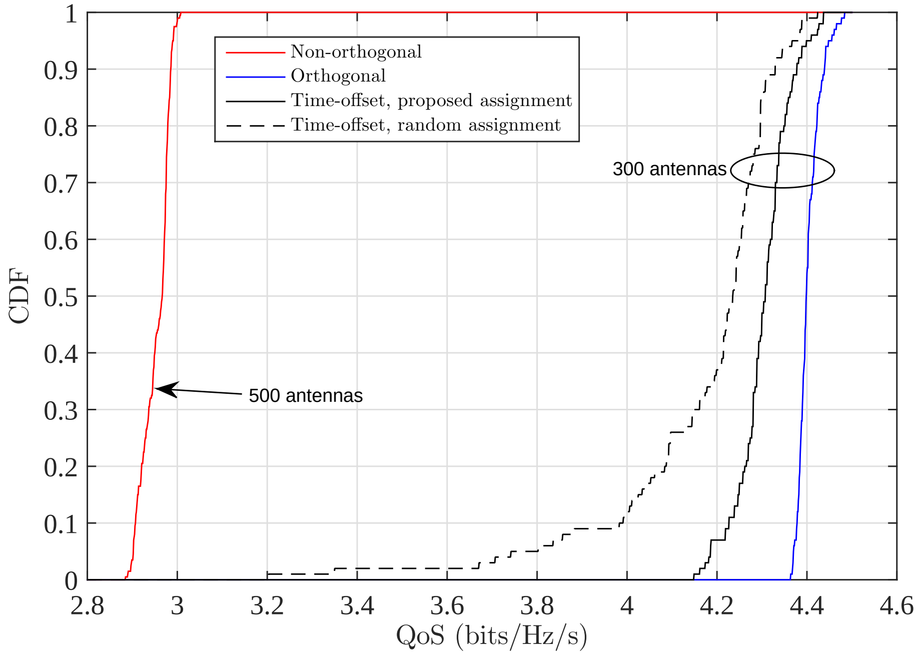


Figure 3.5 Cumulative distribution function of the max-min QoS ($N = 10$ users, $\tau_c = 120$).

The impact of coherence interval on the optimal power allocation is illustrated in Fig. 3.7. With the required QoS of 1.5 bits/sec/Hz and the number of antennas is 200 or 400, the power consumptions of all three pilot methods reduce when the coherence interval increases. Specifically, the power consumption of the time-offset pilot method decreases by 10 mW when the coherence interval increases from 60 to 120 symbols in both cases. With non-orthogonal pilots, the power minimization problem is infeasible with $M = 300$ antennas. When $M = 400$, this problem is solvable only when $\tau_c \geq 70$ symbols. In contrast, the transmit power reduction of the orthogonal pilot method is not very significant. This reduction in power consumption can be explained as a result of the lower required SINR when there are more time slots for data transmission.

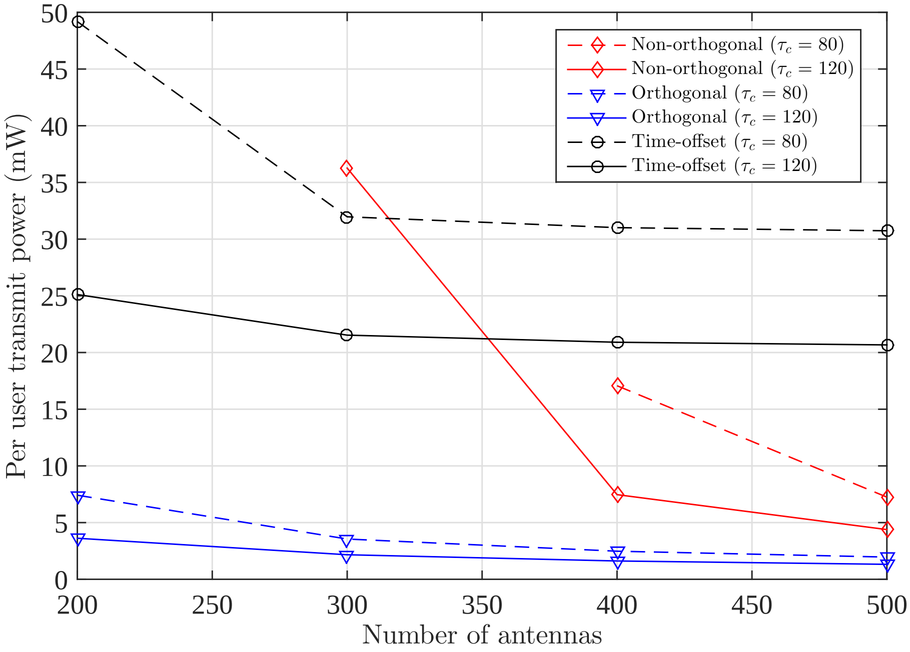


Figure 3.6 Per-user average transmit power versus the number of antennas (QoS=1.5 bit/Hz/s, $N = 30$ users).

Fig. 3.8 compares the change in per-user transmit power of the three pilot methods with respect to different required QoS levels when the BS has 300 antennas and the coherence interval is set at 60 and 120 symbols. Obviously, the transmit power increases when the required QoS increases. It can be seen that the slope of the curve under the non-orthogonal pilot method is much sharper than that of the two other methods. Noticeably, the curve with the time-offset pilot method only increases slightly when the required QoS increases from 0.5 to 1.5 bits/sec/Hz. The same tendency can also be observed in the case of the orthogonal pilot method but the change is larger.

Finally, Fig. 3.9 compare the sum SE of all users in the system between the orthogonal and proposed time-offset pilot methods. Although the per-user UL SE is lower with the time-offset pilot method than the orthogonal pilot method, with a fixed number of pilots

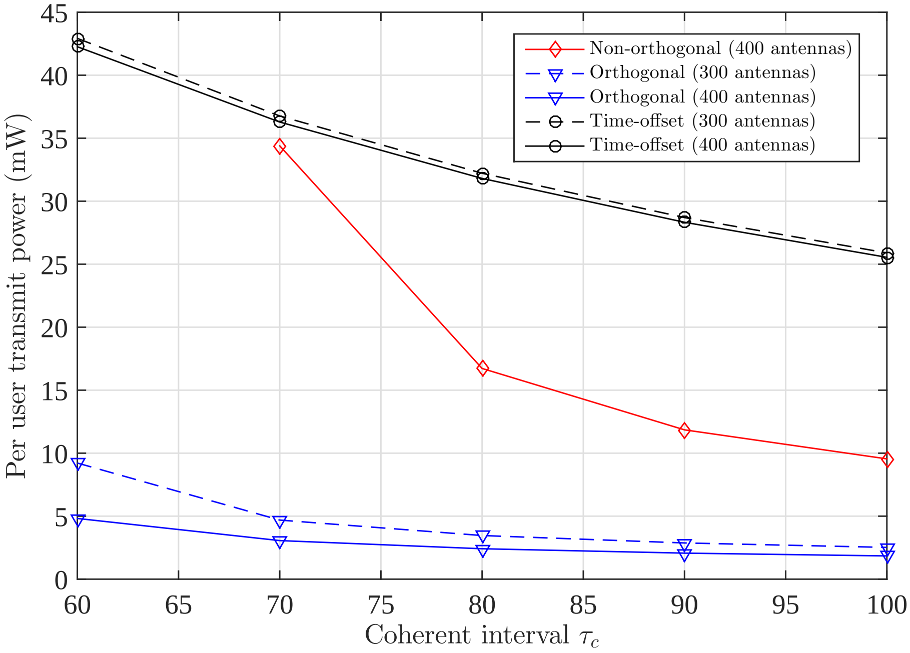


Figure 3.7 Per-user average transmit power versus coherence interval (QoS=1.5 bit/Hz/s, $N = 30$ users).

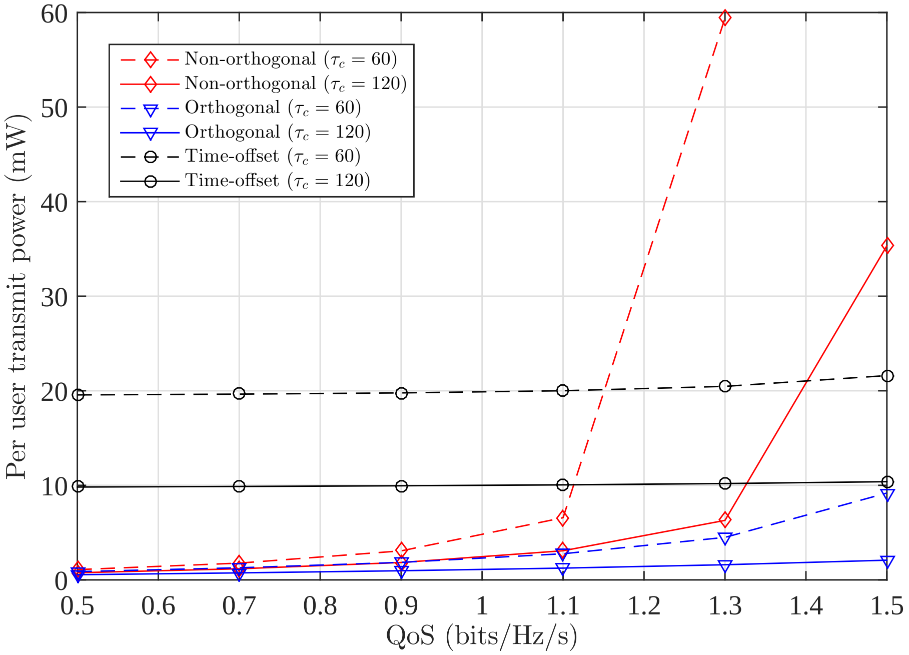


Figure 3.8 Per-user average transmit power versus required QoS level ($M = 300$ antennas, $N = 30$ users).

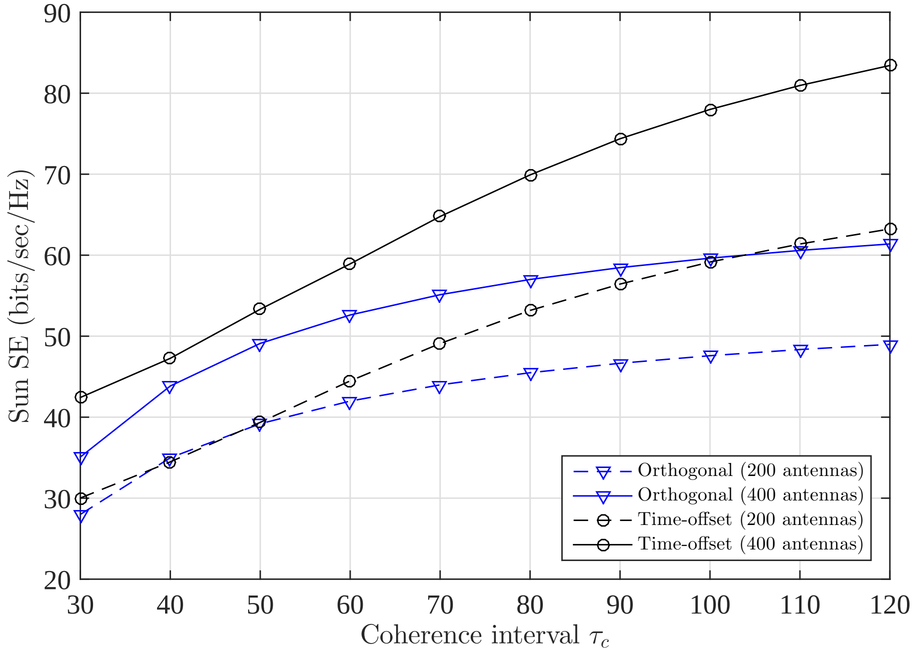


Figure 3.9 Comparison of the sum SE: $N = 15$ users with orthogonal pilots and $N = 30$ users with time-offset pilots.

sequences (here $\tau_p = 15$), the time-offset pilot method can serve twice the number of users (30 users) as compared to the conventional orthogonal pilot method (15 users). As a result, the sum SE is significantly larger with the time-offset pilot method than the orthogonal pilot method.

3.6 Conclusions

This work investigated performance of time-offset pilots in the UL of a single-cell multiuser massive MIMO system. It is shown that the correlated interference, a consequence of the correlation between pilots of one group and UL data of the other group, can be effectively removed by applying successive interference cancellation. We further formulate power control problems for two different cost functions: max-min QoS and total power minimization. Due to the signomial constraints, these two problems are NP-hard and hence very computationally demanding. Therefore, we proposed algorithms to find the suboptimal solutions based on the bisection method, which solve a series of GPs. Numerical results have shown that the time-offset pilot method provides a far better performance than the non-orthogonal pilot method. The time-offset pilot is also better than the orthogonal pilot method when the coherence interval is short.

3.A Appendix I

From the received signal in (3.26), a lower bound on the UL ergodic SE of the qth user in the first group can be obtained based on the definition of the mutual information between the original base-band signal $x_{1,q}^{(\text{tp})}$ and the received signal (after multiplied with the corresponding combining vector) $s_{1,q}$:

$$R_{1,q}^{(\text{tp})} \geq I\left(x_{1,q}^{(\text{tp})}; s_{1,q}, \hat{\mathcal{H}}\right), \tag{3.43}$$

where $\hat{\mathcal{H}}$ denote the knowledge of channel estimation at the BS. Under the input distribution $x_{1,q} \sim \mathcal{CN}(0,1)$, the mutual information can be equivalently expressed as

$$I\left(x_{1,q}^{(\text{tp})}; s_{1,q}, \hat{\mathcal{H}}\right) = h(x_{1,q}^{(\text{tp})}) - h\left(x_{1,q}^{(\text{tp})}|s_{1,q}, \hat{\mathcal{H}}\right) = \log_2(\pi e) - h\left(x_{1,q}^{(\text{tp})}|s_{1,q}, \hat{\mathcal{H}}\right), \tag{3.44}$$

where $h(x_{1,q}^{(\text{tp})})$ is the differential entropy and $h\left(x_{1,q}^{(\text{tp})}|s_{1,q}, \hat{\mathcal{H}}\right)$ is the conditional entropy. Because of the fact that the entropy does not change when subtracting a known variable, $h\left(x_{1,q}^{(\text{tp})}|s_{1,q}, \hat{\mathcal{H}}\right)$ can be bounded from above as

$$\begin{aligned} h\left(x_{1,q}^{(\text{tp})}|s_{1,q}, \hat{\mathcal{H}}\right) &= h\left(x_{1,q}^{(\text{tp})} - \alpha s_{1,q}|s_{1,q}, \hat{\mathcal{H}}\right) \\ &\leq h(x_{1,q}^{(\text{tp})} - \alpha s_{1,q}) \\ &\leq \log_2\left(\pi e \mathbb{E}\left\{|x_{1,q}^{(\text{tp})} - \alpha s_{1,q}|^2\right\}\right), \end{aligned} \tag{3.45}$$

where α is a deterministic scalar. The best upper bound for $h\left(x_{1,q}^{(\text{tp})}|s_{1,q}, \hat{\mathcal{H}}\right)$ can be found by minimizing the expectation in (3.45) with respect to α. Since the UL data signals of users in the system are mutually independent, calculating the statistical average $\mathbb{E}\left\{|x_{1,q}^{(\text{tp})} - \alpha s_{1,q}|^2\right\}$ over $x_{g,k}^{(\text{tp})}$ $(g = 1, 2)$ leads to a quadratic function of α:

$$\begin{aligned} \mathbb{E}\left\{|x_{1,q}^{(\text{tp})} - \alpha s_{1,q}|^2\right\} = 1 - 2\alpha \mathbb{E}\left\{\sum_{l=1}^{L} v_{1,q}^H h_{1,q}\right\}\sqrt{\rho_{1,q}^{(\text{d})}} \\ + \alpha^2 \Bigg[\sum_{k=1}^{K}\left(\rho_{1,k}^{(\text{d})}\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k}\right|^2\right\} + \rho_{2,k}^{(\text{p})}\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k}\right|^2\right\}\right) \\ - \mathbb{E}\left\{\left|\hat{\Upsilon}_{1,q}^{(\text{IP})}\right|^2\right\} + \sigma^2 \mathbb{E}\left\{\|\boldsymbol{v}_{1,q}\|^2\right\}\Bigg]. \end{aligned} \tag{3.46}$$

The minimum value of this quadratic function can be easily obtained as:

$$\mathbb{E}\left\{|x_{1,q}^{(\text{tp})} - \alpha s_{1,q}|^2\right\} \geq \frac{1}{1 + \text{SINR}_{1,q}^{(\text{tp})}}. \tag{3.47}$$

$$\mathrm{SINR}_{1,q}^{(\mathrm{tp})} = \frac{\rho_{1,q}^{(\mathrm{d})}\left|\mathbb{E}\left\{\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\right\}\right|^2}{\sum_{k=1}^K\left(\rho_{1,k}^{(\mathrm{d})}\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,k}\right|^2\right\}+\rho_{2,k}^{(\mathrm{p})}\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\right|^2\right\}\right)-\mathbb{E}\left\{\left|\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right|^2\right\}-\rho_{1,q}^{(\mathrm{d})}\left|\mathbb{E}\left\{\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\right\}\right|^2+\sigma^2\mathbb{E}\left\{\|\boldsymbol{v}_{1,q}\|^2\right\}}. \tag{3.48}$$

where $\mathrm{SINR}_{1,q}^{(\mathrm{tp})}$ is defined in (3.48). By choosing this value, the tightest lower bound of $I\left(x_{1,q}^{(\mathrm{tp})}; s_{1,q}, \hat{\mathcal{H}}\right)$ is obtained. Finally, plugging the result from (3.44) to (3.47) into (3.43) we obtain the lower bound for UL SE that the qth user of the first group can achieve as in *Theorem 1.*

With the MRC, the combining vector for the qth user of the first group is $\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q}$, and we can calculate the closed-form SINR expression as follows.

The expected squared norm of the Rayleigh-distributed channel between the BS and the qth user is

$$\mathbb{E}\left\{\|\boldsymbol{v}_{1,q}\|^2\right\} = \mathbb{E}\left\{\|\hat{\boldsymbol{h}}_{1,q}\|^2\right\} = \gamma_{1,q}M. \tag{3.49}$$

and

$$\mathbb{E}\left\{\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\right\} = \mathbb{E}\left\{\hat{\boldsymbol{h}}_{1,q}^H(\hat{\boldsymbol{h}}_{1,q}+\boldsymbol{e}_{1,q})\right\} = \mathbb{E}\left\{\|\hat{\boldsymbol{h}}_{1,q}\|^2\right\} = \gamma_{1,q}M, \tag{3.50}$$

The expectation $\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\right|^2\right\}$ is:

$$\begin{aligned}\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\right|^2\right\} &= \gamma_{1,q}\beta_{2,k}M + \frac{\rho_{2,k}^{(\mathrm{d})}\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q}^2}{\left(\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q}+\sum_{k=1}^K\rho_{2,k}^{(\mathrm{d})}\beta_{2,k}+\sigma^2\right)^2}\mathbb{E}\left\{\|\boldsymbol{h}_{2,k}\|^4\right\}\\ &= \gamma_{1,q}\beta_{2,k}M + \frac{\rho_{2,k}^{(\mathrm{d})}\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q}^2}{\left(\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q}+\sum_{k=1}^K\rho_{2,k}^{(\mathrm{d})}\beta_{2,k}+\sigma^2\right)^2}\beta_{2,k}^2\frac{\Gamma(M+2)}{\Gamma(M)}\\ &= \gamma_{1,q}\beta_{2,k}M + \frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}}\tau_p\gamma_{1,q}^2\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^2M(M+1).\end{aligned} \tag{3.51}$$

Consider the interference within the first group, when $k = q$, one has:

$$\begin{aligned}\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\right|^2\right\} &= \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\left(\hat{\boldsymbol{h}}_{1,q}+\boldsymbol{e}_{1,q}\right)\right|^2\right\} = \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\hat{\boldsymbol{h}}_{1,q}\right|^2\right\} + \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{e}_{1,q}\right|^2\right\}\\ &= \gamma_{1,q}^2(M+M^2) + \gamma_{g',q}\left(\beta_{1,q}-\gamma_{1,q}\right)M\\ &= \left(\gamma_{1,q}M\right)^2 + \beta_{1,q}\gamma_{1,q}M.\end{aligned} \tag{3.52}$$

In the case when $k \neq q$, one has:

$$\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,k}\right|^{2}\right\} = \gamma_{1,q}\beta_{1,k}M. \tag{3.53}$$

With $\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}$ being defined as in (3.17), the reduced amount of interference is:

$$\mathbb{E}\left\{\left|\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right|^{2}\right\} = \sum_{k=1}^{K}\rho_{2,k}^{(\mathrm{p})}\frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}\tau_{p}}\gamma_{1,q}^{2}\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^{2}M^{2}. \tag{3.54}$$

Substituting (3.49) to (3.54) into (3.48), one obtains the SINR as in (3.21).

References

[1] F. Fernandes, A. Ashikhmin, and T. L. Marzetta, “Inter-Cell Interference in Noncooperative TDD Large Scale Antenna Systems,” in *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 192–201, Feb. 2013.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up MIMO: Opportunities and Challenges with Very Large Arrays,” in *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[3] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, “Massive MIMO Performance Evaluation Based on Measured Propagation Data,” in *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3899–3911, 2015.

[4] E. Björnson, E. G. Larsson, and T. L. Marzetta, “Massive MIMO: ten myths and one critical question,” in *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[5] T. L. Marzetta, “Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas,” in *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[6] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-Free Massive MIMO Versus Small Cells,” in *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems,” in *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[8] T. V. Chien and E. Björnson, “Massive MIMO Communication”, in *5G Mobile Communications*, pp. 77–116. Cham, Switzerland: Springer International Publishing, 2017.

[9] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO.* Cambridge, U.K: Cambridge University Press, 2016.

[10] R. R. Müller, L. Cottatellucci, and M. Vehkaperä, "Blind Pilot Decontamination," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 773–786, Oct. 2014.

[11] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot Contamination and Precoding in Multi-Cell TDD Systems," in *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[12] H. Q. Ngo, T. L. Marzetta, and E. G. Larsson, "Analysis of the pilot contamination effect in very large multicell multiuser MIMO systems for physical channel models," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3464–3467, Prague 2011.

[13] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO-5G system," in *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 905–923, 2016.

[14] N. Krishnan, R. D. Yates, and N. B. Mandayam, "Uplink linear receivers for multi-cell multiuser MIMO with pilot contamination: Large system analysis," in *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4360–4373, Aug. 2014.

[15] J. Shen, J. Zhang, and K. B. Letaief, "Downlink user capacity of massive MIMO under pilot contamination," in *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3183–3193, June 2015.

[16] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.

[17] J. Li, D. Wang, P. Zhu, J. Wang, and X. You, “Downlink spectral efficiency of distributed massive MIMO systems with linear beamforming under pilot contamination,” in *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1130–1145, Feb. 2018.

[18] T. V. Chien, E. Björnson, and E. G. Larsson, “Joint pilot design and uplink power allocation in multi-cell massive MIMO systems,” in *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2000–2015, Mar. 2018.

[19] E. Björnson, E. G. Larsson, and M. Debbah, “Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?,” in *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[20] H. Al-Salihi, T. V. Chien, L. A. Tuan, and M. R. Nakhai, “A successive optimization approach to pilot design for multi-cell massive mimo systems,” in *IEEE Communications Letters*, vol. 22, no. 5, pp. 1086-1089, May 2018.

[21] W. A. W. M. Mahyiddin, P. A. Martin, and P. J. Smith, “Performance of synchronized and unsynchronized pilots in finite massive MIMO systems,” *IEEE Transactions on Wireless Communications*, vol. 14, pp. 6763–6776, Dec. 2015.

