Resource Allocation in Multi-user MIMO Networks: Interference Management and Cooperative Communications

By

Susanna Mosleh

Submitted to the Department of Electrical Engineering and Computer Science and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Prof. Erik Perrins, Chairperson

Prof. Lingjia Liu, Co-chair

Committee members

Prof. Shannon Blunt

Prof. Victor Frost

Prof. Jian Li

Date defended: December 19, 2018

The Dissertation Committee for Susanna Mosleh certifies
that this is the approved version of the following dissertation :

Resource Allocation in Multi-user MIMO Networks:
Interference Management and Cooperative Communications

Prof. Erik Perrins, Chairperson

Date approved: _____

Abstract

Nowadays, wireless communications are becoming so tightly integrated in our daily lives, especially with the global spread of laptops, tablets and smartphones. This has paved the way to dramatically increasing wireless network dimensions in terms of subscribers and amount of flowing data. Therefore, the two important fundamental requirements for the future 5G wireless networks are abilities to support high data traffic and exceedingly low latency. A likely candidate to fulfill these requirements is multi-cell multi-user multi-input multiple-output (MU-MIMO); also termed as coordinated multi-point (CoMP) transmission and reception. To achieve the highest possible performance in MU-MIMO networks, a properly designed resource allocation algorithm is needed. Moreover, with the rapidly growing data traffic, interference has become a major limitation in wireless networks. Interference alignment (IA) has been shown to significantly manage the interference and improve the network performance. However, how practically use IA to mitigate interference in a downlink MU-MIMO network still remains an open problem. In this dissertation, we improve the performance of MU-MIMO networks in terms of spectral efficiency, by designing and developing new beamforming algorithms that can efficiently mitigate the interference and allocate the resources. Then we mathematically analyze the performance improvement of MU-MIMO networks employing proposed techniques. Fundamental relationships between network parameters and the network performance is revealed, which provide guidance on the wireless networks design. Finally, system level simulations are conducted to investigate the performance of the proposed strategies.

Acknowledgements

I would like to express my deepest appreciation to my advisors: Professors Erik Perrins and Lingjia Liu, for their support and encouragement on daily basis from the start until date. Under their guidance I learnt a lot and overcame many difficulties. They also have taught me another aspect of life "*goodness can never be defied and good human beings can never be denied*". For all these, I'm deeply indebted to them throughout my life.

I would also like to extend my deepest gratitude to the member of my dissertation advisory and exam committee: Professors Victor Frost, Shannon Blunt, and Jian Li for generously giving their time to offer me valuable comments toward improving my work.

The completion of this dissertation would not have been possible without the support, sacrifices and nurturing of my loving husband, Javad, and my wonderful son, Adrian. I also feel a deep sense of gratitude for my parents and my grandmother for the selfless love, care, and sacrifice they did to shape my life.

Last but not the least, I would like to extend my sincere thanks to all of those with whom I have had the pleasure to work during my Ph.D. journey for their friendship and the warmth they extended to me during my time at KU.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Nowadays, wireless communications are becoming so tightly integrated in our daily lives, especially with the global spread of laptops, tablets and smartphones. This has paved the way to dramatically increasing wireless network dimensions in terms of subscribers and amount of flowing data. Precisely, the number of mobile-connected devices per capita is expected to reach 1.5 by 2021 and global mobile data traffic is forecast to increase sevenfold between 2016 and 2021. In other words, global mobile data traffic will grow at a compound annual growth rate (CAGR) of 47 percent from 2016 to 2021, reaching 49.0 exabytes per month by 2021 [1]. The volume, velocity, and variety of data from both mobile users and communication networks follow an exponential increase pattern. Consequently, big data will further be entrenched in the upcoming fifth-generation (5G) wireless networks.

The two important fundamental requirements for the future $5G$ wireless networks are abilities to support high data traffic and exceedingly low latency [2]. Likely candidates to fulfill these requirements are multi-cell multi-user multi-input multiple-output (MIMO); also termed as coordinated multi-point (CoMP) transmission and reception; massive MIMO, cell-free massive MIMO, and Cloud Radio Access Networks (Cloud-RAN). These technologies are introduced as promis-

ing technologies in the 3GPP LTE-advanced standard, to manage interference, improve the overall system performance, and enhance system reliability.