[22] I. Atzeni, J. Arnau, and M. Debbah, “Fractional pilot reuse in massive MIMO systems,” *2015 IEEE International Conference on Communication Workshop (ICCW)*, London, 2015, pp. 1030-1035.

[23] X. Zhu, Z. Wang, L. Dai, and C. Qian, “Smart pilot assignment for massive MIMO,” in *IEEE Communications Letters*, vol. 19, no. 9, pp. 1644–1647, Sept. 2015.

[24] H. V. Cheng, E. Björnson, and E. G. Larsson, “Optimal pilot and payload power control in single-cell massive MIMO systems,” in *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2363–2378, May 2017.

[25] E. Björnson, E. G. Larsson, and M. Debbah, “Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?,” in *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, 2016.

[26] D. Hu, L. He, and X. Wang, “Semi-blind pilot decontamination for massive MIMO systems,” in *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 525–536, 2016.

[27] D. Kong, D. Qu, K. Luo, and T. Jiang, “Channel estimation under staggered frame structure for massive MIMO system,” in *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1469–1479, Feb. 2016.

[28] S. Jin, X. Wang, Z. Li, K. Wong, Y. Huang, and X. Tang, “On massive MIMO zero-forcing transceiver using time-shifted pilots,” in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 1, pp. 59–74, Jan. 2016.

[29] B. Sun, Y. Zhou, L. Tian, and J. Shi, “Successive interference cancellation based channel estimation for massive MIMO systems,” *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Singapore, 2017, pp. 1-6.

[30] Liang Wu, Zaichen Zhang, Jian Dang, and Huaping Liu, “Enhanced time-shifted pilot based channel estimation in massive MIMO systems with finite number of antennas,” *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, Paris, 2017, pp. 222-227.

[31] X. Xia, K. Xu, D. Zhang, Y. Xu, and Y. Wang, “Beam-domain full-duplex massive mimo: Realizing co-time co-frequency uplink and downlink transmission in the cellular system,” in *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 8845–8862, 2017.

[32] X. Xia, K. Xu, Y. Wang, and Y. Xu, “A 5G-enabling technology: Benefits, feasibility, and limitations of in-band full-duplex mmimo,” in *IEEE Vehicular Technology Magazine*, vol. 13, no. 3, pp. 81–90, 2018.

[33] K. Xu, Z. Shen, Y. Wang, X. Xia, and D. Zhang, “Hybrid time-switching and power splitting swipt for full-duplex massive mimo systems: A beam-domain approach,” in *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7257–7274, 2018.

[34] T. V. Chien, E. Björnson, and E. G. Larsson, “Joint power allocation and user association optimization for massive MIMO systems,” in *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6384–6399, Sept. 2016.

[35] S. Kay, *Fundamentals of statistical signal processing: Estimation theory.* Englewood Cliffs, NJ, USA: Prentice Hall, 1995.

[36] J. Li, E. Björnson, T. Svensson, T. Eriksson, and M. Debbah, “Joint precoding and load balancing optimization for energy-efficient heterogeneous networks,” in *IEEE Transactions on Wireless Communications*, vol. 14, no. 10, pp. 5810–5822, 2015.

[37] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, U.K: Cambridge University Press, 2005.

[38] H. V. Cheng, E. Björnson, and E. G. Larsson, “Uplink pilot and data power control for single cell massive MIMO systems with MRC,” in *2015 International Symposium on Wireless Communication Systems (ISWCS)*, Brussels, 2015, pp. 396-400.

[39] M. Chiang, C. W. Tan, D. P. Palomar, D. O’neill, and D. Julian, “Power control by geometric programming,” in *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[40] M. Grant, S. Boyd, and Y. Ye, “CVX: Matlab software for disciplined convex programming, 2009,” *Available online from www.stanford.edu/boyd/cvx*, 2015.

[41] *Further advancements for E-UTRA physical layer aspects (Release 9).* 3GPP TS 36.814, Mar. 2010.

4. Max-Min QoS Power Control in Generalized Cell-Free Massive MIMO-NOMA with Optimal Backhaul Combining

Published as:

The Khai Nguyen, Ha H. Nguyen, and Hoang Duong Tuan, “Max-Min QoS Power Control in Generalized Cell-Free Massive MIMO-NOMA with Optimal Backhaul Combining”, *IEEE Transactions on Vehicular Technology*, accepted, Jun. 2020 (available in IEEE Xplore as an Early Access Article).

In the previous chapter, the use of time-offset pilots is investigated and compared with synchronous pilots in a single-cell massive MIMO system. Despite the fact that employing successive interference cancellation can bring significant performance improvement, the correlated interference, which grows proportionally with the number of antennas, cannot be eliminated for all users. Motivated by this shortcoming, in this chapter, we extend our work to a cell-free massive MIMO system with the aid of NOMA in the form of nonorthogonal pilots. Considering the structure of a cell-free network, in which there is not cell boundary and a user can be simultaneous served by multiple base stations, we propose an optimal backhaul combining method to maximize the signal-to-interference-plus-noise ratios for users in the system. It is shown that the proposed scheme is capable of eliminate correlated interference for all users by effectively combining the received signals from multiple base stations at the backhaul central processing unit.

Max-Min QoS Power Control in Generalized Cell-Free Massive MIMO-NOMA with Optimal Backhaul Combining

The Khai Nguyen, Ha H. Nguyen, and Hoang Duong Tuan

Abstract

This paper studies the uplink (UL) transmission of a *generalized* cell-free massive multiple-input multiple-output (massive MIMO) system in which multiple base stations (or access points), each equipped with a multiple-antenna array and connected to a central processing unit (CPU) over a backhaul network, simultaneously serve multiple users in a cell-free service area. The paper focuses on the non-orthogonal multiple access (NOMA) approach for sharing pilot sequences among users. Unlike the conventional cell-free massive MIMO-NOMA systems in which the UL signals from different access points are equally combined over the backhaul network, this paper first develops an optimal backhaul combining (OBC) method to maximize the UL signal-to-interference-plus-noise ratio (SINR). It is shown that, by using OBC, the correlated interference can be effectively mitigated if the number of users assigned to each pilot sequence is less than or equal to the number of base stations (BSs). As a result, the cell-free massive MIMO-NOMA system with OBC can enjoy unlimited performance when the number of antennas at each BS tends to infinity. A closed-form SINR expression is derived under Rayleigh fading and used to formulate a max-min quality-of-service (QoS) power control problem to further enhance the system performance. To deal with the NP-hardness of the concerned optimization problem, a successive inner approximation technique is applied to convert the original problem into a series of convex optimizations, which can be solved iteratively. In addition, a user grouping algorithm is also developed and shown to be better than random user grouping and a grouping method recently proposed in the literature. Numerical results are presented to corroborate the analysis and demonstrate the superiority of the proposed optimal backhaul combining over both equal-gain backhaul combining and zero-forcing backhaul combining.

4.1 Introduction

Over the last decade, the demand for high-speed wireless communication services has grown tremendously. The next generations of communication systems, the fifth-generation (5G) and beyond, require 1000 times higher network capacity. This presents a huge challenge for the limited frequency resource. Among many potential enabling techniques, massive multiple-input multiple-output (MIMO) and non-orthogonal multiple access (NOMA) have emerged as key solutions to address the problem of limited spectrum [1–4].

NOMA exploits the power domain to enable users to effectively share the same system's resources (such as frequency, time slots and spreading codes) [2, 5–8]. By allocating different power levels to users and using superposition coding and successive interference cancellation (SIC), NOMA allows the network's resources to be efficiently used and hence increases the number of connections, as well as the network's sum spectral efficiency (SE) [2, 5–10].

On the other hand, by using hundreds antennas, a massive MIMO BS can serve multiple users in the same time-frequency resources with very high spectral efficiency, thanks to its high array gain and robustness against noise and interference [11–13]. Recently, a new setup of massive MIMO networks, called cell-free massive MIMO, has been shown to significantly enhance the network's SE as well as energy efficiency (EE) [14–18]. Cell-free networks imply that there is no cell classification, no cell boundaries and a user can be served by multiple BSs at a time and the signals received at BSs are gathered, combined and processed by a backhaul network [14]. In particular, the conventional cell-free massive MIMO setup has a massive number of single-antenna "base stations", which are more appropriately called access points (APs), that are geographically distributed over the service area. The cell-free setup is shown to bring up to 5 folds better performance as compared to the small-cell setup [14].

Given the distinctive benefits of massive MIMO and NOMA, the integration of these two techniques is expected to inherit important advantages of both techniques: high SE and massive connectivity [19, 20]. In [21], the authors propose a limited-feedback NOMA scheme. By decomposing a massive MIMO system into multiple single-input single-output NOMA channels, system design is significantly simplified. In [22], the user pairing problem for superposition coding in massive MIMO-NOMA is considered. By proper scheduling the detection order in each group and designing the interference cancellation matrix, an excellent sum rate can be achieved. The incorporation of massive MIMO, NOMA and interleaved division-multiple access (IDMA) is studied in [23]. This scheme is proved to be capable of offering high throughput and robustness against pilot contamination. The authors in [24] propose and analyze performance of a new convergent Gaussian message passing (GMP) multi-user detection method (called scale-and-add GMP) for a coded massive MIMO-NOMA system under the scenario that the number of users is larger than the number of BS antennas (i.e., an user overloaded scenario). The NOMA approach in pilot design for massive MIMO is investigated in [25]. In that paper, UL pilot and data transmission are scheduled in parallel, which allows the system to serve as many users as the length of a coherence interval. The authors in [26] exploit the channel's covariance matrix to group users into clusters. Thanks to the linear-independence property between clusters' covariance matrices [27], inter-cluster interference can be mitigated while data detection within each cluster can be improved with SIC.

All the aforementioned works examine single-cell and multi-cell systems. There are only a few studies of NOMA under the cell-free setup. Specifically, the application of NOMA in cell-free massive MIMO is considered in [28, 29] in terms of reusing pilots. By grouping users into clusters, in which users in the same cluster use the same pilot sequence, this scheme can serve significantly more users than the conventional orthogonal multiple-access (OMA) method. However, the trade-off is a decrease in the sum rate due to the intra-cluster interference. To maximize the achievable rate that can be equally served to all users, a NOMA/OMA mode selection for cell-free massive MIMO is proposed in [30]. This hybrid technique, when combined with SIC, can yield better performance as compared to each individual single-mode (NOMA or OMA) system.

For a cell-free massive MIMO system, there are generally two stages of signal combining: one at each BS for signals received over multiple antenna elements, and one at the backhaul central processing unit (CPU) for signals sent by all BSs. To avoid confusion, signal combining taking place at the backhaul CPU is called *backhaul combining* (BC).

Existing works in cell-free massive MIMO consider either equal-gain backhaul-combining (EBC) [14, 15, 17, 31] or zero-forcing backhaul-combining (ZFBC) [18, 32]. In particular, the ZFBC method in [18, 32] is performed on the signals that are forwarded directly from all the antennas of all BSs to the backhaul CPU (i.e., no combining is performed at each BS). Such a method requires the instantaneous channel state information (CSI) from all users to all BSs in the system, which presents a very large overhead for the backhaul network.

Against the above background, we investigate in this paper the UL data transmission of a *generalized* cell-free massive MIMO-NOMA network with two stages of signal combining. In the first state, the signals received at each BS are combined using the maximum-ratio-combining (MRC) technique. The resulting signals from all the BSs are then *optimally* combined over the backhaul network. The considered cell-free massive MIMO setup becomes the conventional cell-free massive MIMO system when the number of BSs (or APs) is massive and each AP has a single antenna [14]. On the other hand, it becomes a cooperative massive MIMO (or network MIMO) system when there are few BSs, each equipped with a massive antenna array [33,34]. It is pointed out that, while the system model considered in this paper is similar to the one in [14,31], an important difference is that OBC is developed and employed instead of the equal-gain backhaul-combining (EBC). Furthermore, for completeness and to illustrate the superiority of the OBC method, we also develop a ZFBC method that, similar to the OBC, does not require the instantaneous CSI at the backhaul CPU. As such, this ZFBC is markedly different than the ZFBC method in [18, 32].

The optimal backhaul-combining method developed in this paper is to maximize the worst UL SINR among all users in the system without the requirement for instantaneous CSI at backhaul CPU. The analysis focuses on a NOMA scenario, where users are assigned into groups and users in the same group share the same pilot sequence. Such a NOMA approach in pilot design was also employed in [35] and allows more users to be served as compared to OMA. This is very desirable for a cell-free system, in which the number of users tends to be very large due to the large co-coverage of multiple BSs and reusing pilot is inevitable. This approach is especially effective in the scenario that coherence interval is short (e.g., due to high velocity of mobile users), since there would be a short interval available for data transmission if orthogonal pilots are used. The technique of SIC is carried out within each group to improve the achievable UL SE. Although optimal signal combining has been widely studied in adaptive antenna arrays [36], its application in cell-free massive MIMO was only recently examined in [16]. However, no analytical expression for the combining is given in [16]. Instead, the authors formulate an optimal combining optimization problem, and refine it iteratively. As a result, the expression for the achievable UL rate is a function of both transmit power and combining coefficients. In contrast, we provide a tight closed-form lower-bound expression for the achievable UL SE that depends on the users' UL transmit powers only. The closed-form expression reveals many useful observations regarding the performance of a cell-free massive MIMO-NOMA system.

Focusing on the case that there are a few BSs, each equipped with a massive number of antennas (i.e., the cell-free cooperative MIMO setup), the paper also examines the asymptotic behavior of UL SINR when the number of antennas goes to infinity. It was shown in [12,17,37] that correlated interference, as a consequence of using non-orthogonal pilots, cannot be asymptotically mitigated by using a large antenna array at each BS and causes saturation of the system performance. Our recent work concerning time-offset pilots in [35] shows that the correlated interference caused by transmitting pilots simultaneously with data in the training phase can be effectively removed with SIC thanks to the knowledge of all pilot sequences at the BS. However, in the data phase, SIC cannot be applied since the UL data is unknown at the BS. As a consequence, performance in the data phase is still saturated when the number of antennas increases. It shall be shown in this paper that, by using OBC, performance of a cell-free massive MIMO-NOMA system increases proportionally with the number of antennas.

Finally, the paper formulates and solves a power control problem to optimize the system's performance. Unlike most of existing works in NOMA that focus on the maximization of the sum SE, we formulate a max-min QoS optimization problem that maximizes the QoS value that can be equally served to all users in the network. Due to the non-convexity of such an optimization problem, an inner approximation algorithm is developed to solve the optimization problem iteratively. In each iteration, the non-concave cost function is approximated by a concave one so that an alternative convex optimization problem is obtained, whose optimal solution is feasible for the original problem. The proposed method is shown to converge to a suboptimal point of the original problem.

In summary, the main contributions of the paper are as follows[1]:

- We develop the optimal backhaul combining method in a NOMA cell-free massive MIMO system that does not require the instantaneous CSI. In addition to the optimal combining weight vector, the resulting uplink SINR expression is obtained in a closed form as a function of the transmit powers.
- We perform asymptotic analysis of the system's uplink SINR under different backhaul combining methods and show that, unlike EBC, the system's performance of both OBC and ZFBC is not saturated when the number of antennas tends to infinity. Nevertheless, OBC always outperforms ZFBC.
- We propose a user grouping method to further improve the system's performance. The proposed user grouping algorithm is based on minimizing the similarity between large-scale fading profiles of users within a group, which helps to reduce pilot contamination and correlated interference.
- We formulate and solve a max-min QoS power control problem to optimize the system's performance. Due to the non-concave nature of the cost function, an inner approximation is applied to solve the power control problem iteratively, whose solution converges to a suboptimal solution of the original problem.

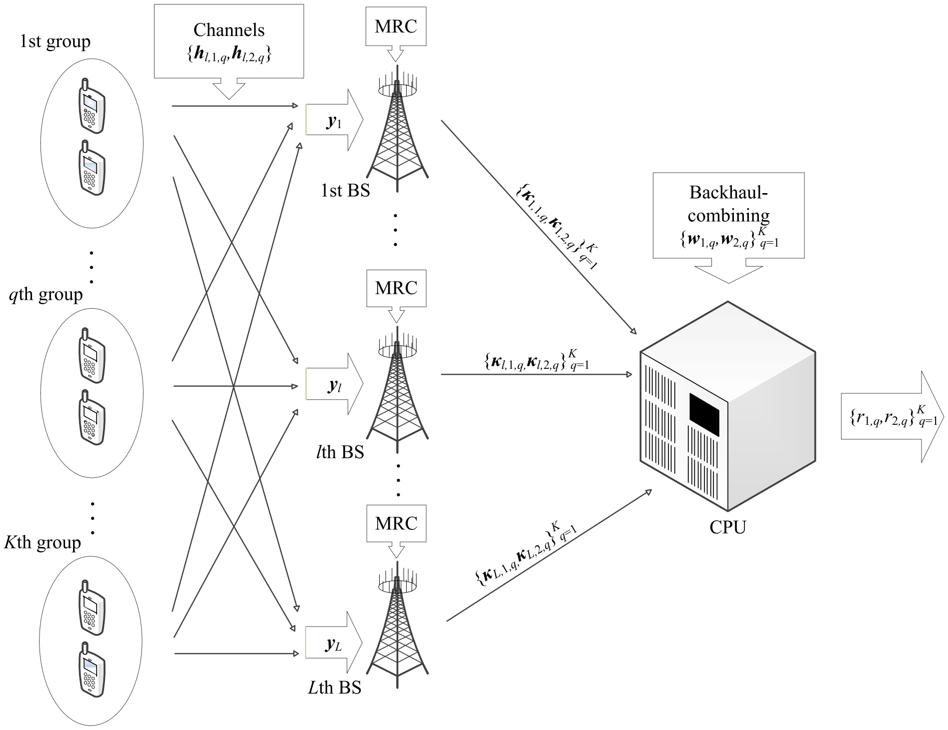
The remainder of this paper is organized as follows. Section 4.2 introduces the system model, including channel estimation and uplink data transmission. Section 4.3 investigates derives a closed-form expression for the optimal combining vector as well as a lower bound of the ergodic uplink spectral efficiency. Section 4.4 performs asymptotic analysis and compares performance of the OBC method to that of the EBC and ZFBC methods. Section 4.5

[1] Some preliminary results are briefly presented in a conference paper [38].

proposes a user grouping algorithm to enhance the system's performance. Section 4.6 studies power optimization problems. Section 4.7 presents simulation results and discussion. Section 4.8 concludes the paper.

4.2 System Model

4.2.1 Generalized Cell-Free Massive MIMO System



1st group
Channels
$\{\boldsymbol{h}_{l,1,q}, \boldsymbol{h}_{l,2,q}\}$
MRC
$\boldsymbol{y}_1$
1st BS
qth group
MRC
$\boldsymbol{y}_l$
lth BS
Kth group
MRC
$\boldsymbol{y}_L$
Lth BS
$\{\boldsymbol{\kappa}_{1,1,q}, \boldsymbol{\kappa}_{1,2,q}\}_{q=1}^{K}$
$\{\boldsymbol{\kappa}_{l,1,q}, \boldsymbol{\kappa}_{l,2,q}\}_{q=1}^{K}$
$\{\boldsymbol{\kappa}_{L,1,q}, \boldsymbol{\kappa}_{L,2,q}\}_{q=1}^{K}$
Backhaul-combining
$\{\boldsymbol{w}_{1,q}, \boldsymbol{w}_{2,q}\}_{q=1}^{K}$
CPU
$\{r_{1,q}, r_{2,q}\}_{q=1}^{K}$

Figure 4.1 System model.

Consider a generalized cell-free massive MIMO system as illustrated in Fig. 4.1, which has L BSs, each equipped with M antennas, to serve $2K$ users. All BSs are connected to a backhaul network over which the signals from all L BSs are sent to and processed at a CPU. As discussed before, such a model becomes the conventional cell-free massive MIMO system when L is very large and $M = 1$ [14], whereas it is a cooperative MIMO system when L is small and M is very large [33, 34]. Similar to [14, 31], in order to accommodate more users with a fixed number of mutually-orthogonal pilot sequences, the users are arranged into K groups with two users in each group who share the same pilot sequence. The pilot sequences assigned to different groups are pairwise orthogonal. The channels between BSs and users are assumed to be flat fading, mutually independent and stay constant within a coherence interval of τ_c symbols satisfying $\tau_c \geq K$.

4.2.2 Channel Estimation

Assuming that the system works in the time-division duplex (TDD) mode, a set of K length-τ_p pilot sequences is used for UL channel estimation. With τ_p symbols used for pilot, the maximum number of pairwise orthogonal pilots available is τ_p. As a result, in order to have enough orthogonal pilots to assign to all K groups, we set $\tau_p = K$. These pilots are collectively represented by a $\tau_p \times K$ pilot matrix $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K]$ which satisfies $\mathbf{\Phi}^{\mathsf{H}}\mathbf{\Phi} = \tau_p \mathbf{I}_K$. With 2 users using the same pilot sequence and different groups using orthogonal pilots, the signal matrix $\mathbf{Y}_l \in \mathbb{C}^{M\times\tau_p}$ received at the lth BS over τ_p time slots (symbols) is given as:

$$\mathbf{Y}_l = \sum_{k=1}^{K} \left(\boldsymbol{h}_{l,1,k}\sqrt{p_{1,k}^{(\mathrm{p})}}\boldsymbol{\phi}_k^{\mathsf{H}} + \boldsymbol{h}_{l,2,k}\sqrt{p_{2,k}^{(\mathrm{p})}}\boldsymbol{\phi}_k^{\mathsf{H}} \right) + \mathbf{N}_l, \tag{4.1}$$

where $p_{g,k}^{(\mathrm{p})}$, $g = 1, 2$, denotes pilot power, $\boldsymbol{h}_{l,g,k} \sim \mathcal{CN}(0, \beta_{l,g,k}\mathbf{I}_M)$ is the uncorrelated Rayleigh fading channel between the gth user of the kth group and the lth BS, and $\beta_{l,g,k}$ is the large scale fading coefficient. Here, $\mathbf{N}_l \in \mathbb{C}^{M\times\tau_p}$ represents AWGN noise.

To estimate the channel for users in the qth group, the lth BS multiplies the received signal with the corresponding pilot of the qth group. This results in:

$$\mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \boldsymbol{h}_{l,1,q}\sqrt{p_{1,q}^{(\mathrm{p})}\tau_p} + \boldsymbol{h}_{l,2,q}\sqrt{p_{2,q}^{(\mathrm{p})}\tau_p} + \mathbf{N}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}. \tag{4.2}$$

Then, the estimate of $\boldsymbol{h}_{l,g,q}$ ($g = 1, 2$) can be obtained by using the minimum mean squared error (MMSE) estimator as [39]:

$$\hat{\boldsymbol{h}}_{l,g,q} = \frac{\mathrm{cov}\left\{\boldsymbol{h}_{l,g,q}, \mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right\}}{\mathrm{var}\left\{\mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right\}} \frac{\mathbf{Y}_l\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \mu_{l,g,q}\frac{\mathbf{Y}_l\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \tag{4.3}$$

where $\mu_{l,g,q} = \frac{\sqrt{p_{g,q}^{(p)}\tau_p}\beta_{l,g,q}}{p_{1,q}^{(p)}\tau_p\beta_{l,1,q} + p_{2,q}^{(p)}\tau_p\beta_{l,2,q} + \sigma_{\mathrm{UL}}^2}$. As a result, the estimated channel is a random vector with distribution $\hat{\boldsymbol{h}}_{l,g,q} \sim \mathcal{CN}(0, \gamma_{l,g,q}\mathbf{I}_M)$, where

$$\gamma_{l,g,q} = \frac{p_{g,q}^{(\mathrm{p})}\tau_p\beta_{l,g,q}^2}{p_{1,q}^{(\mathrm{p})}\tau_p\beta_{l,1,q} + p_{2,q}^{(\mathrm{p})}\tau_p\beta_{l,2,q} + \sigma_{\mathrm{UL}}^2}. \tag{4.4}$$

Furthermore, the channel estimation error $\boldsymbol{e}_{l,g,q} = \boldsymbol{h}_{l,g,q} - \hat{\boldsymbol{h}}_{l,g,q}$ is independent of the estimated channel and distributed as $\boldsymbol{e}_{l,g,q} \sim \mathcal{CN}\left(0, (\beta_{l,g,q} - \gamma_{l,g,q})\mathbf{I}_M\right)$.

4.2.3 UL Data Transmission

Once channel estimation has been acquired in the training phase, uplink data transmission is carried out. As discussed before, in a cell-free massive MIMO system a user can be served by multiple BSs. The signals received by multiple antennas at each BS are first combined. Then the combined signal is sent by each BS over the backhaul network to the CPU. At the CPU, the multiple signals sent by all BSs are then combined again (i.e., backhaul combining) for detecting each user's signal.

The UL data signal received at the lth BS over each symbol time can be presented as:

$$\boldsymbol{y}_l = \sum_{g=1}^{2}\sum_{k=1}^{K} \boldsymbol{h}_{l,g,k}\sqrt{p_{g,k}}x_{g,k} + \boldsymbol{n}_l, \tag{4.5}$$

where $p_{g,k}$ is the UL transmit power of the gth user of the kth group, $x_{g,k}$ represents its data signal which has zero mean and unit power and $\boldsymbol{n_l}$ denotes AWGN noise. In order to extract the signal of the first user of the qth group, the signals received by different antenna elements of the BS are combined using the MRC rule. This is achieved by multiplying $\boldsymbol{y}_l$ with the MRC combining vector $\boldsymbol{v}_{l,1,q} = \hat{\boldsymbol{h}}_{l,1,q}$, which yields:

$$\kappa_{l,1,q} = \boldsymbol{v}_{l,1,q}^{\mathsf{H}} \left(\sum_{g=1}^{2}\sum_{k=1}^{K} \boldsymbol{h}_{l,g,k}\sqrt{p_{g,k}}x_{g,k} + \boldsymbol{n}_l \right). \tag{4.6}$$

The signal components sent by all L BSs over the backhaul network are received by the CPU. Assuming error-free transmission over the backhaul network, the L signal components received at the CPU are collected in a $L \times 1$ signal vector $\boldsymbol{\kappa}_{1,q} = [\kappa_{1,1,q}, \kappa_{2,1,q}, \ldots, \kappa_{L,1,q}]^{\mathsf{T}}$. These L components are then combined together using a $L \times 1$ weighting vector $\boldsymbol{w}_{1,q}$, which leads to:

$$r_{1,q} = \boldsymbol{w}_{1,q}^{\mathsf{T}}\boldsymbol{\kappa}_{1,q} = \sum_{l=1}^{L} w_{l,1,q}\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{y}_l. \tag{4.7}$$

Our objective is to find the OBC weights to maximize the SINR for each user. This is accomplished in the next section.

4.3 Optimal Backhaul Combining

Finding the OBC weights to maximize SINR amounts to finding the covariance matrix of the signal vector $\boldsymbol{\kappa}_{1,q}$. In essence, this means that we need to find the correlation between any two components in $\boldsymbol{\kappa}_{1,q}$.

First, decompose the signal in Eqn. (4.6) as:

$$\begin{aligned}\kappa_{l,1,q} = &\underbrace{\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}\sqrt{p_{1,q}}x_{1,q}}_{\text{DS}_{l,1,q}\text{- Desired signal}} + \underbrace{\left(\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q} - \mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}\right)\sqrt{p_{1,q}}x_{1,q}}_{\text{CU}_{l,1,q}\text{- Channel gain uncertainty}} \\ &+ \underbrace{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,2,q}\sqrt{p_{2,q}}x_{2,q}}_{\text{IwG}_{l,1,q}\text{- Interference within group}} + \underbrace{\sum_{g=1}^{2}\sum_{k=1,k\neq q}^{K}\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,g,k}\sqrt{p_{g,k}}x_{g,k}}_{\text{IoG}_{l,1,q}\text{- Interference from other group}} + \underbrace{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{n}_{l}}_{\text{N}_{l,1,q}\text{- Noise}},\end{aligned} \tag{4.8}$$

The decomposition of the received signal in Eqn. (4.8) has an intuitive structure. The first component, $\text{DS}_{l,1,q}$, is the desired signal, which experiences a constant gain $s_{l,1,q} = \mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}\sqrt{p_{1,q}}$. Due to imperfect CSI at the BSs, the second term $\text{CU}_{l,1,q}$ is the interference originating from the desired signal itself, which is independent from the first term. The last three terms represent interference from other users and thermal noise [14–16, 40].

The analysis in Appendix 4.A reveals that, except the third component in Eqn. (4.8), all other components are uncorrelated across BSs. The third component is correlated across the BSs, i.e., the correlation between $\text{IwG}_{l,1,q}$ and $\text{IwG}_{l',1,q}$ is non-zero whenever $l \neq l'$. To see what the correlation value is, expand the term $\text{IwG}_{l,1,q}$ as:

$$\begin{aligned}\text{IwG}_{l,1,q} = &\mu_{l,1,q}\boldsymbol{h}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,2,q}\sqrt{p_{1,q}^{(\text{p})}p_{2,q}\tau_p}x_{2,q} + \underbrace{\mu_{l,1,q}\mathbb{E}\left\{\boldsymbol{h}_{l,2,q}^{\mathsf{H}}\boldsymbol{h}_{l,2,q}\right\}\sqrt{p_{2,q}^{(\text{p})}p_{2,q}\tau_p}}_{c_{l,1,q}}x_{2,q} \\ &+ \mu_{l,1,q}\left(\boldsymbol{h}_{l,2,q}^{\mathsf{H}}\boldsymbol{h}_{l,2,q} - \mathbb{E}\left\{\boldsymbol{h}_{l,2,q}^{\mathsf{H}}\boldsymbol{h}_{l,2,q}\right\}\right)\sqrt{p_{2,q}^{(\text{p})}p_{2,q}\tau_p}x_{2,q} \\ &+ \mu_{l,1,q}\left(\mathbf{N}_l\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^{\mathsf{H}}\boldsymbol{h}_{l,2,q}\sqrt{p_{2,q}}x_{2,q}.\end{aligned} \tag{4.9}$$

In Eqn. (4.9), the second term is correlated across the BSs since:

$$\text{cov}\left\{c_{l,1,q}x_{2,q}, c_{l'1,q}x_{2,q}\right\} = \mu_{l,1,q}\mu_{l',1,q}p_{2,q}^{(\text{p})}p_{2,q}\tau_p\beta_{l,2,q}\beta_{l',2,q}M^2, \forall l \neq l', \tag{4.10}$$

whereas all other components are uncorrelated across BSs. As a result, all the interference and noise terms in $\boldsymbol{\kappa}_{1,q}$ can be grouped into two length-L vectors: the uncorrelated interference-plus-noise $\boldsymbol{u}_{1,q}$, and the correlated interference-plus-noise $\boldsymbol{c}_{1,q}$. Specifically,

$$\boldsymbol{\kappa}_{1,q} = \boldsymbol{s}_{1,q} x_{1,q} + \boldsymbol{c}_{1,q} x_{2,q} + \boldsymbol{u}_{1,q}, \tag{4.11}$$

where $\boldsymbol{s}_{1,q} = [s_{1,1,q}, s_{2,1,q}, \ldots, s_{L,1,q}]$, and the elements of $\boldsymbol{u}_{1,q}$ and $\boldsymbol{c}_{1,q}$ are as follows:

$$u_{l,1,q} = \mathrm{CU}_{l,1,q} + \mathrm{IoG}_{l,1,q} + \mathrm{N}_{l,1,q} + \mathrm{IwG}_{l,1,q} - \mu_{l,1,q} \mathbb{E}\left\{\boldsymbol{h}_{l,2,q}^{\mathsf{H}} \boldsymbol{h}_{l,2,q}\right\} \sqrt{p_{2,q}^{(\mathrm{p})} p_{2,q} \tau_p} x_{2,q}, \tag{4.12}$$

$$c_{l,1,q} = \mu_{l,1,q} \mathbb{E}\left\{\boldsymbol{h}_{l,2,q}^{\mathsf{H}} \boldsymbol{h}_{l,2,q}\right\} \sqrt{p_{2,q}^{(\mathrm{p})} p_{2,q} \tau_p}. \tag{4.13}$$

Next, it is convenient to normalize (i.e., scale) the signal vector in Eqn. (4.11) so that the uncorrelated interference-plus-noise term, $u_{l,1,q}$ is normalized to $\hat{u}_{l,1,q}$, which has a unit power (i.e., $\mathbb{E}\left\{|\hat{u}_{l,1,q}|^2\right\} = 1$). This is achieved by simply diving the lth element by $\mathbb{E}\left\{|u_{l,1,q}|\right\}$. It follows from the analysis in Appendix 4.D that

$$\begin{aligned}
\mathbb{E}\left\{|u_{l,1,q}|^2\right\} &= \left(\sum_{g=1}^{2}\sum_{k=1}^{K} p_{g,k}\beta_{l,g,k} + \sigma_{\mathrm{UL}}^2\right)\gamma_{l,1,q}M, \\
\mathbb{E}\left\{|s_{l,1,q}|^2\right\} &= p_{1,q}\gamma_{l,1,q}^2 M^2, \\
\mathbb{E}\left\{|c_{l,1,q}|^2\right\} &= p_{2,q}\frac{p_{2,q}^{(\mathrm{p})}}{p_{1,q}^{(\mathrm{p})}}\left(\gamma_{l,1,q}\frac{\beta_{l,2,q}}{\beta_{l,1,q}}\right)^2 M^2.
\end{aligned}$$

The normalization produces the following equivalent signal vector:

$$\hat{\boldsymbol{\kappa}}_{1,q} = \hat{\boldsymbol{s}}_{1,q} x_{1,q} + \hat{\boldsymbol{c}}_{1,q} x_{2,q} + \hat{\boldsymbol{u}}_{1,q}, \tag{4.15}$$

where $\mathbb{E}\left\{\hat{\boldsymbol{u}}_{1,q}\hat{\boldsymbol{u}}_{1,q}^{\mathsf{H}}\right\} = \mathbf{I}_L$. Furthermore, it can be shown that the variance of the normalized effective channel gain $\hat{s}_{l,1,q}$ is exactly the signal to uncorrelated-interference-plus noise ratio of the first user of the qth group at the lth BS:

$$\mathbb{E}\left\{|\hat{s}_{l,1,q}|^2\right\} = \frac{M p_{1,q}\gamma_{l,1,q}}{\sum_{g=1}^{2}\sum_{k=1}^{K} p_{g,k}\beta_{l,g,k} + \sigma_{\mathrm{UL}}^2} \triangleq \mathrm{SNR}_{l,1,q}. \tag{4.16}$$

Likewise, the variance of each element in the normalized correlated interference term $\hat{c}_{l,1,q}$ is

$$\mathbb{E}\left\{|\hat{c}_{l,1,q}|^2\right\} = \frac{M p_{2,q}\frac{p_{2,q}^{(\mathrm{p})}}{p_{1,q}^{(\mathrm{p})}}\gamma_{l,1,q}\left(\frac{\beta_{l,2,q}}{\beta_{l,1,q}}\right)^2}{\sum_{g=1}^{2}\sum_{k=1}^{K} p_{g,k}\beta_{l,g,k} + \sigma_{\mathrm{UL}}^2}. \tag{4.17}$$

If the pilot power is set at the maximum UL transmit power, i.e., $p_{g,q}^{(\mathrm{p})} = p_{\max}, \forall g, q$, we obtain:

$$\mathbb{E}\left\{|\hat{c}_{l,1,q}|^2\right\} = \frac{M p_{2,q}\gamma_{l,2,q}}{\sum_{g=1}^{2}\sum_{k=1}^{K} p_{g,k}\beta_{l,g,k} + \sigma_{\mathrm{UL}}^2} \triangleq \mathrm{SNR}_{l,2,q}, \tag{4.18}$$

which is exactly the signal-to-uncorrelated-interference ratio of the second user of the qth group at the lth BS.

Finally, the covariance matrix of the total interference-plus-noise in Eqn. (4.15) is given as

$$\begin{aligned}\hat{\mathbf{R}}_{1,q} &= \mathbb{E}\left\{\left(\hat{\boldsymbol{\kappa}}_{1,q} - \hat{\boldsymbol{s}}_{1,q}x_{1,q}\right)\left(\hat{\boldsymbol{\kappa}}_{1,q} - \hat{\boldsymbol{s}}_{1,q}x_{1,q}\right)^{\mathsf{H}}\right\} \\ &= \mathbb{E}\left\{\hat{\boldsymbol{u}}_{1,q}\hat{\boldsymbol{u}}_{1,q}^{\mathsf{H}} + \hat{\boldsymbol{c}}_{1,q}\hat{\boldsymbol{c}}_{1,q}^{\mathsf{H}}\right\} = \hat{\boldsymbol{c}}_{1,q}\hat{\boldsymbol{c}}_{1,q}^{\mathsf{H}} + \mathbf{I}_L. \end{aligned}\tag{4.19}$$

Following [36], the optimal combining coefficients to maximize the individual user's SINR and the corresponding effective SINR expression are given in Theorem 1.

Theorem 1: The combining vector that maximizes the SINR of the combined signal in Eqn. (4.7), in which the desired signal vector is $\hat{\boldsymbol{s}}_{1,q}x_{1,q}$ and the covariance matrix of the total interference-plus-noise component is $\hat{\mathbf{R}}_{1,q}$, is

$$\boldsymbol{w}_{1,q}^{(\mathrm{OBC})} = \alpha\hat{\mathbf{R}}_{1,q}^{-1}\hat{\boldsymbol{s}}_{1,q} = \alpha\left(\hat{\boldsymbol{c}}_{1,q}\hat{\boldsymbol{c}}_{1,q}^{\mathsf{H}} + \mathbf{I}_L\right)^{-1}\hat{\boldsymbol{s}}_{1,q}, \tag{4.20}$$

where α is a constant. Furthermore, the resulting maximum effective SINR is given as:

$$\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} = \hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\hat{\mathbf{R}}_{1,q}^{-1}\hat{\boldsymbol{s}}_{1,q} = \sum_{l=1}^{L}\mathrm{SNR}_{l,1,q} - \frac{\left(\sum_{l=1}^{L}\sqrt{\mathrm{SNR}_{l,1,q}}\sqrt{\mathrm{SNR}_{l,2,q}}\right)^2}{1 + \sum_{l=1}^{L}\mathrm{SNR}_{l,2,q}}. \tag{4.21}$$

Proof: Please see Appendices 4.B and 4.C. ■

Corollary 1: By considering uncorrelated Gaussian noise as the worst-case distribution of noise and interference [14–16, 40], a lower bound on the UL spectral efficiency of the first user of the qth group with OBC is given as:

$$R_{1,q}^{(\mathrm{OBC})} \geq \left(1 - \frac{\tau_p}{\tau_c}\right)\log_2\left(1 + \mathrm{SINR}_{1,q}^{(\mathrm{OBC})}\right). \tag{4.22}$$

It is pointed out that the above lower bound on spectral efficiency is based on the property of the mutual information as established in [41]. Such a lower bound is optimized so that the gap between this lower bound and the actual value is minimized. This lower bound is widely used for performance analysis in massive MIMO research (e.g., see references [11–18, 25–31, 37, 40–42]). It shall also be used as the QoS metric for power control optimization in Section 4.6, as well as performance analysis and comparison in Section 4.8.

Remark 1: The system model considered in this section can be extended to accommodate more than two users in one group, which allows more users to be served by the system [2, 5–7, 10]. However, there are three main drawbacks of assigning more than two users into one group. First, having more users in one group makes it complicated to find a closed-form SINR expression for the UL SE with OBC because of the high complexity in the structure of the eigenvalues of the interference-plus-noise covariance matrix. Second, as shown in Appendix 4.F, in order to keep the SINR proportional to the number of BSs' antennas, the number of users in each group should not exceed the number of participating BSs. Since the last user to detect data in each group will have to decode and subtract the signals from all other users, the third disadvantage is the larger delay and signal processing overhead. As a result, assigning 2 users in a group appears most attractive and practical.

Although all the analysis and expressions obtained in the previous subsections are for the first user in each group, the same results apply to the second user in the group as well.

In case the detection of one user (say, without loss of generality, the first user) is very good and can be assumed ideal, then it is possible to apply SIC to subtract the detected signal of the first user from the UL signal before detecting the second user in each group. This is a reasonable assumption and the SIC technique is commonly used in various works on massive MIMO-NOMA [21, 26]. As can be seen from Eqn. (4.13), by subtracting the correlated interference term, the correlation matrix of the normalized interference-plus-noise component for the second user of the qth group becomes an identity matrix, i.e., $\hat{\mathbf{R}}_{2,q} = \mathbb{E}\left\{\hat{\boldsymbol{u}}_{2,q}\hat{\boldsymbol{u}}_{2,q}^{\mathsf{H}} + \hat{\boldsymbol{c}}_{2,q}\hat{\boldsymbol{c}}_{2,q}^{\mathsf{H}}\right\} = \mathbf{I}_L$. It then follows that the OBC vector for the second user of the qth group simplifies to:

$$\boldsymbol{w}_{2,q}^{(\mathrm{SIC-OBC})} = \alpha\hat{\mathbf{R}}_{2,q}^{-1}\hat{\boldsymbol{s}}_{2,q} = \alpha\mathbf{I}_L^{-1}\hat{\boldsymbol{s}}_{2,q} = \alpha\hat{\boldsymbol{s}}_{2,q}, \tag{4.23}$$

which is equivalent to a maximum ratio combining method. As a result, the SINR of the

$$\mathrm{SINR}_{1,q}^{(\mathrm{EBC})}$$

$$= \frac{Mp_{1,q}\left(\sum_{l=1}^{L}\gamma_{l,1,q}\right)^2}{\sum_{l=1}^{L}\sum_{g=1}^{2}\sum_{k=1}^{K}p_{g,k}\beta_{l,g,k}\gamma_{l,1,q} + Mp_{2,q}\frac{p_{2,q}^{(\mathrm{p})}}{p_{1,q}^{(\mathrm{p})}}\left(\sum_{l=1}^{L}\gamma_{l,1,q}\frac{\beta_{l,2,q}}{\beta_{l,1,q}}\right)^2 + \sigma_{\mathrm{UL}}^2\sum_{l=1}^{L}\gamma_{l,1,q}}. \tag{4.25}$$

second user of the qth group with SIC becomes:

$$\mathrm{SINR}_{2,q}^{(\mathrm{SIC-OBC})} = \sum_{l=1}^{L}\mathrm{SNR}_{l,2,q}. \tag{4.24}$$

Remark 2: If users are assigned mutually orthogonal pilot sequences (i.e., in an OMA cell-free massive MIMO system), correlated interference does not exist, i.e., $\hat{\mathbf{R}}_{1,q} = \hat{\mathbf{R}}_{2,q} = \mathbf{I}_L, \forall q$. In this case, SIC is not needed and the optimal combining vector and the corresponding SINR are similar to Eqns. (4.23) and (4.24), respectively.

4.4 Asymptotic Analysis

4.4.1 Equal-Gain Backhaul Combining and Zero-Forcing Backhaul Combining

In order to illustrate the advantage of OBC in the considered cell-free massive MIMO-NOMA system, it is of interest to compare its performance with those of EBC and ZFBC.

First, with EBC, the combining vector at the CPU is simply $\boldsymbol{w}_{g,k}^{(\mathrm{EBC})} = [1, 1, \ldots, 1]^{\mathsf{T}}$. Appendix 4.D shows that the resulting SINR is given as in Eqn. (4.25), shown on top of the next page.

On the other hand, the principle of ZFBC is to null the interference in the received signal. It follows from Eqn. (4.15) that the ZFBC can be obtained as:

$$\left[\boldsymbol{w}_{1,k}^{(\mathrm{ZFBC})}, \boldsymbol{w}_{2,k}^{(\mathrm{ZFBC})}\right] = \Theta\left(\Theta^{\mathsf{H}}\Theta\right)^{-1} \in \mathbb{C}^{L\times 2}, \tag{4.26}$$

where $\Theta = [\hat{\boldsymbol{s}}_{1,q}, \hat{\boldsymbol{c}}_{1,q}] \in \mathbb{C}^{L\times 2}$. With this combining vector, Appendix 4.E shows that the resulting SINR is given as:

$$\mathrm{SINR}_{1,q}^{(\mathrm{ZFBC})} = \sum_{l=1}^{L} \mathrm{SNR}_{l,1,q} - \frac{\left(\sum_{l=1}^{L} \sqrt{\mathrm{SNR}_{l,1,q}}\sqrt{\mathrm{SNR}_{l,2,q}}\right)^2}{\sum_{l=1}^{L} \mathrm{SNR}_{l,2,q}}. \tag{4.27}$$

4.4.2 Asymptotic Analysis

The asymptotic analysis in this section focuses on the cell-free cooperative MIMO setup, i.e., when L is small and $M \to \infty$. Similar results can be obtained for the conventional cell-free massive MIMO setup (i.e., $M = 1$ and $L \to \infty$).

First, for EBC, by dividing both the numerator and denominator of Eqn. (4.25) to M, both the desired signal power and interference power from users in the same group remain finite when the number of antennas goes to infinity. This means that both the SINR and the UL SE are saturated when the number of antennas grows without bound. The saturated value of the SINR can be found by calculating the limit of Eqn. (4.25) when $M \to \infty$:

$$\lim_{M\to\infty} \mathrm{SINR}_{l,1,q}^{(\mathrm{EBC})} = \frac{p_{1,q}\left(\sum_{l=1}^{L} \gamma_{l,1,q}\right)^2}{p_{2,q}\frac{p_{2,q}^{(\mathrm{p})}}{p_{1,q}^{(\mathrm{p})}}\left(\sum_{l=1}^{L} \gamma_{l,1,q}\frac{\beta_{l,2,q}}{\beta_{l,1,q}}\right)^2}. \tag{4.28}$$

The above result can be explained by the fact that both the desired signal $\mathrm{DS}_{l,1,q}$ and correlated interference originating from pilot contamination are proportional with the number of antennas. As a consequence, when the number of antennas goes to infinity, the SINR achieved with EBC is saturated around the value in Eqn. (4.28). It should be also noted that the same effect happens with the max-SNR association method, where only one BS (with the highest channel quality) serves each user [31, 41]. By plugging $L = 1$ into (4.28), it can be seen that the resulting SINR is also bounded at a finite value.

For the case of OBC, with some simple manipulations, the SINR of the first user of the qth group in Eqn. (4.21) can be rewritten as in Eqn. (4.29).