A multi-cell multi-user MIMO system allows each base station (BS) to communicate with several co-channel mobile stations (MSs) simultaneously and thereby significantly increases the system throughput. MIMO systems with a large array antenna at the BSs are known as Massive MIMO. In massive MIMO network each BS is equipped with a large array antenna to improve the spectral and energy efficiency of wireless systems with simple signal processing. Massive MIMO allows a BS to simultaneously serve many number of user along with time-frequency resources to improve the overall system performance. In general, depending on the antenna arrays setup at the BSs, massive MIMO can be categorized into the following two architectures: distributed massive MIMO and co-located massive MIMO. While the latter locates the service antennas in a compact area, the former spreads antennas all over a large area. The network architecture of Cloud-RAN spreads several low-cost low-power BSs all over a small area as an alternative to a high-power BS. In order to have an efficient resource allocation and interference management among multiple BSs, digital backhaul links connect all these low-power BSs to a central computing unit (cloud).

Since the wireless medium is inherently shared, allocating resource efficiently and mitigating the interference are essential in the multi-user cellular environment. Interference has become a major limitation in wireless networks especially with the rapidly growing data traffic. To avoid interference, orthogonalization methods (both in time and frequency) are introduced which are practical, however, these methods drastically degrade the spectral efficiency. Numerous techniques allow the increase of the multiplexing gain or the sum rate in multi-user single cell and multi-user multi-cell networks in order to improve spectral efficiency. In a multi-user single cell network all the interference is intra-cell which is due to the transmission by a single base station of multiple streams intended to multiple users. This kind of interference can be taken care of by simple means of zero forcing (ZF) or by the more complex, but optimal, dirty paper coding (DPC) [3]. While in a multi-user multi-cell network, the situation is more complex as the interference at each receiver come from other cells. In these networks, inter-cell interference is a major drawback

and dealing with interference management and joint processing between nodes to suppress inter-cell interference is crucial. Moreover, in order to achieve the highest possible performance of the aforementioned candidate technologies, a properly designed resource allocation algorithm is needed. By designing a resource allocation algorithm, which maximize the network throughput, these technologies are able to manage the exponential growth of wireless network dimensions in terms of both subscribers and amount of flowing data.

To address the above listed problems, the first target of this dissertation is to improve the performance of wireless networks, in terms of spectral efficiency, by developing new algorithms and protocols that can efficiently mitigate the interference and allocate the resources. In particular, we will focus on designing new beamforming algorithms in downlink multi-cell multi-user MIMO networks. Furthermore, we mathematically analyze the performance improvement of massive MIMO networks employing proposed technique. Fundamental relationships between network parameters and the network performance will be revealed, which will provide guidance on the wireless networks design. Finally, the results of theoretical study will be demonstrated using simulation platform such as MATLAB.

1.2 Proposed Research

In this dissertation, our aim is to improve both the cell-edge and the cell-average user spectral-efficiency in a multi-cell multi-user MIMO network by developing a new protocol to suppress the co-channel interference issue. Therefore, in Chapter 2, a novel interference alignment transceiver beam-forming design along with a low complexity iterative coordinated beam-forming scheme is introduced. While the latter combats the intra-cell interference, the former is utilized to mitigate the inter-cell interference. The proposed schemes consider the codebook-based feedback, which is adopted in the LTE/LTE-advanced systems. Optimal downlink user-specific and cell-specific beam-forming matrices are characterized to maximize the lower bound of expected signal-to-leakage-plus noise ratio and to minimize the residual inter-cell interference, respectively. More-

over, closed-form expressions for these beamforming matrices under limited channel state information feedback and in the presence of the quantization error are identified. Simulations are conducted to investigate the performance of the proposed strategy. The results indicate that our scheme can significantly improve the average spectral-efficiency of the underlying network when compared with existing ones where the quantization error is neglected. Furthermore, for a fixed payload size of the codebook, unlike zero-forcing beam-forming in which the sum throughput is bounded as the signal-to-noise ratio (SNR) increases, in our scheme, the performance gap between rank 2 feedback and perfect feedback remains approximately constant as SNR increases.

In Chapter 3, a resource allocation problem is studied for downlink CoMP coordinated beamforming systems, where each BS serves its own MSs. MIMO transmit precoding and resource allocation are linked to the underlying proportional-fair scheduling to ensure a good trade-off between cell-average and cell-edge user spectral-efficiency. Due to the coupled interference among mobile stations, the resulting proportional-fair resource allocation optimization problem becomes nonconvex. To solve for optimal operating point for MIMO CoMP network, a parallel successive convex approximation-based algorithm is introduced. The introduced scheme enables all BSs to update their optimization variables in parallel by solving a sequence of strongly convex subproblems. Closed-form expressions of the locally optimal solution in both the high and low signal-to-noise regimes are characterized. The performance of the introduced scheme is also investigated through simulations. Numerical results show the efficiency of the introduced algorithm.

In order to support ever-growing end-users' needs, in Chapter 4, a Cloud-RAN has been considered. To reduce the backhaul traffic and aid coordinated multi-point (CoMP) transmission, the BS-level caching technique is utilized, where popular contents are pre-fetched at each BS. Then the MIMO operation and user association policy are linked to the underlying cache placement strategy to ensure a good trade-off between load balancing and backhaul traffic taking into account the underlying wireless channel and the finite cache capacity at BSs. Due to the coupled interference among mobile stations, the binary nature of the underlying cache placement and user association matrices, the resulting mixed-timescale mixed integer optimization problem is non-

convex and NP-hard. To solve this problem, we decompose the joint optimization problem into a long-term content placement sub-problem and a short-term content delivery sub-problem. A novel iterative algorithm is introduced by leveraging the alternating direction method of multipliers together with a parallel successive convex approximation-based algorithm. The introduced scheme enables all BSs to update their optimization variables in parallel by solving a sequence of convex subproblems. Simulation evaluation demonstrates the efficiency of our strategy.

Finally, Chapter 5 studies a resource allocation problem for downlink cell-free massive MIMO network. Transmit precoding and power allocation are linked to the underlying max-min scheduling to ensure uniform and excellent service throughout the coverage area. Due to the coupled interference among UEs, the resulting max-min resource allocation optimization problem becomes nonconvex. We demonstrate the uplink-downlink duality and propose an iterative algorithm which solves the primal downlink problem efficiently. By utilizing the max-min beamformer and taking the channel estimation error into account, we further derive the capacity lower bound of the underlying cell-free massive MIMO network. The performance of the introduced scheme is also investigated through simulations. Numerical results show the efficiency of the introduced algorithm.

1.3 Research Impact and Contributions

By developing new protocols to mitigate the co-channel interference and efficiently allocate the resources, we improve both the cell-edge and the cell-average user spectral-efficiency for the ever growing multi-cell multi-user networks. By completing these research tasks, the proposed research will be of value for the design and analysis of the current and emerging wireless networks communications.

- The novelty of this research lies in the ability to develop new beamforming design algorithms which take full advantage of intelligent interference mitigation techniques to meet the 5G technology requirements. We expect that this research would provide great benefits to the

next generation networks by providing the better spectral efficiency.

- The proposed limited feedback-based IA scheme along with the low complexity iterative leakage-based coordinated beam-forming strategy can be considered as a promising interference management technique to achieve the minimum overall residual co-channel interference in a downlink multi-cell multi-user MIMO network. The introduced algorithm efficiently reduces the effect of inter-cell interference from BSs in other cells, and eliminates the intra-cell interference due to the spatial streams dedicated to other users in the same cell.

- The introduced proportional-fair resource allocation is able to simultaneously improve both cell-edge and cell-average user spectral efficiency by combining MIMO transmit precoder and power allocation.

- The proposed mixed-timescale content delivery (precoding and user association) and content placement algorithm in cached cloud radio access networks improves the system performance by considering a long-term content placement and a short-term content delivery. The cache placement reduces the backhaul consumption and provides more CoMP opportunities while the content delivery guarantees to provide a better average throughput to each user.

- The introduced max-min resource allocation in cell-free massive MIMO networks provides a uniform service throughout the coverage area by linking the transmit precoding and power allocation to the underlying max-min scheduling. This algorithm enhances user experience and improve the overall system performance.

- Furthermore, fundamental relationships between network performance and network parameters is revealed, which facilitates new system design. Interestingly, there is possibility of applying many of this research results in various wireless networks such as C-RAN and CoMP.

- System level simulation using MATLAB platform is carried out to verify the analysis in realistic network scenarios.