It can be easily seen that the first terms in both the numerator and the denominator of Eqn. (4.29) are proportional to M. Meanwhile, the second term of the numerator is in the form of a product of M^2 with a nonnegative scalar, say $\nu_{1,q}$ (corresponding to the first

$$
\begin{aligned}
&\mathrm{SINR}_{1,q}^{(\mathrm{OBC})}\\
&\quad= \frac{\sum_{l=1}^{L}\mathrm{SNR}_{l,1,q}+\frac{1}{2}\sum_{l=1}^{L}\sum_{l'=1}^{L}\left(\sqrt{\mathrm{SNR}_{l,1,q}}\sqrt{\mathrm{SNR}_{l',2,q}}-\sqrt{\mathrm{SNR}_{l',1,q}}\sqrt{\mathrm{SNR}_{l,2,q}}\right)^2}{1+\sum_{l=1}^{L}\mathrm{SNR}_{l,2,q}}.
\end{aligned}
\tag{4.29}
$$

user of the qth group). As a result, if $\nu_{1,q} \neq 0$, the SINR increases without bound when the number of antennas goes to infinity, i.e.,

$$\lim_{M\to\infty} \mathrm{SINR}_{1,q}^{(\mathrm{OBC})} = \infty. \tag{4.30}$$

Similar analysis shows that ZFBC can also enjoy unlimited performance like OBC when $M \to \infty$. However, it is simple to see that the SINR corresponding to OBC in Eqn. (4.29) is always higher than the SINR corresponding to ZFBC in Eqn. (4.27).

4.5 User Grouping

User grouping plays an important role in NOMA. In general, performance gain with NOMA can only be obtained when the channel conditions of users are different [2]. Moreover, when reusing pilots, the distance between users using the same pilot strongly affects the quality of channel estimation [12, 14, 17]. Therefore, an optimal user group assignment is desired to further improve the system's performance.

Unfortunately, optimization of user grouping is a combinatorial problem and hence, cannot be solved in polynomial time. In the conventional NOMA approach, user grouping and decoding order are usually based on large-scale fading. In particular, for a single-cell system, users with the best channels are usually grouped with users with the worst channels. Within each group, the users whose channels are better (closer to the BS) will have their signals detected first, followed by users having the next best channels, and so on. As a result, an arbitrary user can remove the known UL signals of all users which have been already detected before by using SIC. However, this approach is not possible in the cell-free setup, where it is not straightforward to determine which users have better channels because each user communicates with multiple BSs.

Furthermore, in massive MIMO, users with similar channel conditions tend to severely contaminate each other's channel estimation when using the same pilot. For the case of cell-free massive MIMO where there are multiple BSs, strong pilot contamination appears when assigning the same pilot for users in close vicinity of each other [12,14,17,26,42]. In NOMA, this problem is usually addressed by exploiting the similarity in channel statistics among users. For example, in [26], the authors propose a grouping method based on the similarity among the channel matrices of users in a cell to the serving BS. In a NOMA cell-free system, the authors in [42] utilize the Jaccard coefficient to calculate the similarity between each user's large-scale fading profile with a predetermined centroid. Then users having strong similarity coefficients will be assigned into different groups.

For the system model considered in this paper, in order to mitigate the effect of pilot contamination, the similarity of large-scale fading profiles of two users within a group shall be minimized. Before acquiring any group assignment, define $\boldsymbol{\beta}_i \in \mathbb{C}^{L\times 1}$ as the vector containing the large-scale fading coefficients from the ith user to all L BSs (for simplicity, the group index has been dropped). Inspired by the channel matrix similarity in [26,27], the similarity of the large-scale fading profiles of the ith and the jth users can be quantified by the following correlation coefficient:

$$\lambda_{i,j} = \frac{\left\| \boldsymbol{\beta}_i\boldsymbol{\beta}_i^{\mathsf{H}} \left[\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathsf{H}}\right]^{\mathsf{H}} \right\|}{\left\|\boldsymbol{\beta}_i\boldsymbol{\beta}_i^{\mathsf{H}}\right\| \left\|\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathsf{H}}\right\|}, \tag{4.31}$$

For $i = j$, the correlation coefficient equals 1, i.e., $\lambda_{i,i} = 1, \forall i$. For $i \neq j$, the smaller $\lambda_{i,j}$ is, the less pilot contamination occurs between the two users. Therefore, the grouping problem becomes choosing K out of $(2K)^2$ values of $\{\lambda_{i,j}\}$, $\forall i, j$, such that the maximum value is minimized. This assignment problem can be accomplished by Algorithm 3, whose main steps are described below.

Initially, all possible pairs of users are saved into a class named ϑ with three properties:

- $\vartheta.user1$ is the first user of the pair.
- $\vartheta.user2$ is the second user of the pair.
- $\vartheta.value$ is the correlation value between the two users' large-scale fading profiles.

Algorithm 3 User Grouping

```
Require: Large-scale fading correlation coefficients $\{\lambda_{i,j}\} \forall i, j$.
Let $\mathcal{S}$ be the set of grouped users. Initially, $\mathcal{S} = \emptyset$
Step 1: Save all possible pairs $\{\lambda_{i,j}\}, \forall i < j$ into a class named $\vartheta$
$n = 0$;
for $i = 1 : 2K$ do
    for $j > i$ do
        $n = n + 1$;
        $\vartheta(n).value = \lambda_{i,j}$;    //Correlation value
        $\vartheta(n).user1 = i$;    //First user
        $\vartheta(n).user2 = j$;    //Second user
    end for
end for
Step 2: Sort $\vartheta$ in ascending order of $\vartheta.value$.
Step 3: User grouping
Define $\chi$ as the position of the worst pair of users after grouping, initially: $\chi = 0$
$k = 0$;
Let $\mathcal{S}_c = \{1, \ldots, 2K\}$ be the set of users who are not in any pairs of the first $k$ pairs.
while $\bar{\mathcal{S}} \neq \emptyset$ do {Stop when all users are grouped}
    Scan from the first pair until all users are in at least one pair
    while $\mathcal{S}_c \neq \emptyset$ do
        $k = k + 1$;
        if $\{\{\vartheta(k).user1\}, \{\vartheta(k).user2\}\} \in \mathcal{S}_c$ then
            $\mathcal{S}_c = \mathcal{S}_c \setminus \{\{\vartheta(k).user1\}, \{\vartheta(k).user2\}\}$
        end if
    end while
    if $\chi = 0$ then
        $\chi = k$; // Save worst pair's position.
    end if
    if $k \leq \chi$ then {If new pair is better than the worst pair}
        $\mathcal{S} = \mathcal{S} \cup \{\vartheta(k).user1, \vartheta((k).user2)\}$; //Add new pair
    else {If new pair is worse that the worst pair}
        $\mathcal{S} = \emptyset$; // Clear all current assignment.
        $\chi = \chi + 1$; //Update new worst pair's position.
        $\mathcal{S} = \mathcal{S} \cup \{\vartheta(\chi).user1, \vartheta((\chi).user2)\}$; // Add this pair as the worst pair
    end if
    $\mathcal{S}_c = \{1, \ldots, 2K\} \setminus \mathcal{S}$; //Paired users are not re-scanned; in the next loop.
    $k = 0$;
end while
return $\mathcal{S}$
```

After that, all $(2K)^2$ elements of the class are sorted in an ascending order of $\vartheta.value$. Then, the sorted class is scanned from the first pair (i.e., the pair having the lowest correlation value between two users' channels) until every user appears in at least one pair. The pair where the scanning stops is chosen as the worst group position, denoted χ. This pair is then added into the set of grouped users, denoted as $\mathcal{S}$. In the next iteration, the class is re-scanned to find the next worst pair, but all pairs containing a user in the worst pair of the previous loop will not be scanned again. If the worst pair of the current loop is worst than the pair at position χ, all group assignment is cleared. The worst pair position is updated as $\chi = \chi + 1$ before the next iteration. The algorithm continues until all users are grouped. With this algorithm, K pairs of users are formed such that the maximum correlation value is minimized.

4.6 Power Control Optimization

For the max-min QoS power control problem considered in this section, it is assumed that some method of user grouping has been applied and the objective is to maximize the minimum rate among users subject to a maximum power constraint. Focusing on the case that SIC is applied to the second user in each group, the power control problem at hand can be formulated as:

$$\begin{aligned} &\underset{p_{g,q}}{\text{maximize}} \min_{q=1,\ldots,K} \left\{ R_{1,q}^{(\mathrm{OBC})}, R_{2,q}^{(\mathrm{SIC-OBC})} \right\} \\ &\text{subject to} \quad 0 \leq p_{g,q} \leq p_{\max}, \forall g, q, \end{aligned} \tag{4.32}$$

where $p_{\max}$ is the maximum transmit power of each user. The above min-QoS maximization is equivalent to maximizing the minimum UL SINR, which is stated as:

$$\max_{p_{g,q}} \min_{q=1,\ldots,K} \{ M \cdot R_{1,q}(p), M \cdot R_{2,q}(p) \} \tag{4.33a}$$

$$\text{subject to} \quad 0 \leq p_{g,q} \leq p_{\max}, \forall p, q. \tag{4.33b}$$

where

$$R_{2,q}(p) \triangleq \frac{1}{M} \mathrm{SINR}_{2,q}^{(\mathrm{SIC-OBC})} = p_{2,q} \sum_{l=1}^{L} \frac{\gamma_{l,2,q}}{\eta_l(p)},$$

$$R_{1,q}(p) \triangleq \frac{1}{M} \mathrm{SINR}_{1,q}^{(\mathrm{OBC})} = p_{1,q} \sum_{l=1}^{L} \frac{\gamma_{l,1,q}}{\eta_l(p)} - p_{1,q} \frac{\left(\displaystyle\sum_{l=1}^{L} \frac{\gamma_{l,q}}{\eta_l(p)} \right)^2}{\dfrac{1}{M p_{2,q}} + \displaystyle\sum_{l=1}^{L} \frac{\gamma_{l,2,q}}{\eta_l(p)}}.$$

In the above expressions, $\eta_l(p) \triangleq \sum_{g=1}^{2}\sum_{k=1}^{K} p_{g,k}\beta_{l,g,k} + \sigma_{\mathrm{UL}}^2$ is an affine positive function, whereas $\gamma_{l,q} \triangleq \sqrt{\gamma_{l,1,q}\gamma_{l,2,q}} > 0$. The above optimization problem is obviously equivalent to

$$\begin{aligned}\max_{p_{q,q}} \varphi(p) \triangleq \min_{q=1,\ldots,K} &\{R_{1,q}(p), R_{2,q}(p)\}\\ \text{subject to} \quad &(4.33b),\end{aligned} \tag{4.34}$$

which is a nonconvex problem because its objective function $\varphi(p)$ is nonconcave.

The remaining of this section presents a method to solve the above optimization problem. First, let $p^{(\varpi)}$ be a feasible point for (4.34) that is found from the $(\varpi - 1)$th iteration. By using the inequalities (4.76) and (4.77) in Appendix 4.G, we have

$$\begin{aligned}
& R_{2,q}(p) \\
\geq\ & p_{2,q}\sum_{l=1}^{L}\left(\frac{2\gamma_{l,2,q}}{\eta_l(p^{(\varpi)})} - \frac{\gamma_{l,2,q}\eta_l(p)}{(\eta_l(p^{(\varpi)}))^2}\right) \\
=\ & p_{2,q}\sum_{l=1}^{L}\frac{2\gamma_{l,2,q}}{\eta_l(p^{(\varpi)})} - p_{2,q}\sum_{l=1}^{L}\frac{\gamma_{l,2,q}\eta_l(p)}{(\eta_l(p^{(\varpi)}))^2} \\
\geq\ & p_{2,q}\sum_{l=1}^{L}\frac{2\gamma_{l,2,q}}{\eta_l(p^{(\varpi)})} \\
-\ & \frac{R_{2,q}(p^{(\varpi)})}{4}\left(\frac{p_{2,q}}{p_{2,q}^{(\varpi)}} + \frac{p_{2,q}^{(\varpi)}}{R_{2,q}(p^{(\varpi)})}\sum_{l=1}^{L}\frac{\gamma_{l,2,q}\eta_l(p)}{(\eta_l(p^{(\varpi)}))^2}\right)^2 \\
\triangleq\ & R_{2,q}^{(\varpi)}(p),
\end{aligned} \tag{4.35}$$

where $R_{2,q}^{(\varpi)}(p)$ is concave quadratic. Analogously, for $\tilde{R}_{1,q}(p) \triangleq p_{1,q}\sum_{l=1}^{L}\frac{\gamma_{l,1,q}}{\eta_l(p)}$, it is true that

$$\tilde{R}_{1,q}(p) \geq \tilde{R}_{1,q}^{(\varpi)}(p) \triangleq p_{1,q}\sum_{l=1}^{L}\frac{2\gamma_{l,1,q}}{\eta_l(p^{(\varpi)})} - \frac{\tilde{R}_{1,q}(p^{(\varpi)})}{4}\left(\frac{p_{1,q}}{p_{1,q}^{(\varpi)}} + \frac{p_{1,q}^{(\varpi)}}{\tilde{R}_{1,q}(p^{(\varpi)})}\sum_{l=1}^{L}\frac{\gamma_{l,1,q}\eta_l(p)}{(\eta_l(p^{(\varpi)}))^2}\right)^2. \tag{4.36}$$

Now, introduce a positive variable x_q satisfying

$$\bar{R}_{1,q}(p) \triangleq p_{1,q}\frac{\left(\sum_{l=1}^{L}\frac{\gamma_{l,q}}{\eta_l(p)}\right)^2}{\frac{1}{p_{2,q}} + \sum_{l=1}^{L}\frac{\gamma_{l,2,q}}{\eta_l(p)}} \leq (x_q)^2. \tag{4.37}$$

Then one has

$$R_{1,q}(p) \geq R_{1,q}^{(\varpi)}(p, x_q) \triangleq \tilde{R}_{1,q}^{(\varpi)}(p) - (x_q)^2, \tag{4.38}$$

where $R_{1,q}^{(\varpi)}(p, x_q)$ is a concave quadratic function.

The next step is to handle the nonconvex constraint (4.37), which is equivalent to

$$\frac{\left(\sum_{l=1}^{L} \frac{\gamma_{l,q}}{\eta_l(p)}\right)^2}{\frac{1}{p_{2,q}} + \sum_{l=1}^{L} \frac{\gamma_{l,2,q}}{\eta_l(p)}} \leq \frac{(x_q)^2}{p_{1,q}}. \tag{4.39}$$

By using inequality (4.78) in Appendix 4.G, one has

$$\text{RHS of (4.39)} = \frac{(x_q)^2}{p_{1,q}} \geq \frac{2x_q^{(\varpi)}}{p_{1,q}^{(\varpi)}} x_q - \frac{(x_q^{(\varpi)})^2}{(p_{1,q}^{(\varpi)})^2} p_{1,q}. \tag{4.40}$$

Likewise,

$$\text{LHS of (4.39)} \leq \Lambda_q^{(\varpi)}(p) = \frac{\Lambda_{q,\mathrm{NUM}}^{(\varpi)}(p)}{\Lambda_{q,\mathrm{DEN}}^{(\varpi)}(p)} \triangleq \frac{\left(\sum_{l=1}^{L} \frac{\gamma_{l,q}}{\eta_l(p)}\right)^2}{\frac{2}{Mp_{2,q}^{(\varpi)}} - \frac{p_{2,q}}{M(p_{2,q}^{(\varpi)})^2} + \sum_{l=1}^{L} \gamma_{l,2,q} \left(\frac{2}{\eta_l(p^{(\varpi)})} - \frac{\eta_l(p)}{(\eta_l(p^{(\varpi)}))^2}\right)} \tag{4.41}$$

under the trust region

$$\Lambda_{q,\mathrm{DEN}}^{(\varpi)}(p) > 0. \tag{4.42}$$

The function $\Lambda^{(\varpi)}$ is convex over the trust region (4.42). Thus, the nonconvex constraint (4.39) is innerly approximated by the following convex constraint

$$\Lambda_q^{(\varpi)}(p) \leq \frac{2x_q^{(\varpi)}}{p_{1,q}^{(\varpi)}} x_q - \frac{(x_q^{(\varpi)})^2}{(p_{1,q}^{(\varpi)})^2} p_{1,q}. \tag{4.43}$$

Initialized by a feasible $p^{(0)}$ for the power constraint (4.33b) and $x_q^{(0)} = \sqrt{\bar{R}_{2,q}(p^{(0)})}$, $q = 1, \ldots, K$, at the ϖth iteration we solve the following convex optimization problem to generate the next feasible point $(p^{(\varpi+1)}, x^{(\varpi+1)})$ for (4.34):

$$\begin{aligned} &\max_{p, x=(x_1,\ldots,x_K)} \min_{q=1,\ldots,K} \min\{R_{1,q}^{(\varpi)}(p, x), R_{2,q}^{(\varpi)}(p)\} \\ &\text{subject to} \quad (4.33b), (4.42), (4.43), \end{aligned} \tag{4.44}$$

which is equivalently expressed as in Eqn. (4.45).

$$\begin{aligned} \max_{p,x,y=(y_1,\ldots,L)} \varphi^{(\varpi)}(p,x) &\triangleq \min_{q=1,\ldots,K} \min\{R_{1,q}^{(\varpi)}(p,x), R_{2,q}^{(\varpi)}(p)\} \\ \text{subject to}\quad & 0 \leq p_{g,q} \leq p_{\max}, \forall p,q \\ & \begin{bmatrix} y_l & 1 \\ 1 & \eta_l(p) \end{bmatrix} \succeq 0, \ell = 1,\ldots,L, \\ & \begin{bmatrix} \frac{2x_q^{(\varpi)}}{p_{1,q}^{(\varpi)}} x_q - \frac{\left(x_q^{(\varpi)}\right)^2}{(p_{1,q}^{(\varpi)})^2} p_{1,q}. & \sum_{l=1}^{L} \gamma_{l,q} y_l \\ \sum_{l=1}^{L} \gamma_{l,q} y_l & \Lambda_{q,\mathrm{DEN}}^{(\varpi)}(p) \end{bmatrix} \succeq 0. \end{aligned} \tag{4.45}$$

The above optimization problem can be easily solved by convex optimization tools such as CVX. Note that $\varphi^{(\varpi)}(p^{(\varpi+1)}, x^{(\varpi+1)}) > \varphi^{(\varpi)}(p^{(\varpi)}, x^{(\varpi)})$ as far as $(p^{(\varpi+1)}, x^{(\varpi+1)}) \neq (p^{(\varpi)}, x^{(\varpi)})$ because $(p^{(\varpi+1)}, x^{(\varpi+1)})$ is the optimal solution of (4.44) but $(p^{(\varpi)}, x^{(\varpi)})$ is only its feasible point. Therefore we have

$$\varphi(p^{(\varpi+1)}) \ \ \geq \ \ \varphi^{(\varpi)}(p^{(\varpi+1)}, x^{(\varpi+1)}) > \varphi^{(\varpi)}(p^{(\varpi)}, x^{(\varpi)}) = \varphi(p^{(\varpi)}),$$

i.e., the sequence $\{p^{(\varpi)}\}$ is of improved feasible points for the nonconvex problem (4.34) and as such, it converges at least to a locally optimal solution of (4.34), which satisfies the Karush-Kuh-Tucker optimality condition [43].

Remark 3: By focusing on the case that SIC is applied for the second user in each group, it is implicitly assumed that the decoding order in each group has been determined. In practice, this is an important step and could strongly affect the system's performance. Here we propose to determine the decoding order for SIC implementation by first solving the power control problem for the non-SIC case, i.e., by replacing $R_{2,q}^{(\mathrm{SIC-OBC})}$ with $R_{2,q}^{(\mathrm{OBC})}$ in Eqn. (4.32). Once such a power control problem is solved, we can choose the user having a higher transmit power in each group as the worse user to perform SIC (i.e., it is the second user as in the analysis in Section 4.2), whereas the other user having a lower transmit power is treated as a better user. With such a decoding order in each group, the worse user is granted a more favorable signal detection as compared to the better user. This enhances fairness in the system and therefore, a better max-min QoS can be achieved.

4.7 Simulation Results

Although all the numerical expressions (except the asymptotic analysis) and the power control problem are valid for a general case of finite L (number of BSs or APs) and M (number of antennas on each BS), all the results are presented for the cell-free cooperative MIMO setup, i.e., when L is small and M is large. The only exception is the last figure, which presents results for different combinations of M and L while $M \times L$ is fixed. The massive MIMO system considered in the simulation consists of L multi-antenna BSs and $2K$ users. In each iteration, locations of users are randomly generated within the co-coverage area of all BSs (a 400m × 400m square) and all numerical results are averaged over 300 iterations. The large-scale fading coefficients are molded according to the 3GPP LTE standard. In particular, the large scale fading is defined as $\beta_{l,g,k} = -131 - 42.8\log10 d_{l,g,k} + z_{l,g,k}\text{dB}$, where $d_{l,g,k}$ is the distance from the lth BS to the kth user of the gth group and $z_{l,g,k}$ is the standard deviation of the shadowing variable. The noise figure of 5dB translates to a noise variance of −96dBm. The simulation parameters are summarized in Table 4.1. The QoS performance metric used to compare different backhaul combining methods is the tight lower bound of the UL SE, which is directly related to the SINR. In all simulation scenarios, the number of pilot sequences used for NOMA is exactly half of that used for OMA.

Table 4.1 Simulation parameters.

Parameter	Value
Peak UL transmit power	23 dBm
Shadowing standard deviation	10 dB
Penetration loss (indoor users)	20 dB
Noise figure	5 dB
Coherence interval	100 symbols
Pathloss	$131 + 42.8\log10 d$

Except for Fig. 4.2, all the results presented in this section are obtained with max-min QoS optimization of power control[2]. Without a proper power control, it is not possible to

[2] The power optimization problems for EBC and max-SNR association are quasi-convex problems, which

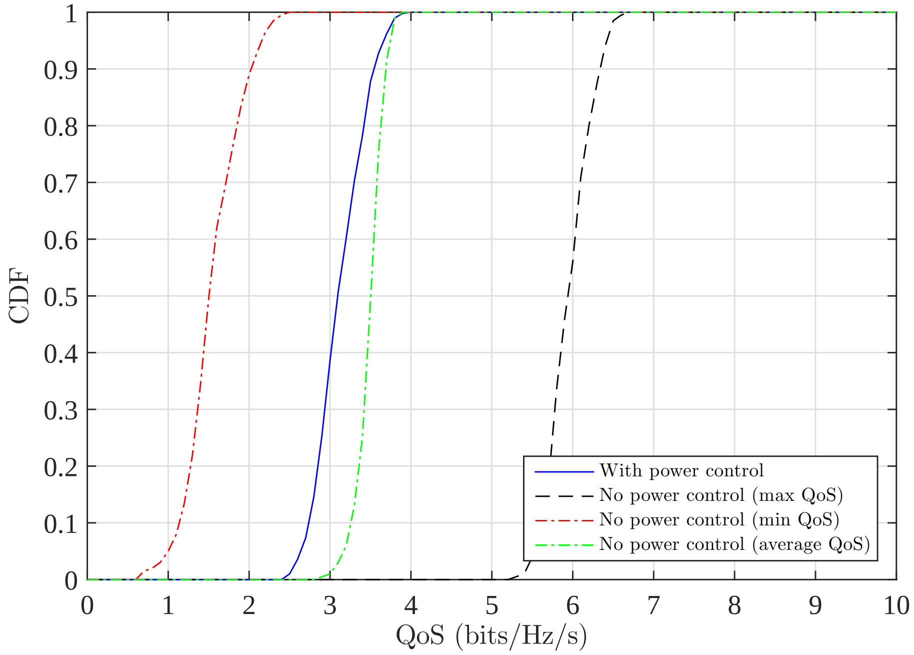


Figure 4.2 Cumulative distribution function of UL SE with and without power control (3 BSs, each having 300 antennas).

ensure that all users are equally served with a predetermined QoS value. An illustration on the effect of power control can be seen in Fig. 4.2, where the performance obtained with OBC is compared between the two cases of max-min QoS power control and equal power allocation. As expected, with equal transmit powers, some users enjoy very good performance, while others severely suffer from interference, which means very low QoS.

Fig. 4.3 shows the cumulative distribution functions (CDFs) of UL SE achieved with NOMA using different combining methods when the network has 3 BSs, each having $M = 300$ antennas, and 30 users randomly grouped in 15 pairs. For the results in this figure, SIC is implemented to subtract the correlated interference from the first user before detecting the second user in every group. As can be seen, OBC is always better than EBC and max-SNR association. The average value of the UL SE achieved with OBC is around 3.1 bits/s/Hz, which is about 30% higher than that achieved by EBC (2.6 bits/s/Hz) and almost twice the

can be solved by the bi-section method [14]. On the other hand, the max-min QoS power control problem for ZFBC can be solved similarly as with the proposed power control algorithm for OBC in Section 4.6.

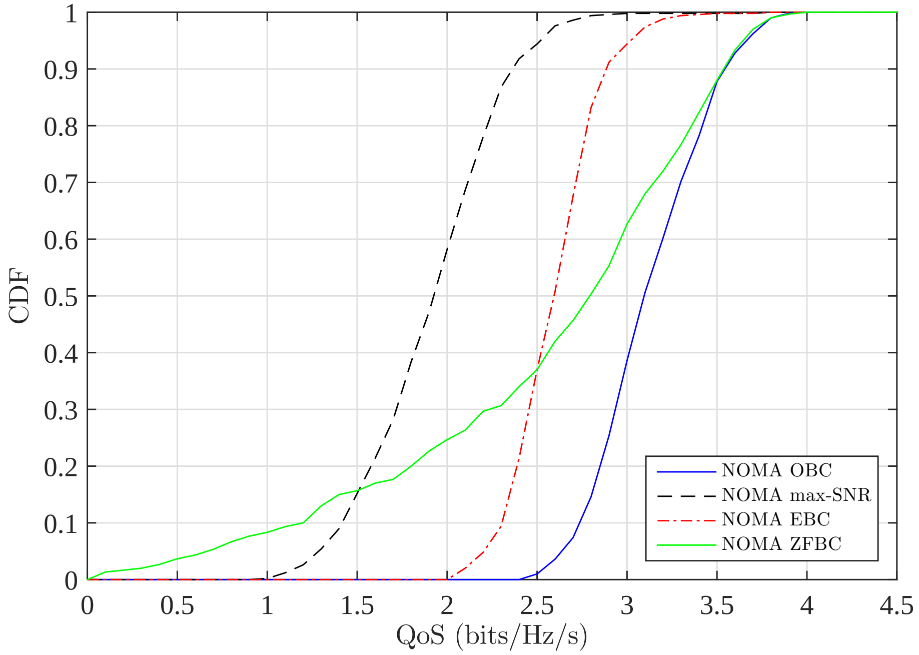


Figure 4.3 Cumulative distribution functions of UL SE for three different backhaul combining methods (3 BSs, each having 300 antennas).

value of the max-SNR method (1.8 bits/s/Hz). With a favorable channel condition, OBC could achieve up to about 4 bits/s/Hz in UL SE. The ZFBC method can also achieve peak performance similar to that of OBC. However, the UL SE of ZFBC is widely distributed between 0 to 4 bits/Hz/s. The reason is that ZFBC ignores uncorrelated noise-and-interference, hence its performance is poor in the low SINR, while it can achieve similar performance as that of OBC in the high SINR.

The effect of user grouping and SIC is illustrated in Fig. 4.4 for 10 users and with three BSs, each having $M = 300$ antennas. It is clear from the figure that the proposed user grouping method yields a significant improvement in UL SE when compared to random user grouping. The average UL SE obtained with OBC and the proposed grouping is approximately 5.6 bits/s/Hz, which is 12% higher than 5 bits/s/Hz obtained with OBC and random user grouping. Although Section 4.2 shows that the SINR achieved with OBC and without SIC is not limited by the impact of correlated interference when the number of antennas tends to infinity, applying SIC yields a noticeable improvement. It can be seen from the

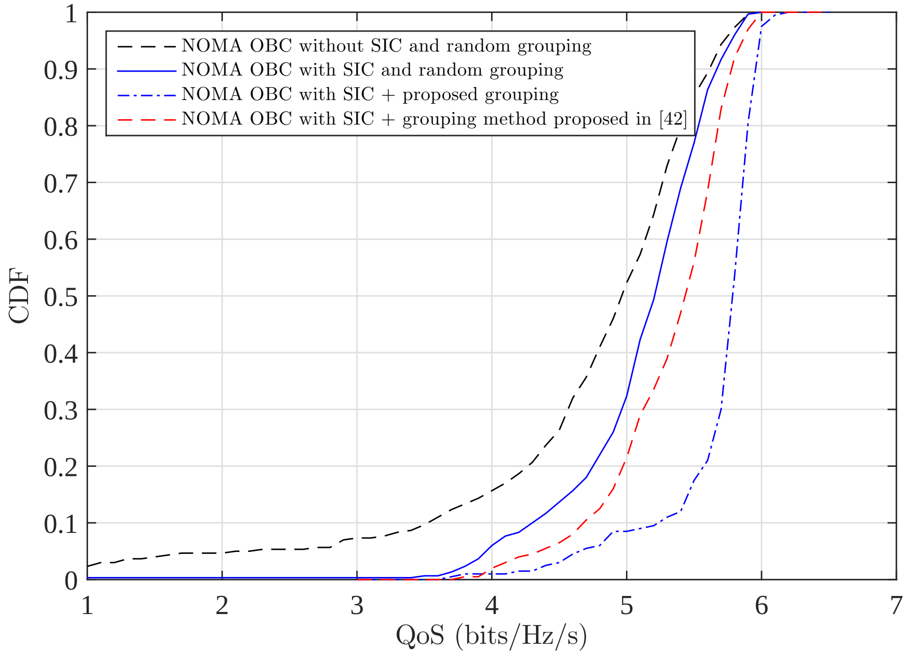


Figure 4.4 Cumulative distribution function of UL SE with and without grouping (3 BSs, each having 300 antennas).

figure that the CDF obtained with SIC is consistently better than the CDF obtained without SIC. Moreover, while the UL SE obtained without SIC varies widely around its median value and can be sometimes below 1 bit/Hz/s, the UL SE obtained with SIC is almost always greater than 3.5 bits/s/Hz and concentrates more around the median value. The figure also shows that our proposed grouping method performs better than the method in [42]. However, the tradeoff is the higher complexity since the method in [42] needs to calculate only $2K$ coefficients, whereas our method requires to calculate $(2K)^2$ coefficients.

Given the advantages of the proposed user grouping and SIC, these two methods are always implemented to obtain the results presented in the remaining figures of this section. Fig. 4.5 compares the average UL SE obtained with OMA and NOMA that can be equally served to 30 users in the network versus the number of antennas in each BS. With OMA, the number of pilot sequences required is $\tau_p = 30$, which is equal to the number of users. In contrast, NOMA only needs half the number ($\tau_p = 15$) of pilot sequences, which saves more symbol times for data transmission. For both OMA and NOMA, OBC provides the best

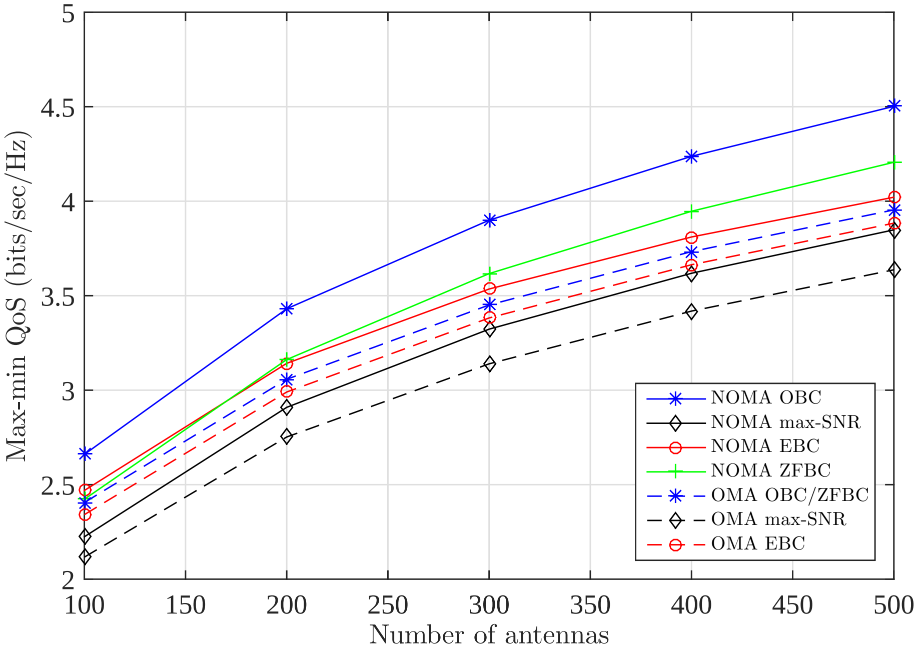


Figure 4.5 Max-min QoS versus the number of antennas (30 users, 3 BSs).

performance, followed by EBC and max-SNR schemes. In addition, the performance gaps among different combining methods under NOMA are noticeably greater than that under OMA. For OMA, the performance gap between OBC and EBC is consistently 0.1 bit/Hz/s, whereas it is 0.4 bits/s/Hz between OBC and max-SNR association. With NOMA, the slope of the performance curve obtained with OBC is much sharper than the slopes of other curves. Compared to EBC and max-SNR association, the performance gains provided by OBC are about 0.5 bit/Hz/s and 1 bit/Hz/s, respectively. The performance gain also increases with the number of antennas. This is expected since the SINR obtained with OBC increases without a bound with increasing number of antennas, whereas the SINRs obtained with both EBC and max-SNR methods are saturated when the number of antennas tends to infinity as analyzed in Section 4.2. The performance of ZFBC is eventually better than that of EBC and max-SNR association when the number of antennas is large enough. However, it is always worse than the performance of OBC. Also note that in the case of OMA, ZFBC and OBC are the same, and so are their performance curves.

Since for both NOMA and OMA, OBC always yields the best performance, from now

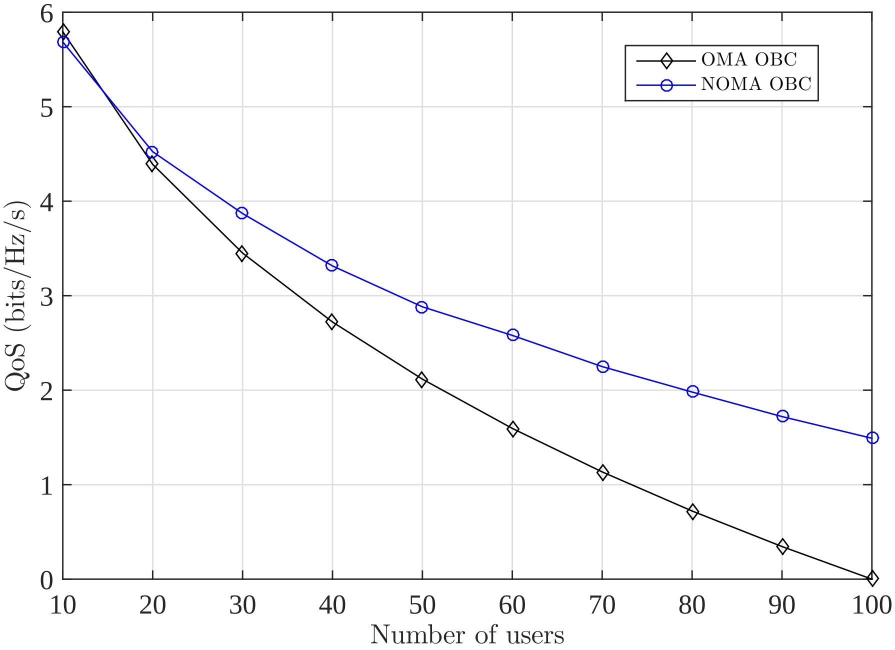


Figure 4.6 Max-min QoS versus the number of users (3 BSs, each having 300 antennas).

on, we only consider OBC at the backhaul network. In Fig. 4.6, the max-min QoS is plotted versus the number of users with three BSs, each having 300 antennas. Obviously, when there are more users in the network, the max-min QoS value gets smaller. When the number of users increases to about 20, NOMA starts to outperform OMA, which is because the effect of using less symbol times for pilot sequences starts to kick in. The performance gap between the two methods widens when the number of users increases.

Fig. 4.7 shows the CDFs of UL SE for different numbers of participating BSs with $M = 300$ antennas at each BS and 30 users. As expected, the system's performance is enhanced when there are more BSs serving each user. The average UL SE is approximately 2.7, 3.9 and 4.2 bits/s/Hz when there are 2, 3 and 4 BSs, respectively. Compared to OMA in terms of average performance, NOMA performs poorer if there are only 2 BSs, but better when there are 3 or 4 BSs. Moreover, the CDF curves for the cases of having 3 and 4 BSs vary less and concentrate more around the median values as opposed to the CDF for the case of having 2 BSs.

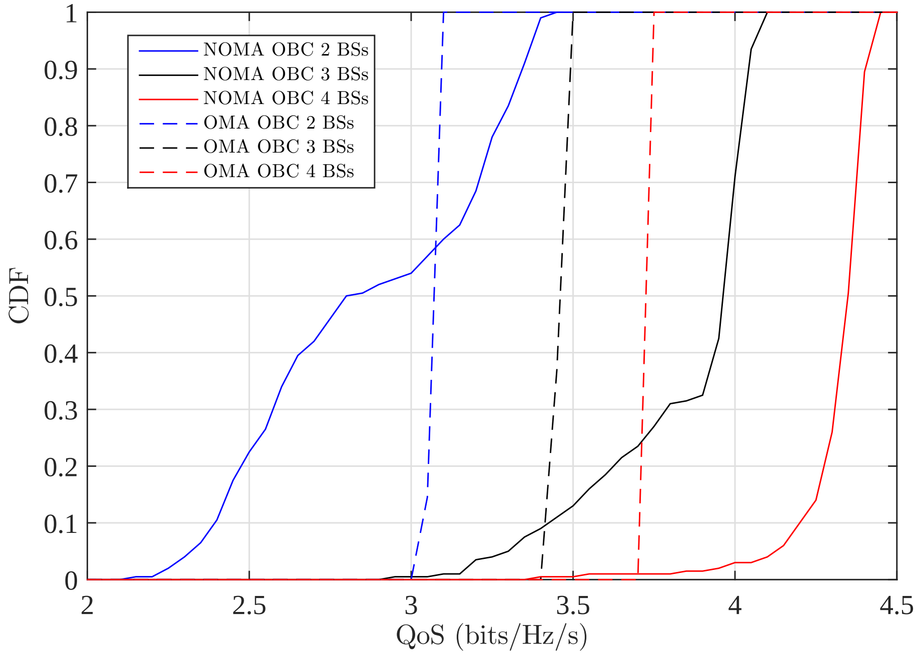


Figure 4.7 Cumulative distribution function of UL SE versus the number of serving BSs (30 users, 300 antennas at each BS).

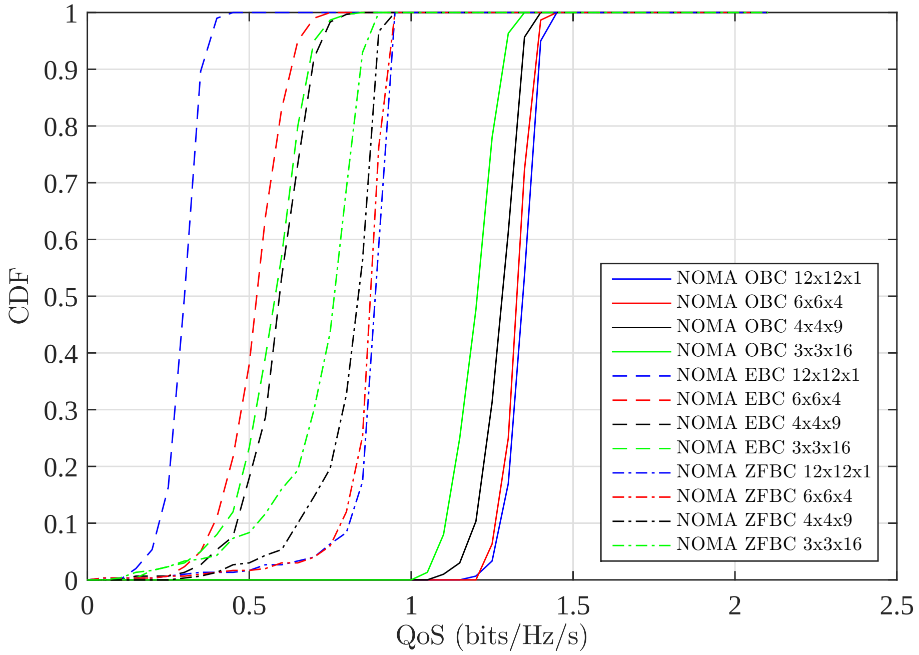


Figure 4.8 Cumulative distribution functions of UL SE for a cell-free massive MIMO system: $M \times L = 144$ and 50 users.

Finally, Fig. 4.8 compares the CDFs of QoS obtained with OBC and EBC with a fixed total number of antennas, namely $M \times L = 144$. The BSs are equally distanced in grid sizes 12×12, 6×6, 4×4 and 3×3 with the numbers of antennas being 1, 4, 9, 16, respectively. It can be seen that the performance with EBC deteriorates when L increases and M decreases. This is because the signal after the first stage of combining (at each BS) contains mostly noise and uncorrelated interference as the channel hardening property is weak with a small number of antennas. When M grows and L decreases, noise and uncorrelated interference declines, which is consistent with the results for EBC in Fig. 4.8 when M increases from 1 to 9. However, when M increases, not only the desired signal but the correlated interference also grows and at some value of M, it surpasses noise and uncorrelated interference and becomes dominant. At that point, increasing M further may cause the SINR to decline and gradually converge to the saturation value as analyzed with large M in Section 4.4. Unlike EBC, with ZFBC and OBC, when M is small, the amount of correlated interference is insignificant as compared to noise and uncorrelated interference, and hence can be ignored. When M increases, correlated interference becomes dominant, which causes more degradation to the SINR as can be seen from the SINR expressions for OBC and ZFBC in (4.21) and (4.27), respectively. That explains why OBC and ZFBC are better in low M and large L regime. Another reason is that when deploying more BSs in a fixed area, the distances between a user and BSs become shorter, which leads to better channel conditions.

4.8 Conclusions

This paper has considered the uplink transmission in a generalized cell-free massive MIMO-NOMA system and developed an optimal combining method for the signals forwarded by multiple BSs to the CPU over the backhaul network. The proposed combining method has been shown to provide unlimited uplink SINR when the number of antennas at each BS tends to infinity, despite the existence of pilot contamination originating from sharing pilot sequences. To optimize system performance, a max-min QoS power control problem was formulated and solved in which a desired QoS value that can be equally served to all users is maximized, subject to a maximum transmit power for every user. Because of the non-convexity of the problem, an inner approximation method is developed to convert it into a series of convex optimization, which can be solved iteratively. In addition, to further improve the achievable max-min QoS, a user grouping algorithm is introduced and shown to be better than random user grouping and a user grouping method recently proposed in the literature. Numerical results were presented to corroborate the analysis and demonstrate that the proposed optimal backhaul combining method outperforms both equal-gain combining and zero-forcing combining. Moreover, simulation results also show that, by using the proposed optimal backhaul combining, cell-free massive MIMO-NOMA is superior than cell-free massive MIMO-OMA in both max-min QoS and connectivity.

4.A Examination of correlation property

This appendix examines correlation properties of different signal components in Eqn. (4.8). First, the channel gain uncertainty $\mathrm{CU}_{l,1,q}$ is uncorrelated interference since:

$$\begin{aligned}\operatorname{cov}\left\{\mathrm{CU}_{l,1,q},\mathrm{CU}_{l',1,q}\right\} &= p_{1,q}\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}(\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{h}_{l',1,q})\right\}\\ &- p_{1,q}\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}(\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{h}_{l',1,q})\right\}\\ &- p_{1,q}\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\mathbb{E}\left\{\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{h}_{l',1,q}\right\}\right\}\\ &+ p_{1,q}\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}\mathbb{E}\left\{\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{h}_{l',1,q}\right\} = 0,\end{aligned} \tag{4.46}$$

due to the fact that $\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}$ and $\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{h}_{l',1,q}$ are independent when $l \neq l'$. Similarly, it is easy to verify that:

$$\operatorname{cov}\left\{\mathrm{IoG}_{l,1,q},\mathrm{IoG}_{l',1,q}\right\} = \sum_{g=1}^{2}\sum_{k=1,k\neq q}^{K} p_{g,k}\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,g,k}(\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{h}_{l',g,k})\right\} = 0, \tag{4.47}$$

and

$$\operatorname{cov}\left\{\mathrm{N}_{l,1,q},\mathrm{N}_{l',1,q}\right\} = \mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{n}_{l}(\boldsymbol{v}_{l',1,q}^{\mathsf{H}}\boldsymbol{n}_{l'})\right\} = 0. \tag{4.48}$$

Performing the same analysis for other terms of Eqn. (4.9), on arrives at the conclusion that the only correlated interference term is $c_{l,1,q}x_{2,q}$.

4.B Proof of Theorem 1: Optimal weight combining vector

The proof can be carried out by applying the Schwartz inequality [44]. With the signal model in Eqn. (4.7), the SINR at the BS after combining signals with the OBC vector $\boldsymbol{w}_{1,q}$ is:

$$\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} = \frac{\left|\boldsymbol{w}_{1,q}^{\mathsf{T}}\boldsymbol{s}_{1,q}\right|^2}{\mathbb{E}\left\{\left|\boldsymbol{w}_{1,q}^{\mathsf{T}}(\boldsymbol{\kappa}_{1,q} - \boldsymbol{s}_{1,q}x_{1,q})\right|^2\right\}} = \frac{\left|\boldsymbol{w}_{1,q}^{\mathsf{T}}\boldsymbol{s}_{1,q}\right|^2}{\boldsymbol{w}_{1,q}^{\mathsf{H}}\mathbf{R}_{1,q}\boldsymbol{w}_{1,q}}. \tag{4.49}$$

By the definition in Eqn. (4.19), $\mathbf{R}_{1,q}$ is a positive-definite Hermitian matrix. Therefore, it can be rewritten via a unitary decomposition as follows:

$$\mathbf{R}_{1,q} = \boldsymbol{\Psi}^{\mathsf{H}}\mathbf{D}_{1,q}^2\boldsymbol{\Psi} \tag{4.50}$$

where $\mathbf{D}_{1,q}$ is a diagonal matrix whose elements are the square roots of the eigenvalues of $\mathbf{R}_{1,q}$ and $\boldsymbol{\Psi}$ is a unitary matrix. As a result, the SINR with OBC is:

$$\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} = \frac{\left|\boldsymbol{w}_{1,q}^{\mathsf{T}}\left(\mathbf{D}\boldsymbol{\Psi}\right)^{\mathsf{T}}\left[\left(\mathbf{D}\boldsymbol{\Psi}\right)^{\mathsf{T}}\right]^{-1}\boldsymbol{s}_{1,q}\right|^2}{\boldsymbol{w}_{1,q}^{\mathsf{H}}\boldsymbol{\Psi}^{\mathsf{H}}\mathbf{D}_{1,q}^2\boldsymbol{\Psi}\boldsymbol{w}_{1,q}} = \frac{\left|\left(\mathbf{D}\boldsymbol{\Psi}\boldsymbol{w}_{1,q}\right)^{\mathsf{T}}\left[\left(\mathbf{D}\boldsymbol{\Psi}\right)^{\mathsf{T}}\right]^{-1}\boldsymbol{s}_{1,q}\right|^2}{\left\|\mathbf{D}\boldsymbol{\Psi}\boldsymbol{w}_{1,q}\right\|^2}. \tag{4.51}$$

Applying the Schwartz inequality leads to:

$$\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} \leq \frac{\left\|\mathbf{D}\boldsymbol{\Psi}\boldsymbol{w}_{1,q}\right\|^2\left\|\left[\left(\mathbf{D}\boldsymbol{\Psi}\right)^{\mathsf{T}}\right]^{-1}\boldsymbol{s}_{1,q}\right\|^2}{\left\|\mathbf{D}\boldsymbol{\Psi}\boldsymbol{w}_{1,q}\right\|^2} = \left\|\left[\left(\mathbf{D}\boldsymbol{\Psi}\right)^{\mathsf{T}}\right]^{-1}\boldsymbol{s}_{1,q}\right\|^2 = \boldsymbol{s}_{1,q}^{\mathsf{H}}\mathbf{R}_{1,q}^{-1}\boldsymbol{s}_{1,q}. \tag{4.52}$$

The equality holds when $\mathbf{D}\boldsymbol{\Psi}\boldsymbol{w}_{1,q} = \alpha\left[\left(\mathbf{D}\boldsymbol{\Psi}\right)^{\mathsf{T}}\right]^{-1}\boldsymbol{s}_{1,q}$, which is equivalent to the backhaul combining vector $\boldsymbol{w}_{1,q} = \alpha\mathbf{R}_{1,q}^{-1}\boldsymbol{s}_{1,q}$, where α is a constant.