1.4 Notation

Throughout this dissertation, normal letters are used for scalars. Boldface capital and lower case letters denote matrices and vectors, respectively. The transposition, the Hermitian transposition, and the determinant of a complex matrix $\mathbf{A}$ are denoted by $\mathbf{A}^T$, $\mathbf{A}^H$ and $|\mathbf{A}|$, respectively. An $N \times K$ matrix, with ones on its main diagonal and zeros on its off-diagonal entries, is denoted by $\mathbf{I}_{N\times K}$, while the identity matrix of size N is simply denoted by $\mathbf{I}_N$. An $N \times K$ all-zeros matrix is denoted by $\mathbf{0}_{N\times K}$. The sets of complex and real numbers are denoted by $\mathbb{C}$ and $\mathbb{R}$, respectively. A circularly symmetric complex Gaussian random variable (r.v.) is represented by $Z = X + jY \sim \mathscr{CN}(0, \sigma^2)$, where X and Y are independent and identically distributed (i.i.d.) normal r.v.'s from $\mathscr{N}(0, \frac{\sigma^2}{2})$. $\mathbb{E}[\cdot]$ represents the expectation operator. The trace of a square matrix $\mathbf{A} = [a_{ij}]_{n\times n}$ is defined as $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$. The $\mathrm{vec}(\cdot)$ operator aligns all the elements of a matrix into a column vector by stacking the column vectors of the matrix, i.e., for $\mathbf{A} \in \mathbb{C}^{M\times N}$ then $\mathrm{vec}(\mathbf{A}) \in \mathbb{C}^{MN\times 1}$. Moreover, the Kronecker and Hadamard product between two matrices $\mathbf{A}$ and $\mathbf{B}$ are symbolized by $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{A} \odot \mathbf{B}$, respectively. $\|\mathbf{a}\|_2$ and $\|\mathbf{A}\|_F$ denote the two-norm of the vector $\mathbf{a}$ and the Frobenius norm of the matrix $\mathbf{A}$, respectively. The orthogonality, inner product and angle are symbolized by $\perp\!\!\!\perp$, $< \cdot, \cdot >$, and $\angle$, respectively. The operator $[x]^+$, for $x \in \mathbb{R}$, is equivalent to the operation $\max(0, x)$.

Chapter 2

Multi-cell Multi-user Networks: Interference Mitigation

2.1 Introduction

A multi-cell multi-user multiple-input multiple-output (MU-MIMO) system allows each base station (BS) to communicate with several co-channel mobile stations (MSs) simultaneously and thereby significantly increases the system throughput. For a multi-cell MU-MIMO system, it is essential to mitigate co-channel interference (CCI) in multi-user cellular environments in which each receiver may suffer from intra-cell and inter-cell interferences [4, 5].

Much work has been done to investigate the methodologies that combat these two types of interference in cellular networks [6–12]. To mitigate inter-cell interference, recently, the attention of researchers has been drawn to a relatively new technique, namely interference alignment (IA) [13, 14]. The IA technique refers to the construction of signals such that the resulting interference signal lies in a subspace that is orthogonal to the one spanned by the signal of interest at each receiver [15]. Therefore, IA can be adopted to enhance the cell-edge user throughput in a cellular network. A lot of attention has been given to IA techniques [10,11,16] which show how to mitigate inter-cell interference in cellular uplink and downlink networks. However, these schemes usually

assume global perfect channel state information (CSI) available at the transmitter which is usually unrealistic: Global perfect CSI requires significant feedback/control overhead. On the other hand, unlike point-to-point MIMO systems where the imperfect CSI causes only a signal-to-noise ratio (SNR) offset in the capacity vs. SNR curve, the accuracy of the CSI in the IA systems affects the slope of the curve, i.e., the degree of freedom (DoF) [17]. Given the difficulty of achieving near perfect CSI at the transmitters in reality, there has been much work aimed at relaxing this assumption. An analysis on the effect of imperfect CSI on the mutual information of the underlying IA scheme is provided in [18]. IA with analog feedback is considered in [19] and the performance degradation is investigated. In [20], a method that reduces the quantization error with respect to the classical scheme is introduced. This method involves a computationally heavy iterative algorithm which must be run for each codeword and for each channel realization. There are several existing works introducing IA with limited feedback to cellular networks. In [21–24], the authors considered a multi-cell MIMO interfering uplink network where the BS is serving as the receiver. Ideal receiver CSI is assumed at the BS so that IA is used to generate transmit beam-forming vectors. The generated transmit beam-forming vectors are then fed back to MSs (transmitters) using codebook-based feedback.

In this research proposal and in order to complete tasks one and two, we consider a downlink multi-cell MIMO network and aim to design the beam-forming matrices based on both the global CSI and the limited feedback. To be specific, assuming each user has the ideal receiver CSI, while in the latter one receiver feeds back the limited CSI to the BS, in the former one receiver utilizes the global CSI. Moreover, in the limited feedback case, upon reception of the limited CSI feedback from MSs, the BS utilizes IA-based precoding to minimize the residual inter-cell interference. The proposed scheme takes into account the effects of finite-rate codebook based feedback adopted in the LTE-Advanced system.

Among many CCI suppression schemes that mitigate the intra-cell interference, linear precoding gains popularity on account of its simplicity of implementation and good performance. To design the optimal linear precoding scheme, it is often desirable to maximize the output signal-

to-interference-plus-noise ratio (SINR) for each user. However, this problem is known to be challenging due to its coupled nature and unavailability of closed-form solutions. A more tractable but suboptimal design is to enforce a zero-CCI requirement for each user, such as the zero-forcing (ZF) beam-forming [7, 8]. However, the sum-rate of ZF beam-forming under codebook-based feedback usually suffers significantly from the feedback error [25]. In [9], the authors introduced the concept of signal-to-leakage-and-noise ratio (SLNR) as the optimization metric for a linear precoder design. This metric transforms a coupled optimization problem into a completely decoupled one, for which a closed-form solution is available. To further improve the system capacity, a coordinated beam-forming (CB) algorithm is proposed in [6] that jointly optimized the transmit and receive beam-forming through iteration. In [12], a low complexity leakage-based CB scheme under the LTE-Advanced feedback framework is introduced and shown a better performance than ZF-based CB under the same feedback overhead. However, these work did not consider the existence of quantization errors.

This research proposal differs from previously mentioned studies particularly in its aim to improve both the cell-edge and the cell-average user spectral-efficiency by introducing a novel protocol to mitigate the co-channel interference issue. Furthermore, this work also takes into account the effects of finite-rate codebook based feedback adopted in the frequency division duplex (FDD) LTE-Advanced system. Main contributions of this research proposal can be summarized as follows.

- A limited feedback-based IA scheme along with a novel low complexity iterative leakage-based coordinated beam-forming strategy is introduced as a promising interference management technique to achieve the minimum overall residual co-channel interference in a downlink multi-cell multi-user LTE-Advanced network. The introduced algorithm efficiently reduces the effect of inter-cell interference from BSs in other cells, and eliminates the intra-cell interference due to the spatial streams dedicated to other users in the same cell.

- A limited feedback strategy based on random vector quantization feedback is introduced. The introduced scheme is capable of minimizing the overall residual CCI for each cell under

the given codebook. In addition, it can be implemented with flexible antenna configurations.

- Optimal closed-form expressions for the transmit and receive beam-forming matrices are characterized under quantization errors due to the finite-rate codebook-based feedback. The optimized transceiver design can be used to effectively mitigate the CCI under given codebooks.

- Evaluation is conducted to show that the introduced scheme significantly outperforms existing interference management strategies in terms of sum-rate in realistic cellular environments making it a promising candidate for LTE-Advanced networks.

2.2 System Setup

2.2.1 System Model and Assumption

As illustrated in Fig. 2.1, a G-cell interfering broadcast channel is considered, where the i-th BS is equipped with N_t transmit antennas and serves I_i users in cell i. A cooperating set is formed by aggregating neighboring BSs that transmission strategies for the BSs within this set are coordinated to effectively mitigate the interference at the MSs. Let us define i_k to be the kth user in the ith cell and N_r be the number of receive antennas at receiver i_k. Throughout this research proposal we consider an arbitrary but fixed power allocation among all users. Although for simplicity of exposition, we consider the case where every transmitter-receiver pair is equipped with the same numbers of antennas, the results can be readily generalized to a network with different numbers of antennas as long as the IA remains feasible [26]. Let us also define $\mathscr{I}$ to be the set of all receivers, i.e.,

$$\mathscr{I} = \{i_k | i \in \mathscr{G} \triangleq \{1,\ldots,G\}, k \in \{1,2,\ldots,I_i\}\}. \tag{2.1}$$

It is assumed that each co-scheduled user operating in the MU-MIMO mode only receives one