4.C Proof of Theorem 1: SINR with OBC

The normalized interference-plus-noise covariance matrix can be rewritten as:

$$\hat{\mathbf{R}}_{1,q} = \mathbb{E}\left\{\hat{\boldsymbol{u}}_{1,q}\hat{\boldsymbol{u}}_{1,q}^{\mathsf{H}} + \hat{\boldsymbol{c}}_{1,q}\hat{\boldsymbol{c}}_{1,q}^{\mathsf{H}}\right\} = \mathbf{I}_L + \hat{\boldsymbol{c}}_{1,q}\hat{\boldsymbol{c}}_{1,q}^{\mathsf{H}}. \tag{4.53}$$

Multiplying this matrix with $\hat{\boldsymbol{c}}_{1,q}$ leads to:

$$\hat{\mathbf{R}}_{1,q}\hat{\boldsymbol{c}}_{1,q} = \hat{\boldsymbol{c}}_{1,q} + \hat{\boldsymbol{c}}_{1,q} * \left\|\hat{\boldsymbol{c}}_{1,q}\right\|^2 = \left(1 + \left\|\hat{\boldsymbol{c}}_{1,q}\right\|^2\right)\hat{\boldsymbol{c}}_{1,q}. \tag{4.54}$$

The above equation means that $\hat{\boldsymbol{c}}_{1,q}$ is an eigenvector of $\hat{\mathbf{R}}_{1,q}$ with the corresponding eigenvalue:

$$\lambda_{1,1,q} = 1 + \|\hat{\boldsymbol{c}}_{1,q}\|^2 . \tag{4.55}$$

Any other vectors which are orthogonal to $\hat{\boldsymbol{c}}_{1,q}$ are also eigenvectors with an unit eigenvalue, which means $\lambda_{l,1,q} = 1, (l = 2, \ldots, L)$. By a unitary decomposition of $\hat{\mathbf{R}}_{1,q}^{-1}$, the SINR in Eqn. (4.21) can be rewritten as:

$$\begin{aligned}\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} &= \hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\boldsymbol{\Psi}^{\mathsf{H}}\mathrm{diag}(\lambda_{1,1,q}^{-1}, \ldots, \lambda_{L,1,q}^{-1})\boldsymbol{\Psi}\hat{\boldsymbol{s}}_{1,q} \\ &= \hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\boldsymbol{\Psi}^{\mathsf{H}}\mathbf{I}_L\boldsymbol{\Psi}\hat{\boldsymbol{s}}_{1,q} - \hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\boldsymbol{\Psi}^{\mathsf{H}}\mathrm{diag}(1-\lambda_{1,1,q}^{-1}, 0, \ldots, 0)\boldsymbol{\Psi}\hat{\boldsymbol{s}}_{1,q} \\ &= \|\hat{\boldsymbol{s}}_{1,q}\|^2 - \hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\boldsymbol{\Psi}^{\mathsf{H}}\mathrm{diag}(1-\lambda_{1,1,q}^{-1}, 0, \ldots, 0)\boldsymbol{\Psi}\hat{\boldsymbol{s}}_{1,q},\end{aligned} \tag{4.56}$$

where $\boldsymbol{\Psi}$ is a unitary matrix. It can be easily seen that the first column $\boldsymbol{\psi}_1$ of $\boldsymbol{\Psi}$ corresponding to $\lambda_{1,1,q}$ is also the eigenvector of $\hat{\mathbf{R}}_{1,q}$ corresponding to eigenvalue $\lambda_{1,1,q}$. With this property, the SINR of the first user in the qth group is simplified to:

$$\begin{aligned}\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} &= \|\hat{\boldsymbol{s}}_{1,q}\|^2 - \left|\boldsymbol{\psi}_1^{\mathsf{H}}\hat{\boldsymbol{s}}_{1,q}\right|^2\left(1-\lambda_{1,1,q}^{-1}\right) \\ &= \|\hat{\boldsymbol{s}}_{1,q}\|^2 - \frac{\left|\hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q}\right|^2}{\|\hat{\boldsymbol{c}}_{1,q}\|^2}\left(1 - \frac{1}{1+\|\hat{\boldsymbol{c}}_{1,q}\|^2}\right) \\ &= \|\hat{\boldsymbol{s}}_{1,q}\|^2 - \frac{\left|\hat{\boldsymbol{s}}_{1,q}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q}\right|^2}{1+\|\hat{\boldsymbol{c}}_{1,q}\|^2} \\ &= \sum_{l=1}^{L}\mathrm{SNR}_{l,1,q} - \frac{\left(\sum_{l=1}^{L}\sqrt{\mathrm{SNR}_{l,1,q}}\sqrt{\mathrm{SNR}_{l,2,q}}\right)^2}{1+\sum_{l=1}^{L}\mathrm{SNR}_{l,2,q}}.\end{aligned} \tag{4.57}$$

4.D Derivation for the SINR of EBC

With the signal decomposition in Eqn. (4.8), the desired signal power equally combined at the CPU is computed as:

$$\begin{aligned}\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathrm{DS}_{l,1,q}\right|^2\right\} &= p_{1,q}\left|\left(\sum_{l=1}^{L}\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}\right)^2\right| \\ = p_{1,q}\left|\left(\sum_{l=1}^{L}\mathbb{E}\left\{\|\hat{\boldsymbol{h}}_{l,1,q}\|^2\right\}\right)^2\right| &= p_{1,q}\left(\sum_{l=1}^{L}\gamma_{l,1,q}M\right)^2 .\end{aligned} \tag{4.58}$$

The power of the channel gain uncertainty, $\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathrm{CU}_{l,1,q}\right|^2\right\}$, is obtained using the same method in [31] as:

$$\begin{aligned}
& p_{1,q}\mathbb{E}\left\{\left|\sum_{l=1}^{L}\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q} - \sum_{l=1}^{L}\mathbb{E}\left\{\boldsymbol{v}_{l,1,q}^{\mathsf{H}}\boldsymbol{h}_{l,1,q}\right\}\right|^2\right\} \\
& = p_{1,q}\left(\sum_{l=1}^{L}\gamma_{l,1,q}^2(M+M^2) + \sum_{l=1}^{L}\gamma_{l,1,q}\left(\beta_{l,1,q}-\gamma_{l,1,q}\right)M\right) \\
& - p_{1,q}\left(\sum_{l=1}^{L}\gamma_{l,1,q}M\right)^2 \le p_{1,q}\sum_{l=1}^{L}\gamma_{l,1,q}\beta_{l,1,q}M.
\end{aligned} \tag{4.59}$$

The interference which is caused by the second user of the qth group can be decomposed as in Eqn. (4.9). As a result, the power of the interference from the user using the same pilot in the group is:

$$\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathrm{IwG}_{l,1,q}\right|^2\right\} = p_{2,q}\sum_{l=1}^{L}\gamma_{l,1,q}\beta_{l,2,q}M + p_{2,q}\frac{p_{2,q}^{(\mathrm{p})}}{p_{1,q}^{(\mathrm{p})}}\left(\sum_{l=1}^{L}\gamma_{l,1,q}\frac{\beta_{l,2,q}}{\beta_{l,1,q}}\right)^2 M^2. \tag{4.60}$$

Due to the fact that the channels of different users are mutually independent and $\mathrm{IoG}_{l,1,q}$ is uncorrelated interference, the total power of $\mathrm{IoG}_{l,1,q}$ after using EBC is:

$$\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathrm{IoG}_{l,1,q}\right|^2\right\} = \sum_{l=1}^{L}\sum_{g=1}^{2}\sum_{\substack{k=1,\\k\neq q}}^{K}p_{g,k}\beta_{l,g,k}\gamma_{l,1,q}. \tag{4.61}$$

Finally, the power of the noise term is easily calculated as:

$$\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathrm{N}_{l,1,q}\right|^2\right\} = \sigma_{\mathrm{UL}}^2\sum_{l=1}^{L}\gamma_{l,1,q}. \tag{4.62}$$

Diving the power of the desired signal, $\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathrm{DS}_{l,1,q}\right|^2\right\}$, by the sum of the powers of remaining terms, we obtain the SINR expression as in Eqn. (4.25).

4.E Derivation for the SINR of ZFBC

From Eqn. (4.15), the SINR with ZFBC can be calculated as:

$$\mathrm{SINR}_{1,q}^{(\mathrm{ZFBC})} = \frac{\left|\left(\boldsymbol{w}_{1,q}^{(\mathrm{ZFBC})}\right)^{\mathsf{T}} \hat{\boldsymbol{s}}_{1,q}\right|^2}{\mathbb{E}\left\{\left|\left(\boldsymbol{w}_{1,q}^{(\mathrm{ZFBC})}\right)^{\mathsf{T}} \left(\hat{\boldsymbol{\kappa}}_{1,q} - \hat{\boldsymbol{s}}_{1,q} x_{1,q}\right)\right|^2\right\}}. \tag{4.63}$$

With the property of ZFBC, we have $\left|\left(\boldsymbol{w}_{1,q}^{(\mathrm{ZFBC})}\right)^{\mathsf{T}} \hat{\boldsymbol{s}}_{1,q}\right|^2 = 1$ and $\left|\left(\boldsymbol{w}_{1,q}^{(\mathrm{ZFBC})}\right)^{\mathsf{T}} \hat{\boldsymbol{c}}_{1,q}\right|^2 = 0$. Substituting into Eqn. (4.63) results in:

$$\mathrm{SINR}_{1,q}^{(\mathrm{ZFBC})} = \frac{1}{\left|\left(\boldsymbol{w}_{1,q}^{(\mathrm{ZFBC})}\right)^{\mathsf{T}} \hat{\boldsymbol{u}}_{1,q}\right|^2} = \frac{1}{\left\|\boldsymbol{w}_{1,q}^{(\mathrm{ZFBC})}\right\|^2} = \frac{1}{\left[\left(\Theta^{\mathsf{H}}\Theta\right)^{-1}\right]_{1,1}}. \tag{4.64}$$

Since:

$$\Theta^{\mathsf{H}}\Theta = \begin{bmatrix} \|\hat{\boldsymbol{s}}_{1,q}\|^2 & \left|\hat{\boldsymbol{s}}_{1,q}^{\mathsf{T}}\hat{\boldsymbol{c}}_{1,q}\right| \\ \left|\hat{\boldsymbol{s}}_{1,q}^{\mathsf{T}} c_{1,q}\right| & \|\hat{\boldsymbol{c}}_{1,q}\|^2, \end{bmatrix} \tag{4.65}$$

the element $\left[\left(\Theta^{\mathsf{H}}\Theta\right)^{-1}\right]_{1,1}$ can be computed as:

$$\left[\left(\Theta^{\mathsf{H}}\Theta\right)^{-1}\right]_{1,1} = \frac{1}{\det\left\{\Theta^{\mathsf{H}}\Theta\right\}}\|\hat{\boldsymbol{c}}_{1,q}\|^2 = \frac{\|\hat{\boldsymbol{c}}_{1,q}\|^2}{\|\hat{\boldsymbol{s}}_{1,q}\|^2\|\hat{\boldsymbol{c}}_{1,q}\|^2 - \left|\hat{\boldsymbol{s}}_{1,q}^{\mathsf{T}}\hat{\boldsymbol{c}}_{1,q}\right|^2}. \tag{4.66}$$

Plugging Eqns. (4.16), (4.18) and (4.66) into (4.64), we have the result for the SINR of ZFBC as in (4.27). ■

4.F Examination on how many BSs should serve a user

This appendix examines the question that how many users should be grouped to share the same pilot sequence. Assuming that N users share one pilot, we have the interference-plus-noise covariance matrix of the first user in the qth group as

$$\hat{\mathbf{R}}_{1,q} = \mathbb{E}\left\{\hat{\boldsymbol{u}}_{1,q}\hat{\boldsymbol{u}}_{1,q}^{\mathsf{H}} + \sum_{n=2}^{N} \hat{\boldsymbol{c}}_{1,q,n}\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\right\} = \mathbf{I}_L + \sum_{n=2}^{N} \hat{\boldsymbol{c}}_{1,q,n}\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}, \tag{4.67}$$

where $\hat{\boldsymbol{c}}_{1,q,n}$ denotes the normalized gain of correlated noise originating from the nth user ($n \geq 1$).

Case 1: $N > L$

In order to analyze the behavior of UL SINR with OBC, the following Lemma is useful.

Lemma 2: Any eigenvectors of Eqn. (4.67) must have the form of:

$$\boldsymbol{a} = \sum_{n=2}^{N} \theta_n \hat{\boldsymbol{c}}_{1,q,n}. \tag{4.68}$$

Proof: If $\boldsymbol{a}$ is an eigenvector of $\hat{\mathbf{R}}_{1,q}$, it satisfies:

$$\hat{\mathbf{R}}_{1,q}\boldsymbol{a} = \boldsymbol{a} + \sum_{n=2}^{N} \hat{\boldsymbol{c}}_{1,q,n}\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\boldsymbol{a} = \lambda_{1,1,q}\boldsymbol{a}, \tag{4.69}$$

which is equivalent to:

$$\boldsymbol{a} = \frac{1}{\lambda_{1,1,q} - 1} \sum_{n=2}^{N} \hat{\boldsymbol{c}}_{1,q,n}\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\boldsymbol{a}. \tag{4.70}$$

By defining $\theta_n = \frac{1}{\lambda_{1,1,q}-1}\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\boldsymbol{a}$, we obtain the result as in Lemma 2. ■

Multiplying $\hat{\mathbf{R}}_{1,q}$ with its eigenvector leads to:

$$\begin{aligned}\hat{\mathbf{R}}_{1,q}\boldsymbol{a} &= \sum_{n=2}^{N}\theta_n\hat{\boldsymbol{c}}_{1,q,n} + \sum_{n=2}^{N}\theta_n\hat{\boldsymbol{c}}_{1,q,n}\left(\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q,n} + \sum_{n'=2,n'\neq n}^{N}\frac{\theta_{n'}}{\theta_n}\hat{\boldsymbol{c}}_{1,q,n'}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q,n}\right)\\ &= \lambda_{l,1,q}\sum_{n=2}^{N}\theta_n\hat{\boldsymbol{c}}_{1,q,n},\end{aligned} \tag{4.71}$$

where $\lambda_{l,1,q}$ is an eigenvalue of $\hat{\mathbf{R}}_{1,q}$. Then $\boldsymbol{a}$ is an eigenvector if and only if:

$$\hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q,n} + \sum_{n'=2,n'\neq n}^{N}\frac{\theta_{n'}}{\theta_n}\hat{\boldsymbol{c}}_{1,q,n'}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q,n} = \mathit{constant}, \forall n. \tag{4.72}$$

As a result, all L eigenvalues of $\hat{\mathbf{R}}_{1,q}$ must have the form of:

$$\lambda_{l,1,q} = 1 + \hat{\boldsymbol{c}}_{1,q,n}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q,n} + \sum_{n'=2,n'\neq n}^{N}\frac{\theta_{n'}}{\theta_n}\hat{\boldsymbol{c}}_{1,q,n'}^{\mathsf{H}}\hat{\boldsymbol{c}}_{1,q,n}, \tag{4.73}$$

The second term in the above equation has the value equal to the sum of SNRs from all L BSs of the nth user ($n \geq 1$) of the qth group, which is a source of correlated interference to the first user of the qth group. According to the analysis in the paper, this value is proportional to the number of antennas M. With these eigenvalues, the *instantaneous* SINR is defined by:

$$\mathrm{SINR}_{1,q}^{(\mathrm{OBC,inst})} = \hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst}),\mathsf{H}} \hat{\mathbf{R}}_{1,q}^{-1} \hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst})} = \hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst}),\mathsf{H}} \boldsymbol{\Psi}^{\mathsf{H}} \mathrm{diag}\left(\lambda_{1,1,q}^{-1}, \ldots, \lambda_{L,1,q}^{-1}\right) \boldsymbol{\Psi} \hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst})}, \tag{4.74}$$

where $\hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst})}$ represents the instantaneous gain vector corresponding to the normalized desired signal and $\boldsymbol{\Psi}$ is a unitary matrix obtained by the unitary transform of $\hat{\mathbf{R}}_{1,q}^{-1}$. Because multiplying with a unitary matrix does not change the distribution of a random vector, the expectation of Eqn. (4.74) can be computed as:

$$\mathbb{E}\left\{\mathrm{SINR}_{1,q}^{(\mathrm{OBC,inst})}\right\} = \mathbb{E}\left\{\hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst}),\mathsf{H}} \mathrm{diag}\left(\lambda_{1,1,q}^{-1}, \ldots, \lambda_{L,1,q}^{-1}\right) \hat{\boldsymbol{s}}_{1,q}^{(\mathrm{inst})}\right\} = \sum_{l=1}^{L} \frac{\left|\hat{s}_{l,1,q}^{(\mathrm{inst})}\right|^2}{\lambda_{l,1,q}}. \tag{4.75}$$

Due to the fact that both the desired signal power and the eigenvalues are proportional to the number of antennas M, all addends of (50) converge when M tends to infinity. This means that with $N > L$, the instantaneous SINR is saturated when M goes to infinity.

Case 2: $N \leq L$

The same analysis can be carried out for the case $N > L$. However, in this case, there are only $N - 1$ eigenvectors lying on the hyperplane defined by $\hat{\boldsymbol{c}}_{1,q,n}, \ldots, \hat{\boldsymbol{c}}_{1,q,N}$. The remaining eigenvectors which are orthogonal to this hyperplane are corresponding to the unit eigenvalue. As a result, at least one addend of Eqn. (4.75) is proportional to M, which means the instantaneous SINR in this case is not bounded when M goes to infinity. This result shows that in order to acquire the array gain in cell-free massive MIMO-NOMA with OBC, the number of BSs serving each user must be equal or greater than the number of users in each group (i.e., $N \leq L$).

4.G Fundamental inequalities for convex approximation

The following fundamental inequalities are used:

$$\frac{1}{x} \geq \frac{2}{\bar{x}} - \frac{x}{\bar{x}^2} \quad \forall\, x > 0, \bar{x} > 0. \tag{4.76}$$

$$xy \leq \frac{\bar{x}\bar{y}}{4}\left(\frac{x}{\bar{x}} + \frac{y}{\bar{y}}\right)^2 \quad \forall x > 0, y > 0, \bar{x} > 0, \bar{y} > 0. \tag{4.77}$$

$$\frac{x^2}{y} \geq \frac{2\bar{x}}{\bar{y}}x - \frac{\bar{x}^2}{\bar{y}^2}y \quad \forall\, x > 0, y > 0, \bar{x} > 0, \bar{y} > 0. \tag{4.78}$$

Note that the functions on the left-hand side (LHS) of (4.76) and (4.78) are convex while the functions on the right-hand side (RHS) are their first-order approximations. As such (4.76) and (4.78) follow from a standard condition of convex functions [45]. Lastly, (4.77) follows from a standard least-square.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?,” *IEEE J. Sel. Areas in Commun.*, vol. 32, pp. 1065–1082, June 2014.

[2] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, “A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends,” *IEEE J. Sel. Areas in Commun.*, vol. 35, pp. 2181–2195, Oct. 2017.

[3] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, “Achieving sustainable ultra-dense heterogeneous networks for 5G,” *IEEE Commun. Mag.*, vol. 55, pp. 84–90, Dec. 2017.

[4] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, and R. Fan, “Non-orthogonal multiple access: Achieving sustainable future radio access,” *IEEE Commun. Mag.*, vol. 57, pp. 116–121, Feb. 2019.

[5] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, “Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends,” *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, 2015.

[6] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, “Joint power and time allocation for NOMA-MEC offloading,” *IEEE Trans. Veh. Technol.*, vol. 68, pp. 6207–6211, June 2019.

[7] N. Ye, X. Li, H. Yu, L. Zhao, W. Liu, and X. Hou, “DeepNOMA: A unified framework for NOMA using deep multi-task learning,” *IEEE Trans. Wireless Commun.*, p. 1, 2020.

[8] H. D. Tuan, A. A. Nasir, H. H. Nguyen, T. Q. Duong, and H. V. Poor, “Non-Orthogonal Multiple Access With Improper Gaussian Signaling,” *IEEE J. Sel. Topics in Signal Process.*, vol. 13, no. 3, pp. 496–507, 2019.

[9] A. A. Nasir, H. D. Tuan, H. H. Nguyen, T. Q. Duong, and H. V. Poor, “Han-Kobayashi

signal superposition in noma with proper and improper Gaussian signaling," *Summitted to IEEE Trans. Wireless Commun.*, 2019.

[10] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2995–3005, 2019.

[11] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3590–3600, Nov. 2010.

[12] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 1293–1308, Feb. 2016.

[13] J. Zhang, L. Dai, X. Zhang, E. Bjrnson, and Z. Wang, "Achievable rate of rician large-scale MIMO channels with transceiver hardware impairments," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 8800–8806, Oct. 2016.

[14] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 1834–1850, Mar. 2017.

[15] T. M. Hoang, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and A. Marshall, "Cell-free massive MIMO networks: Optimal power control against active eavesdropping," *IEEE Trans. Commun.*, vol. 66, pp. 4724–4737, Oct. 2018.

[16] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max-min SINR of cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 2021–2036, Apr. 2019.

[17] T. C. Mai, H. Q. Ngo, M. Egan, and T. Q. Duong, "Pilot power control for cell-free massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 67, pp. 11264–11268, Nov. 2018.

[18] D. Maryopi, M. Bashar, and A. Burr, "On the uplink throughput of zero forcing in cell-free massive MIMO with coarse quantization," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 7220–7224, July 2019.

[19] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation in uplink mmwave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 3000–3004, Mar. 2019.

[20] K. Senel, H. V. Cheng, E. Bjrnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?," *IEEE J. Sel. Topics in Signal Process.*, vol. 13, pp. 597–611, June 2019.

[21] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, pp. 629–633, May 2016.

[22] X. Chen, F. Gong, G. Li, H. Zhang, and P. Song, "User pairing and pair scheduling in massive MIMO-NOMA systems," *IEEE Commun. Lett.*, vol. 22, pp. 788–791, Apr. 2018.

[23] C. Xu, Y. Hu, C. Liang, J. Ma, and L. Ping, "Massive MIMO, non-orthogonal multiple access and interleave division multiple access," *IEEE Access*, vol. 5, pp. 14728–14748, 2017.

[24] X. Liu, Y. Liu, X. Wang, and H. Lin, "Highly efficient 3-D resource allocation techniques in 5G for NOMA-enabled massive MIMO and relaying systems," *IEEE J. Sel. Areas in Commun.*, vol. 35, pp. 2785–2797, Dec. 2017.

[25] J. Ma, C. Liang, C. Xu, and L. Ping, "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems," *IEEE J. Sel. Areas in Commun.*, vol. 35, pp. 2696–2707, Dec. 2017.

[26] D. Kudathanthirige and G. A. A. Baduge, "NOMA-aided multicell downlink massive MIMO," *IEEE J. Sel. Topics in Signal Process.*, vol. 13, pp. 612–627, June 2019.

[27] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, 2017.

[28] Y. Li and G. A. Aruma Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, pp. 950–953, Dec. 2018.

[29] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Spectral efficiency maximization for uplink cell-free massive MIMO-NOMA networks," in *2019 IEEE Int. Conf. Commun. Workshops*, pp. 1–6, May 2019.

[30] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "NOMA/OMA mode selection-based cell-free massive MIMO," in *Proc. ICC 2019 - 2019 IEEE Int. Conf. Commun. (ICC)*, pp. 1–6, May 2019.

[31] T. H. Nguyen, T. K. Nguyen, H. D. Han, and V. D. Nguyen, "Optimal power control and load balancing for uplink cell-free multi-user massive MIMO," *IEEE Access*, vol. 6, pp. 14462–14473, 2018.

[32] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 4445–4459, July 2017.

[33] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun.*, vol. 13, no. 4, pp. 56–61, 2006.

[34] E. Björnson, R. Zakhour, D. Gesbert, and B. Ottersten, "Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical CSI," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4298–4310, 2010.

[35] T. K. Nguyen, H. H. Nguyen, and T. H. Nguyen, "Multiuser massive MIMO systems with time-offset pilots and successive interference cancellation," *IEEE Access*, vol. 7, pp. 132748–132762, 2019.

[36] J. Winter, "Optimum combining in digital mobile radio with cochannel interference," *IEEE J. Sel. Area in Commun.*, vol. 2, pp. 539–583, 1984.

[37] F. Rezaei, C. Tellambura, A. Tadaion, and A. R. Heidarpour, "Rate analysis of cell-free massive MIMO-NOMA with three linear precoders," *IEEE Trans. Commun.*, 2020.

[38] T. K. Nguyen, H. H. Nguyen, and H. D. Tuan, "Cell-Free Massive MIMO-NOMA with Optimal Backhaul Combining", to appear in *Proc. ICCE 2021.*

[39] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* PTR Prentice Hall, 1993.

[40] T. V. Chien and E. Björnson, "Massive MIMO communications," in *5G Mobile Communications*, pp. 77–116, Springer International Publishing, oct 2016.

[41] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Trans. on Wireless Commun.*, vol. 15, pp. 6384–6399, Sept. 2016.

[42] F. Rezaei, A. R. Heidarpour, C. Tellambura, and A. Tadaion, "Underlaid spectrum sharing for cell-free massive MIMO-NOMA," *IEEE Commun. Lett.*, p. 1, 2020.

[43] A. A. Nasir, H. D. Tuan, D. T. Ngo, T. Q. Duong, and H. V. Poor, "Beamforming design for wireless information and power transfer systems: Receive power-splitting vs transmit time-switching," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 876–889, 2017.

[44] H. Tan, "Optimal combining of residual carrier array signals in correlated noises," *The Telecommunications and Data Acquisition Progress Report 42-124, October - December 1995*, February 15, 1996.

[45] H. Tuy, *Convex Analysis and Global Optimization (second edition).* Springer International, 2016.

5. Adaptive Successive Interference Cancellation in Cell-free Massive MIMO-NOMA

Accepted as:

The Khai Nguyen, Ha H. Nguyen and Hoang Duong Tuan, “Adaptive Successive Interference Cancellation in Cell-free Massive MIMO-NOMA”, *IEEE Vehicular Technology Conference (VTC 2020-Fall)*, Victoria, Canada, October 4-7, 2020.

In the previous chapter, a NOMA-aided cell-free massive MIMO system has been considered under the assumption that successive interference cancellation (SIC) can be carried out perfectly. To study the effect of detection error to the system's performance, in this manuscript, we extend the work in Chapter 4 to consider imperfect SIC. By treating noise plus interference as white Gaussian noise, a discrete statistical model between the transmitted signal and received uplink signal is derived and used to develop a more proper implementation of SIC. The developed adaptive SIC method is analytically and numerically shown to be better than the conventional SIC method.

Adaptive Successive Interference Cancellation in Cell-free Massive MIMO-NOMA

The Khai Nguyen, Ha H. Nguyen, and Hoang Duong Tuan

Abstract

This paper proposes a novel successive interference cancellation method to enhance the ergodic spectral efficiency of a cell-free massive multiple-input multiple-output (MIMO) system with non-orthogonal multiple-access (NOMA). Unlike the majority of existing research works on performance evaluation of NOMA, which assume perfect channel state information and perfect data detection for successive interference cancellation, we take into account the effect of practical (hence imperfect) successive interference cancellation (SIC). We show that the received signal at the backhaul network of a cell-free massive MIMO-NOMA system can be effectively treated as a signal received over an AWGN channel. As a result, a discrete joint distribution between the interfering signal and its detected version can be analytically found, from which an adaptive SIC scheme is proposed to improve performance of interference cancellation.

5.1 Introduction

On the journey to the sixth generation (6G) of wireless networks, cell-free massive MIMO has become one of the most promising technological advances to enable very high speed and energy-efficient communications with low latency [1–5]. With a massive number of single-antenna access points (APs) ubiquitously distributed, or a few base stations (BSs) each equipped with a massive number of antennas, a cell-free system is capable of simultaneously serving a large number of users in the same frequency resources [1–5]. Furthermore, when integrated with non-orthogonal multiple-access (NOMA), a cell-free massive MIMO-NOMA system can provide a flexible tradeoff between accommodating a very large number of terminals and having lower per-user spectral efficiency (SE) [4, 6, 7].

One critical issue when incorporating NOMA into a cell-free system (or any other communication systems in general) is how well successive interference cancellation (SIC) can perform [4, 6–8]. Previous works on performance evaluation of NOMA assume that SIC can be carried out perfectly [9–12]. This assumption implies perfect channel state information (CSI) and perfect signal detection before the SIC stage, which are not possible in practical systems.

To deal with imperfect CSI when implementing SIC, the works in [4,13,14] have estimated and used the effective channel gains for cancelling the interfering signals, which are known by the receiver. As a result, after implementing SIC, only a small residue interference remains while a substantial amount of correlated interference is eliminated. Nevertheless, no existing works have considered the effect of errors in data detection before the SIC stage. In recent attempts to address the impact of both imperfect CSI and imperfect data detection, the authors in [6–8, 15] model the decoded signal as a linear function of the transmitted signal. This model depends on the correlation coefficient between the transmitted signal and the decoded signal, which is suggested to be obtained by long time observation. However, this approach does not lead to an accurate analysis of the imperfect SIC and it cannot account for the change when real time power optimization is applied.

Motivated by the above discussions, we propose in this paper a method to analytically obtain the statistical relationship between the transmitted and decoded signals and use that information to develop an adaptive SIC technique. The system model considered in this paper is similar to the one in [16] where a cell-free massive MIMO-NOMA system consisting of a few massive-antenna BSs simultaneously serving all users in the network (i.e., there is no cell boundary). Using a quadrature-amplitude modulation (QAM) constellation, a statistical relationship between the signals at the input and output of such a system is obtained. Due to the combination of signals from multiple massive-antenna BSs, noise and interference at the backhaul central processing unit (CPU) can be effectively modeled as a Gaussian random variable, which allows us to consider the signal detection problem in a generalized cell-free massive MIMO-NOMA system the same as that over an AWGN channel. With such an equivalent model, the correlation coefficient between the transmitted signal and the decoded signal (before applying SIC) is analytically found. Using the obtained correlation information, we then introduce an adaptive SIC algorithm to enhance the system's ergodic spectral efficiency (SE). Our proposed method is analytically and numerically shown to be better than the conventional SIC method.

5.2 System Model

The system model considered in this paper is the same as that in [16]. Specifically, we consider the uplink (UL) of a cell-free massive MIMO-NOMA system, which comprises of L base stations, each having M antennas and simultaneously serving $2K$ users. All BSs are connected to a backhaul network over which the signals from all L BSs are sent to and processed at a CPU. The users are assigned into K groups with two users in each group who share the same pilot, whereas pilot sequences assigned to different groups are pairwise orthogonal. Such pilot sharing allows more users to be served with a fixed number of pilots, and as a consequence, more information symbols are sent in every coherence interval as compared to using orthogonal pilots. As one of the main contributions of [16], an optimal backhaul combining (OBC) method that maximizes the UL signal-to-interference-plus-noise ratio (SINR) was developed and shown to effectively mitigate the correlated interference. In order to present our proposed adaptive SIC method, this section reviews the main results in [16] with respect to the OBC method.

With the system operating in the time-division duplex (TDD) mode, a set of K length-τ_p pilot sequences is used for UL channel estimation. These pilots are collectively represented by a $\tau_p \times K$ pilot matrix $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K]$ which satisfies $\mathbf{\Phi}^H \mathbf{\Phi} = \tau_p \mathbf{I}_K$. With 2 users using the same pilot sequence and different groups using orthogonal pilots, the signal vector $\mathbf{Y}_l \in \mathbb{C}^{M \times \tau_p}$ received at the lth BS over τ_p time slots (symbols) is given as

$$\mathbf{Y}_l = \sum_{k=1}^{K} \left(\boldsymbol{h}_{l,1,k} \sqrt{p_{1,k}^{(\mathrm{p})}} \boldsymbol{\phi}_k^H + \boldsymbol{h}_{l,2,k} \sqrt{p_{2,k}^{(\mathrm{p})}} \boldsymbol{\phi}_k^H \right) + \mathbf{N}_l, \tag{5.1}$$

where $p_{g,k}^{(\mathrm{p})}$ denotes pilot power, $\boldsymbol{h}_{l,g,k} \sim \mathcal{CN}\left(0, \beta_{l,g,k} \mathbf{I}_M\right)$ is the uncorrelated Rayleigh fading channel between the gth user of the kth group and the lth BS, and $\beta_{l,g,k}$ is the large scale fading coefficient.

To estimate the channel for the gth user in the qth group, the lth BS multiplies the received signal with the corresponding pilot of the qth group, which yields

$$\mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \boldsymbol{h}_{l,1,q}\sqrt{p^{(\mathrm{p})}_{1,q}\tau_p} + \boldsymbol{h}_{l,2,q}\sqrt{p^{(\mathrm{p})}_{2,q}\tau_p} + \mathbf{N}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}. \tag{5.2}$$

Then the minimum mean squared error (MMSE) estimate of $h_{l,g,q}$ can be obtained as [17]

$$\hat{\boldsymbol{h}}_{l,1,q} = \frac{\mathrm{cov}\left\{\boldsymbol{h}_{l,1,q}, \mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right\}}{\mathrm{var}\left\{\mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right\}} \mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \mu_{l,1,q} \mathbf{Y}_l \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}, \tag{5.3}$$

where $\mu_{l,g,q} = \frac{\sqrt{p^{(p)}_{g,q}\tau_p}\beta_{l,g,q}}{p^{(p)}_{1,q}\tau_p\beta_{l,1,q}+p^{(p)}_{2,q}\tau_p\beta_{l,2,q}+\sigma^2_{\mathrm{UL}}}$. As a result, the estimated channel is a random variable with distribution $\hat{\boldsymbol{h}}_{l,g,q} \sim \mathcal{CN}\left(0, \gamma_{l,g,q}\mathbf{I}_M\right)$, where $\gamma_{l,g,q} = \frac{p^{(\mathrm{p})}_{g,q}\tau_p\beta^2_{l,g,q}}{p^{(\mathrm{p})}_{1,q}\tau_p\beta_{l,1,q}+p^{(\mathrm{p})}_{2,q}\tau_p\beta_{l,2,q}+\sigma^2_{\mathrm{UL}}}$. Furthermore, the channel estimation error $\boldsymbol{e}_{l,g,q} = \boldsymbol{h}_{l,g,q} - \hat{\boldsymbol{h}}_{l,g,q}$ is independent of the estimated channel and distributed as $\boldsymbol{e}_{l,g,q} \sim \mathcal{CN}\left(0, (\beta_{l,g,q} - \gamma_{l,g,q})\mathbf{I}_M\right)$.

After CSI is acquired, UL data transmission is carried out. The UL data signal received at the lth BS over each symbol time can be presented as:

$$\boldsymbol{y}_l = \sum_{g=1}^{2}\sum_{k=1}^{K} \boldsymbol{h}_{l,g,k}\sqrt{p_{g,k}}x_{g,k} + \boldsymbol{n}_l, \tag{5.4}$$

where $p_{g,k}$ is the UL transmit power of the gth user of the kth group and $x_{g,k}$ represents its data signal. In order to extract the signal of the first user of the qth group, before being sent to the backhaul CPU, the signal in (5.4) is multiplied with $\hat{\boldsymbol{h}}^H_{l,1,q}$, which leads to:

$$\begin{aligned}
\kappa_{l,1,q} &= \hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{y}_l = \hat{\boldsymbol{h}}^H_{l,1,q}\sum_{g=1}^{2}\sum_{k=1}^{K} \boldsymbol{h}_{l,g,k}\sqrt{p_{g,k}}x_{g,k} + \hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{n}_l \\
&= \underbrace{\mathbb{E}\left\{\hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{h}_{l,1,q}\right\}\sqrt{p_{1,q}}x_{1,q}}_{\mathrm{DS}_{l,1,q}\text{- Desired signal}} + \underbrace{\left(\hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{h}_{l,1,q} - \mathbb{E}\left\{\hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{h}_{l,1,q}\right\}\right)\sqrt{p_{1,q}}x_{1,q}}_{\mathrm{CU}_{l,1,q}\text{- Channel gain uncertainty}} \\
&+ \underbrace{\hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{h}_{l,2,q}\sqrt{p_{2,q}}x_{2,q}}_{\mathrm{IwG}_{l,1,q}\text{- Interference within group}} + \underbrace{\sum_{g=1}^{2}\sum_{k=1,k\neq q}^{K} \hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{h}_{l,g,k}\sqrt{p_{g,k}}x_{g,k}}_{\mathrm{IoG}_{l,1,q}\text{- Interference from other group}} + \underbrace{\hat{\boldsymbol{h}}^H_{l,1,q}\boldsymbol{n}_l}_{\mathrm{N}_{l,1,q}\text{- Noise}},
\end{aligned} \tag{5.5}$$

Excluding the desired signal, it not hard to verify that the interference within group is correlated across the BSs since $\mathbb{E}\left\{\mathrm{IwG}_{l,1,q}, \mathrm{IwG}_{l',1,q}\right\} \neq 0$, $\forall l \neq l'$, whereas all other components are uncorrelated across BSs. As a result, the vector $\boldsymbol{\kappa}_{1,q} = \left[\kappa_{1,1,q}, \kappa_{2,1,q} \ldots \kappa_{L,1,q}\right]^T$ can be decomposed into three length-L vectors: the desired signal, the uncorrelated interference-plus-noise $\boldsymbol{u}_{1,q}$, and the correlated interference-plus-noise $\boldsymbol{c}_{1,q}$. Specifically,

$$\boldsymbol{\kappa}_{1,q} = \boldsymbol{s}_{1,q} x_{1,q} + \boldsymbol{c}_{1,q} x_{2,q} + \boldsymbol{u}_{1,q} \tag{5.6}$$

where $\boldsymbol{s}_{1,q} = [s_{1,1,q}, s_{2,1,q} \ldots s_{L,1,q}]$ with $s_{1,1,q}$ being the desired signal gain:

$$s_{l,1,q} = \mathrm{DS}_{l,1,q} = \mathbb{E}\left\{\hat{\boldsymbol{h}}_{l,1,q}^{H} \boldsymbol{h}_{l,1,q}\right\} \sqrt{p_{1,q}}. \tag{5.7}$$

The elements of $\boldsymbol{c}_{1,q}$ and $\boldsymbol{u}_{1,q}$ are as follows:

$$c_{l,1,q} = \mu_{l,1,q} \mathbb{E}\left\{\boldsymbol{h}_{l,2,q}^{H} \boldsymbol{h}_{l,2,q}\right\} \sqrt{p_{2,q}^{(\mathrm{p})} p_{2,q} \tau_p} \tag{5.8}$$

$$u_{l,1,q} = \mathrm{CU}_{l,1,q} + \mathrm{IoG}_{l,1,q} + \mathrm{N}_{l,1,q} + \mathrm{IwG}_{l,1,q} - c_{l,1,q} x_{2,q} \tag{5.9}$$

Next, normalize (i.e., scale) the signal vector in (5.6) so that the uncorrelated interference-plus-noise term has unit power. This is achieved by simply diving the lth element by $\mathbb{E}\left\{|u_{l,1,q}|\right\}$. It can be easily seen that:

$$\begin{aligned}
\mathbb{E}\left\{|u_{l,1,q}|^2\right\} &= \sum_{g=1}^{2} \sum_{k=1}^{K} p_{g,k} \beta_{l,g,k} \gamma_{l,1,q} M + \sigma_{\mathrm{UL}}^2 \gamma_{l,1,q} M && (5.10)\\
\mathbb{E}\left\{|s_{l,1,q}|^2\right\} &= p_{1,q} \gamma_{l,1,q}^2 M^2 && (5.11)\\
\mathbb{E}\left\{|c_{l,1,q}|^2\right\} &= p_{2,q} \frac{p_{2,q}^{(\mathrm{p})}}{p_{1,q}^{(\mathrm{p})}} \left(\gamma_{l,1,q} \frac{\beta_{l,2,q}}{\beta_{l,1,q}}\right)^2 M^2 && (5.12)
\end{aligned}$$

The normalization produces the following equivalent signal vector:

$$\hat{\boldsymbol{\kappa}}_{1,q} = \hat{\boldsymbol{s}}_{1,q} x_{1,q} + \hat{\boldsymbol{c}}_{1,q} x_{2,q} + \hat{\boldsymbol{u}}_{1,q} \tag{5.13}$$

where $\mathbb{E}\left\{\hat{\boldsymbol{u}}_{1,q} \hat{\boldsymbol{u}}_{1,q}^{H}\right\} = \mathbf{I}_L$. Furthermore, it can be shown that the variance of the normalized effective channel gain $\hat{s}_{l,1,q}$ is exactly the signal to uncorrelated-interference-plus noise ratio of the first user of the qth group at the lth BS:

$$\mathbb{E}\left\{|\hat{s}_{l,1,q}|^2\right\} = \frac{M p_{1,q} \gamma_{l,1,q}}{\sum_{g=1}^{2} \sum_{k=1}^{K} p_{g,k} \beta_{l,g,k} + \sigma_{\mathrm{UL}}^2} \triangleq \xi_{l,1,q}. \tag{5.14}$$

Likewise, the variance of each element in the normalized correlated interference term $\hat{c}_{l,1,q}$ is

$$\mathbb{E}\left\{|\hat{c}_{l,1,q}|^2\right\} = \frac{M p_{2,q} \gamma_{l,2,q}}{\sum_{g=1}^{2} \sum_{k=1}^{K} p_{g,k} \beta_{l,g,k} + \sigma_{\mathrm{UL}}^2} \triangleq \xi_{l,2,q}, \tag{5.15}$$

which is exactly the signal-to-uncorrelated-interference ratio of the second user of the qth group at the lth BS.

Finally, the normalized signal vector $\hat{\boldsymbol{\kappa}}_{1,q}$ is combined at the backhaul CPU with a weight vector $\boldsymbol{w}_{1,q} \in \mathbb{C}^{L\times 1}$, which results in:

$$r_{1,q} = \boldsymbol{w}_{1,q}^T \hat{\boldsymbol{\kappa}}_{1,q} = \boldsymbol{w}_{1,q}^T (\hat{\boldsymbol{s}}_{1,q} x_{1,q} + \hat{\boldsymbol{c}}_{1,q} x_{2,q} + \hat{\boldsymbol{u}}_{1,q}) \tag{5.16}$$

Following the same derivation steps in [16, 18], the combining vector that maximizes the effective SINR of the combined signal in (5.16) is

$$\boldsymbol{w}_{1,q}^{(\mathrm{OBC})} = \alpha \hat{\mathbf{R}}_{1,q}^{-1} \hat{\boldsymbol{s}}_{1,q} = \alpha \big(\hat{\boldsymbol{c}}_{1,q} \hat{\boldsymbol{c}}_{1,q}^H + \mathbf{I}_L \big)^{-1} \hat{\boldsymbol{s}}_{1,q} \tag{5.17}$$

where α is a constant [16] and the matrix $\hat{\mathbf{R}}_{1,q}$ is the correlation matrix of the total interference plus noise:

$$\hat{\mathbf{R}}_{1,q} \;=\; \mathbb{E}\left\{ \hat{\boldsymbol{u}}_{1,q} \hat{\boldsymbol{u}}_{1,q}^H + \hat{\boldsymbol{c}}_{1,q} \hat{\boldsymbol{c}}_{1,q}^H \right\} = \hat{\boldsymbol{c}}_{1,q} \hat{\boldsymbol{c}}_{1,q}^H + \mathbf{I}_L \tag{5.18}$$

The resulting maximum effective SINR is given as:

$$\mathrm{SINR}_{1,q}^{(\mathrm{OBC})} = \hat{\boldsymbol{s}}_{1,q}^H \hat{\mathbf{R}}_{1,q}^{-1} \hat{\boldsymbol{s}}_{1,q} = \sum_{l=1}^{L} \xi_{l,1,q} - \frac{\left(\sum_{l=1}^{L} \sqrt{\xi_{l,1,q} \xi_{l,2,q}} \right)^2}{1 + \sum_{l=1}^{L} \xi_{l,2,q}}. \tag{5.19}$$

Of course, the same result can be obtained for the second user in every group [16].

5.3 Successive Interference Cancellation

To further enhance the system's performance, SIC can be carried out by subtracting the decoded data of the first user when detecting the data of the second user in every group. With the conventional SIC method, the decoded data of the first user in each group is directly subtracted from the overall signal, which is an optimal strategy if data detection of the first user is perfect. However, such a direct subtraction operation is not optimal when there are detection errors. Therefore, in this section, we introduce an adaptive method to improve performance of SIC.

Let $\hat{x}_{1,q}$ denote the detected data of the first user in the gth group. In order to minimize the power of the residue interference after performing SIC, instead of directly subtracting

$\hat{x}_{1,q}$ as in the conventional SIC, we propose an adaptive SIC method to improve the ergodic UL spectral efficiency. Specifically, we aim at finding $\alpha_{1,q}^{(\mathrm{I})}$ and $\alpha_{1,q}^{(\mathrm{Q})}$ such that:

$$\underset{\alpha_{1,q}^{(\mathrm{I})},\alpha_{1,q}^{(\mathrm{Q})}}{\text{minimize}} \quad \mathbb{E}\left\{\left|x_{1,q}-\alpha_{1,q}^{(\mathrm{I})}\Re\{\hat{x}_{1,q}\}-j\alpha_{1,q}^{(\mathrm{Q})}\Im\{\hat{x}_{1,q}\}\right|^2\right\}. \tag{5.20}$$

It is not hard to see that the above problem is to minimize a quadratic function and the solution is $\alpha_{1,q}^{(\mathrm{I})}=\rho_{1,q}^{(\mathrm{I})}$ and $\alpha_{1,q}^{(\mathrm{Q})}=\rho_{1,q}^{(\mathrm{Q})}$, where $\rho_{1,q}^{(\mathrm{I})}$ and $\rho_{1,q}^{(\mathrm{Q})}$ are the correlation coefficients between $x_{1,q}$ and $\hat{x}_{1,q}$ on the inphase (I) and quadrature (Q) channels, respectively.

The signal vector after implementing the adaptive SIC is:

$$\hat{\boldsymbol{\kappa}}_{2,q}^{(\mathrm{aSIC})}=\hat{\boldsymbol{s}}_{2,q}x_{2,q}+\underbrace{\hat{\boldsymbol{c}}_{2,q}\left(x_{1,q}-\rho_{1,q}^{(\mathrm{I})}\Re\{\hat{x}_{1,q}\}-j\rho_{1,q}^{(\mathrm{Q})}\Im\{\hat{x}_{1,q}\}\right)}_{\text{Residue interference}}+\hat{\boldsymbol{u}}_{2,q}. \tag{5.21}$$

With the above adaptive SIC, the SINR of the second user of the qth group becomes

$$\mathrm{SINR}_{2,q}^{(\mathrm{OBC-aSIC})}=\sum_{l=1}^{L}\xi_{l,2,q}-\frac{\left(\sum_{l=1}^{L}\sqrt{\xi_{l,1,q}^{(\mathrm{aSIC})}\xi_{l,2,q}}\right)^2}{1+\sum_{l=1}^{L}\xi_{l,1,q}^{(\mathrm{aSIC})}}, \tag{5.22}$$

where:

$$\xi_{l,1,q}^{(\mathrm{aSIC})}\triangleq\frac{M\left(1-\frac{(\rho_{1,q}^{(\mathrm{I})})^2+(\rho_{1,q}^{(\mathrm{Q})})^2}{2}\right)p_{1,q}\gamma_{l,1,q}}{\sum_{g=1}^{2}\sum_{k=1}^{K}p_{g,k}\beta_{l,g,k}+\sigma_{\mathrm{UL}}^2}. \tag{5.23}$$

It is pointed out that the case of conventional SIC corresponds to $\alpha_{1,q}^{(\mathrm{I})}=\alpha_{1,q}^{(\mathrm{Q})}=1$ and we have the same result as in (5.22), but with $\xi_{l,1,q}^{(\mathrm{aSIC})}$ replaced by $\xi_{l,1,q}^{(\mathrm{nSIC})}$, given as

$$\xi_{l,1,q}^{(\mathrm{nSIC})}\triangleq\frac{M\left(2-\rho_{1,q}^{(\mathrm{I})}-\rho_{1,q}^{(\mathrm{Q})}\right)p_{1,q}\gamma_{l,1,q}}{\sum_{g=1}^{2}\sum_{k=1}^{K}p_{g,k}\beta_{l,g,k}+\sigma_{\mathrm{UL}}^2}. \tag{5.24}$$

Comparing (5.24) and (5.23) reveals that the power of the residue correlated interference in the proposed adaptive SIC method is consistently smaller than that in the conventional SIC. Intuitively, the case that $\rho_{1,q}^{(\mathrm{I})}\to 1$ and $\rho_{1,q}^{(\mathrm{Q})}\to 1$ means very good data detection of the first user of the qth group, hence the adaptive SIC will remove most of the correlated interference for the second user. On the other hand, the case that $\rho_{1,q}^{(\mathrm{I})}\to 0$ and $\rho_{1,q}^{(\mathrm{Q})}\to 0$ means very poor data detection of the first user of the qth group, hence the adaptive SIC will not subtract any amount of signal from the overall signal because doing so degrades the quality of the signal used in data detection of the second user.

Obviously, an important information required by the proposed adaptive SIC method is the correlation coefficients. The remainder of this section shows how to obtain such information. From (5.16), it can be seen that all the components are independent and identical distributed. With a very large number of users, the signal combined at the backhaul CPU from the massive number of antennas at the BSs can be effectively approximated by a Gaussian random variable by applying the central limit theorem (see Fig. 5.1 for the verification of such an approximation). As a result, the detection problem of the signal in (5.16) can be simplified to the detection problem over an AWGN channel as:

$$\hat{r}_{1,q} = x_{1,q} + z_{1,q} \tag{5.25}$$

where:

$$z_{1,q} = \frac{\boldsymbol{w}_{1,q}^T(\hat{\boldsymbol{c}}_{1,q}x_{2,q} + \hat{\boldsymbol{u}}_{1,q})}{\boldsymbol{w}_{1,q}^T\hat{\boldsymbol{s}}_{1,q}} \to \mathcal{CN}\left(0, 1 \Big/ \mathrm{SINR}_{1,q}^{(\mathrm{OBC})}\right). \tag{5.26}$$

Consider a 16-QAM constellation. The problem of finding $\rho_{1,q}^{(\mathrm{I})}$ and $\rho_{1,q}^{(\mathrm{Q})}$ can be broken down into considering transmitting two 4-PAM signals over two independent AWGN channels (I and Q channels). In order to have $\mathbb{E}\{|x_{1,q}|^2\} = 1$ the Euclidean distance between two adjacent points in the 4-PAM constellation $\{\pm\Delta, \pm 3\Delta\}$ should be $2\Delta = 2/\sqrt{10}$. The signal after backhaul combining $r_{1,q}$ in (5.16) has the SINR specified by (5.19). As a result, with a 4-PAM signal transmitted over a AWGN channel and experiences the SNR specified in (5.16), the conditional symbol error probabilities are obtained as in Table 5.1. In the table, we have $P_{i\Delta} = Q\left(i\sqrt{\frac{\mathrm{SINR}_{1,q}^{(\mathrm{OBC})}}{5}}\right)$, for $i = 1, 3, 5$.

From Table 5.1, the correlation coefficient between $x_{1,q}$ and $\hat{x}_{1,q}$ on either the I or Q channel can be computed as:

$$\rho_{1,q}^{(\mathrm{I})} = \frac{\mathrm{cov}\left\{\Re\{x_{1,q}\}, \Re\{\hat{x}_{1,q}\}\right\}}{\mathrm{var}\left\{\Re\{x_{1,q}\}\right\}\mathrm{var}\left\{\Re\{\hat{x}_{1,q}\}\right\}} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} s_i s_j P_{s_i \to s_j}}{10\Delta^2} = \frac{10 - 6P_{\Delta} - 8P_{3\Delta} - 6P_{5\Delta}}{10} \tag{5.27}$$

It is pointed out that, since $0 \leq Q(\cdot) \leq 1$, one has $-1 \leq \rho_{1,q}^{(\mathrm{I})} = \rho_{1,q}^{(\mathrm{Q})} \leq 1$, a intuitively satisfying property.

Table 5.1 $P(s_j \text{ decided} \mid s_i \text{ transmitted})$

	s_1	s_2	s_3	s_4
s_1	$1-P_\Delta$	$P_\Delta - P_{3\Delta}$	$P_{3\Delta} - P_{5\Delta}$	$P_{5\Delta}$
s_2	P_Δ	$1-2P_\Delta$	$P_\Delta - P_{3\Delta}$	$P_{3\Delta}$
s_3	$P_{3\Delta}$	$P_\Delta - P_{3\Delta}$	$1-2P_\Delta$	P_Δ
s_4	$P_{5\Delta}$	$P_{3\Delta} - P_{5\Delta}$	$P_\Delta - P_{3\Delta}$	$1-P_\Delta$

5.4 Optimal Power Control with Adaptive SIC

It is of interest to maximize the minimum rate among users subject to a maximum power constraint. Such a max-min QoS power control problem is formulated and solved in [16] for the case of perfect SIC. For the case that the adaptive SIC is applied to the second user in each group, the power control problem is formulated as:

$$\max_{p_{g,q}} \min_{q=1,\ldots,K} \left\{ \mathrm{SINR}_{1,q}^{(\mathrm{OBC})}, \mathrm{SINR}_{2,q}^{(\mathrm{OBC-aSIC})} \right\} \tag{5.28a}$$

$$\text{subject to} \quad 0 \le p_{g,q} \le p_{\max}, \forall p, q \tag{5.28b}$$

According to (5.22) and (5.23), the above power control problem requires the prior knowledge of $\rho_{1,q}^{(\mathrm{I})}$ and $\rho_{1,q}^{(\mathrm{Q})}$. Hence, an iterative algorithm is proposed and summarized in Algorithm 4 to solve the above power control problem, and at the same time obtain the correlation coefficients iteratively.

Algorithm 4 Max-min QoS power control with adaptive SIC

```
Require: Large scale fading coefficients $\beta_{l,g,k}$
Initially set $\rho_{g,q}^{(I)} = \rho_{g,q}^{(Q)} = 0$ (no information of correlation, SIC is not utilized).
while Until convergence do
    Solve (5.28).
    Obtain new $\rho_{g,q}^{(I)}$ and $\rho_{g,q}^{(Q)}$ by (5.27).
end while
return $p_{g,k}$.
```

Initially, since the correlation coefficients between $x_{1,q}$ and $\hat{x}_{1,q}$ are unknown, $\rho_{1,q}^{(\mathrm{I})}$ and $\rho_{1,q}^{(\mathrm{Q})}$ are set to zero, which means that SIC is not utilized in the first iteration. After the first iteration, problem (5.28) is solved by following the method proposed in [16]. With the obtained power allocation, the new value for $\rho_{1,q}^{(\mathrm{I})}$ and $\rho_{1,q}^{(\mathrm{Q})}$ is calculated by (5.27), which is then used to solve (5.28) in the next iteration. The algorithm continues until convergence.

5.5 Numerical Results and Discussion

In this section, simulation results are provided to compare performance of the proposed adaptive SIC and conventional SIC in both perfect and imperfect cases. For simulation, a cell-free system is considered with 9 BSs deployed in a 3×3 grid over a coverage area of 500×500 squared meters, each BS has 16 antennas. The systems simultaneously serves 30 uniformly distributed users. With 2 users assigned into each group, $\tau_p = 15$ orthogonal pilot sequences are required. The large scale fading is defined as $\beta_{l,g,k} = -131 - 42.8\log 10 d_{l,g,k} + z_{l,g,k}$dB, where $d_{l,g,k}$ is the distance from the lth BS to the kth user of the gth group and $z_{l,g,k}$ is the standard deviation of the shadowing variable. The noise figure of 5dB translates to a noise variance of -96dBm. The simulation parameters are summarized in Table 5.2.

Table 5.2 Simulation parameters.

Parameter	**Value**
Peak UL transmit power ($p_{\max}$)	23 dBm
Shadowing standard deviation	10 dB
Penetration loss (indoor users)	20 dB
Noise figure	5 dB
Coherence interval (τ_c)	200 symbols
Pathloss	$131 + 42.8\log 10 d$

First, Fig. 5.1 plots the histogram of noise-plus-interference term in the I channel of an arbitrary user in the system and when the 4-PAM constellation is used. As can be seen, the distribution is very close to the Gaussian fit of the same variance. This means that data detection can be treated as a detection problem over an AWGN channel.

Fig. 5.2 displays the correlation coefficients obtained by our method and the long time

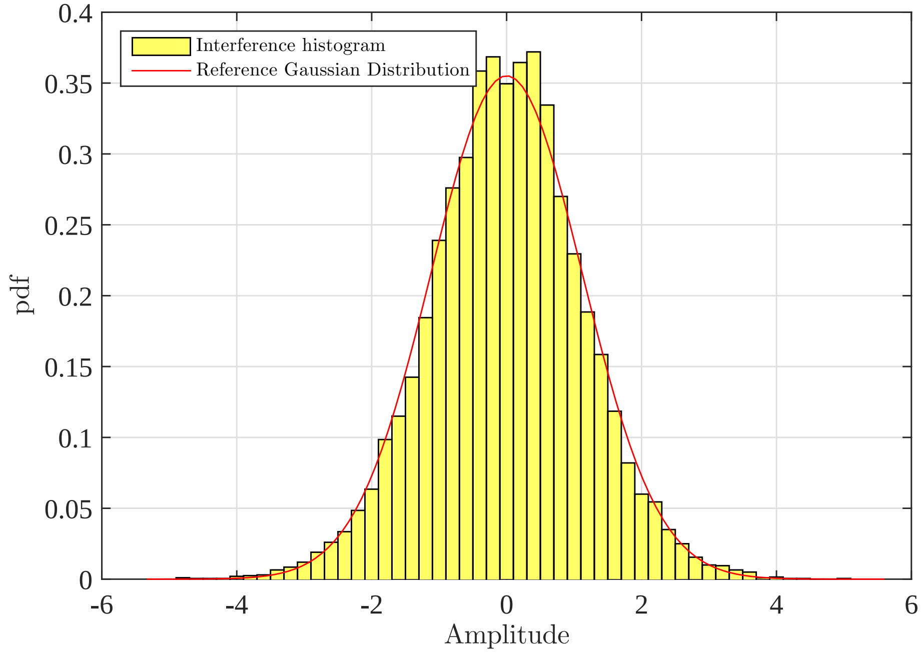


Figure 5.1 Histogram of a user's noise-plus-interference term in the I channel with 4-PAM constellation.

observation method suggested in [6–8, 15]. As expected, the longer the observation time is, the better the estimated values of $\rho_{g,q}^{(\mathrm{I})}$ and $\rho_{g,q}^{(\mathrm{Q})}$ can be obtained. With the assumption that the channels stay unchanged within τ_c symbols, the observation time grows proportionally with τ_c. Another issue with the long time observation approach is that by the time $\rho_{g,q}^{(\mathrm{I})}$ and $\rho_{g,q}^{(\mathrm{Q})}$ are acquired, the locations or transmit powers of users may have changed significantly, which can readily change the correlation coefficients. In contrast, our proposed method can determine the correlation coefficients in every coherence interval and the obtained values are very accurate.

Finally, Fig. 5.3 plots the cumulative distribution functions (CDF) of the achievable QoS for different SIC methods (as noted in the figure's legend). The figure clearly shows that imperfect data detection severely degrades performance of the conventional SIC method. With the proposed adaptive SIC, the CDF curve is much more favorable and closely approaches the CDF curve achieved with the perfect SIC, especially for QoS values higher than 2 bits/sec/Hz.

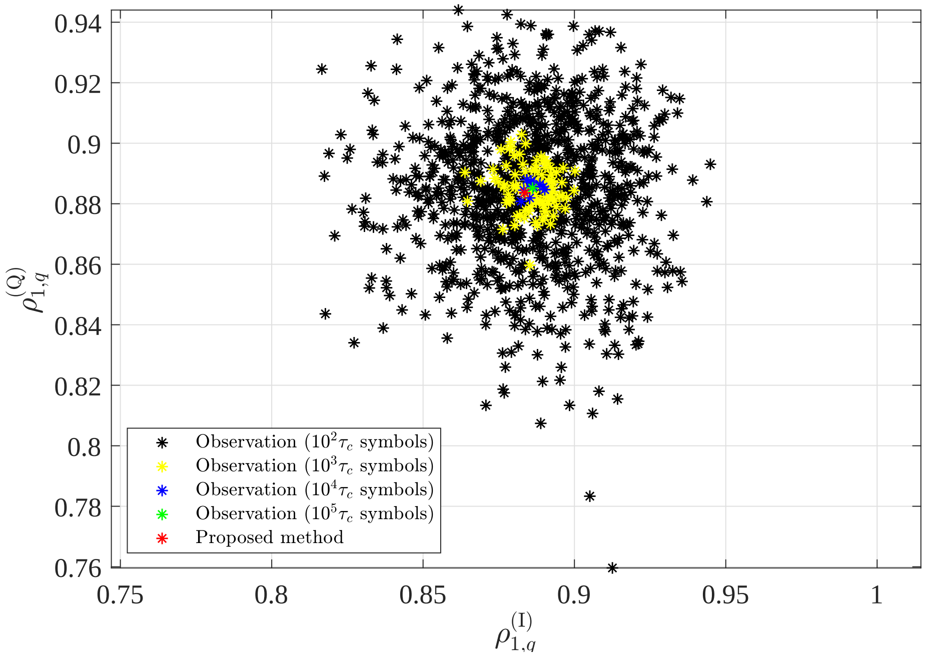


Figure 5.2 Correlation coefficients on the I and Q channels.

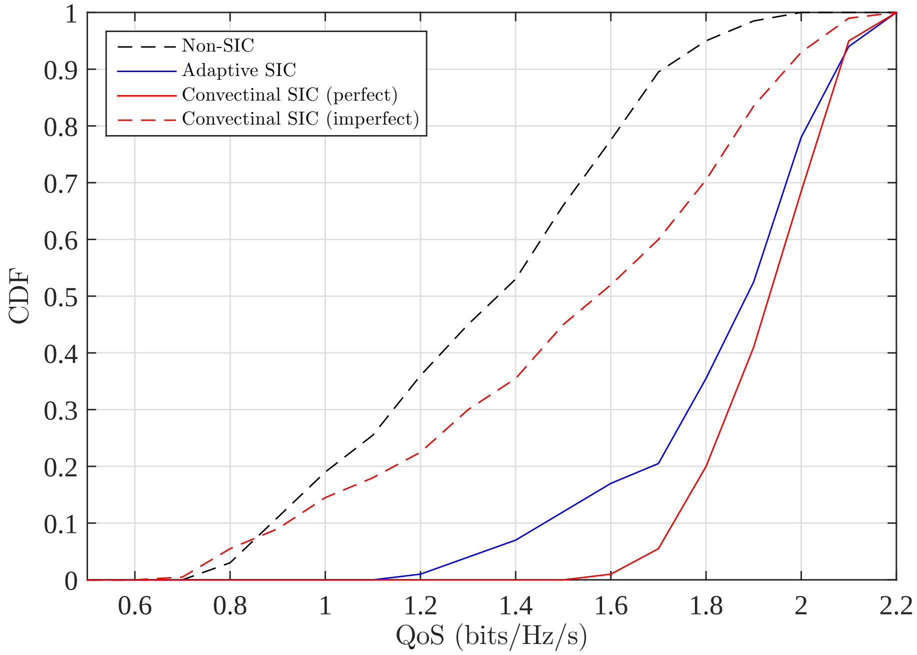


Figure 5.3 Cumulative distribution functions of the achievable QoS.

5.6 Conclusion

The effects of imperfect CSI and data detection errors on the SIC operation have been investigated for a cell-free massive MIMO-NOMA system. Unlike previous works which assume a linear relationship between the transmitted and decoded signals before the SIC stage, we develop a nonlinear model when the transmitted signal is drawn from a QAM constellation. With the interference being effectively treated as Gaussian noise, the relationship between the signals used by the SIC operation can be modeled with a discrete joint probability distribution. From the obtained relationship, we propose an adaptive SIC method to enhance the ergodic uplink spectral efficiency among all users in the network. The proposed method is analytically and numerically shown to be better than the conventional SIC method.

References

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 1834–1850, Mar. 2017.

[2] T. M. Hoang, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and A. Marshall, "Cell-free massive MIMO networks: Optimal power control against active eavesdropping," *IEEE Transactions on Communications*, vol. 66, pp. 4724–4737, Oct. 2018.

[3] T. H. Nguyen, T. K. Nguyen, H. D. Han, and V. D. Nguyen, "Optimal power control and load balancing for uplink cell-free multi-user massive MIMO," *IEEE Access*, vol. 6, pp. 14462–14473, 2018.

[4] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the performance of cell-free massive MIMO relying on adaptive NOMA/OMA mode-switching," *IEEE Trans. Commun.*, p. 1, 2019.

[5] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 8, pp. 616–619, Apr. 2019.

[6] Y. Li and G. A. Aruma Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, pp. 950–953, Dec. 2018.

[7] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Spectral efficiency maximization for uplink cell-free massive MIMO-NOMA networks," in *2019 IEEE Int. Conf. Commun. Workshops*, pp. 1–6, May 2019.

[8] P. Li, R. C. de Lamare, and R. Fa, "Multiple feedback successive interference cancellation detection for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 2434–2439, Aug. 2011.

[9] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A

survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas in Commun.*, vol. 35, pp. 2181–2195, Oct. 2017.

[10] Z. Ding, P. Fan, and H. V. Poor, "On the coexistence between full-duplex and NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, pp. 692–695, Oct. 2018.

[11] H. D. Tuan, A. A. Nasir, H. H. Nguyen, T. Q. Duong, and H. V. Poor, "Non-orthogonal multiple access with improper Gaussian signaling," *IEEE J. Sel. Topics in Signal Process.*, vol. 13, pp. 496–507, June 2019.

[12] X. Liu, J. Wang, N. Zhao, Y. Chen, S. Zhang, Z. Ding, and F. R. Yu, "Placement and power allocation for NOMA-UAV networks," *IEEE Wireless Commun. Lett.*, vol. 8, pp. 965–968, June 2019.

[13] T. K. Nguyen, H. H. Nguyen, and T. H. Nguyen, "Multiuser massive MIMO systems with time-offset pilots and successive interference cancellation," *IEEE Access*, vol. 7, pp. 132748–132762, 2019.

[14] D. Kudathanthirige and G. A. A. Baduge, "NOMA-aided multicell downlink massive MIMO," *IEEE J. Sel. Topics in Signal Process.*, vol. 13, pp. 612–627, June 2019.

[15] X. Chen, R. Jia, and D. W. K. Ng, "On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 67, pp. 2539–2551, Mar. 2019.

[16] T. K. Nguyen, H. H. Nguyen, and H. D. Tuan, "Max-min QoS power control in generalized cell-free massive MIMO-NOMA with optimal backhaul combining," *Summitted to IEEE Trans. Vel. Technol.*, 2019.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* PTR Prentice Hall, 1993.

[18] J. Winter, "Optimum combining in digital mobile radio with cochannel interference," *IEEE J. Sel. Area in Commun.*, vol. 2, pp. 539–583, 1984.

6. Conclusion

6.1 Conclusion

This thesis has studied the integration of NOMA into massive MIMO systems via the use of time-offset pilots and the SIC algorithm. The main objective for this study is to be able to accommodate more users (i.e., a higher number of connections) in the network with limited radio resources. The main contributions of this research are as follows:

- Chapter 3 investigates the performance of a single-cell massive MIMO system with time-offset pilots with the aid of SIC. Numerical results show that the proposed method can achieve a significant performance enhancement as compared to the conventional orthogonal pilot method.
- Chapter 4 proposes the integration of NOMA into a cell-free massive MIMO network with OBC. Analytical and numerical results are provided to demonstrate that with the use of OBC, a NOMA cell-free massive MIMO system can achieve unlimited performance when the number of antennas at each BS goes to infinity.
- Power control problems are formulated and solved to further improve the system's performance. Due to the NP-hardness of the problem, a successive approximation approach is adopted to convert the original optimization problem to a series of convex problems, whose solutions are feasible to the original one and satisfy the KKT conditions. Simulation results have shown that power control not only improves the system's SE but also users' fairness.
- We also investigate the effect of imperfect SIC and introduce an adaptive SIC method to address this imperfection. Simulation results shown the advantage of the proposed

method over the conventional SIC not only in terms of SE but also in time requirement.

6.2 Future Research Topics

Although the integration of NOMA into massive MIMO systems is promising to solve the problem of limited number of connections in wireless networks, there are several issues that can be further studied and they are discussed below.

- Although reusing pilots in massive MIMO systems can improve the connection capability, assigning too many users with the same pilot sequence can result in very poor channel estimation. Hence, how many users should be assigned with the same pilot sequence and how to optimally decide which users should use the same resources remain interesting questions to study.
- The majority of research works on massive MIMO-NOMA systems consider single-carrier transmission. It is relevant and interesting to exploit the structure of multi-carrier transmission to solve the problem of limited connection capability in a massive MIMO system.