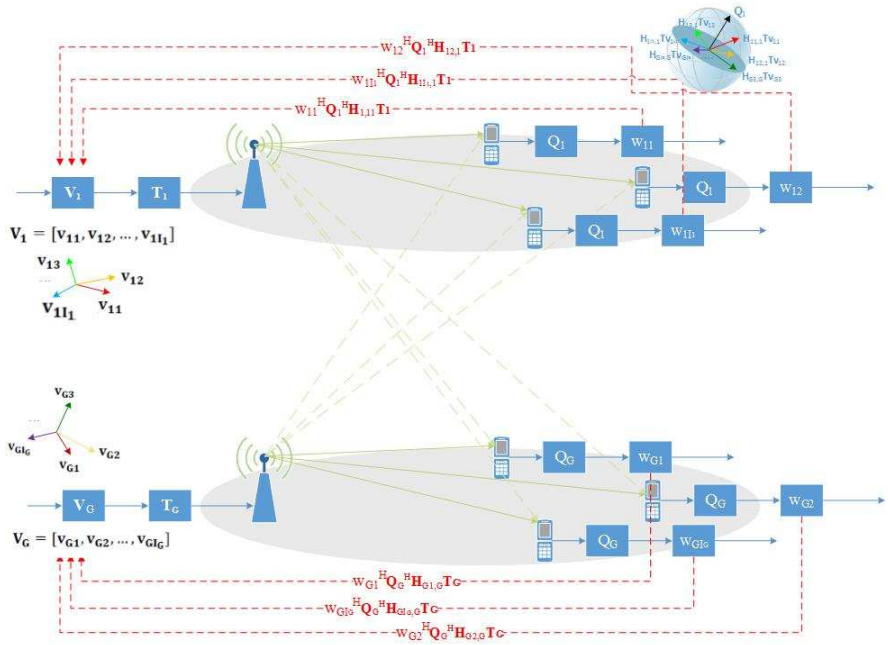

Figure 2.1. G-cell multi-user MIMO downlink system.

spatial stream (rank 1 transmission) as specified by the Rel-10 LTE-Advanced standard [27, Chapter 11]. The channel matrix from the jth BS to the kth user in cell i is denoted by $\mathbf{H}_{i_k,j} \in \mathbb{C}^{N_r \times N_t}$ where $j \in \mathscr{G}$ and $i_k \in \mathscr{I}$. A spatially uncorrelated flat Rayleigh fading channel is assumed. The elements of $\mathbf{H}_{i_k,j}$ are modeled as i.i.d. complex Gaussian variables with zero-mean and unit-variance. It is important to note that the results obtained in this chapter can be generalized to any statistical channel models. Moreover, we assume that the channels are constant over a few time slots with respect to channel estimation and CSI feedback procedures. Frequency-division duplexing mode is assumed in this system. We also assume perfect time and frequency synchronization when expressing received baseband signals. Each BS, say $i \in \mathscr{G}$, plans to communicate a symbol vector $\mathbf{s}_i = [s_{i_1}, \ldots, s_{i_{I_i}}]^T \in \mathbb{C}^{I_i \times 1}$ to its associated receivers, where s_{i_k} is the transmit symbol from the i-th BS to the i_k receiver with unit power of $\mathbb{E}\{|s_{i_k}|^2\} = 1$. Note that the achievable DoF in each cell can be perceived as the number of signal space dimensions free of interference and I_i represents the pursued DoF in the i-th cell, which is the number of symbols that the i-th BS wishes to transmit. Prior to transmitting, the i-th BS, $i \in \mathscr{G}$, linearly precodes its symbol vector $\mathbf{x}_i = \sum_{k=1}^{I_i} \sqrt{\rho_{i_k}} \mathbf{f}_{i_k} s_{i_k}$ where ρ_{i_k} stands for the transmit power for user i_k; $\mathbf{f}_{i_k}$ denotes the beam-former

that BS i uses to transmit the signal s_{i_k} to receiver i_k; and $\mathbf{F}_i = [\mathbf{f}_{i_1}, \ldots, \mathbf{f}_{i_{I_i}}] \in \mathbb{C}^{N_t \times I_i}$ indicates the transmit precoder at BS i with unit-norm columns. Hence, the transmit power at BS i is computed as $p_i = \mathbb{E}\{\|\mathbf{x}_i\|^2\} = \sum_{k=1}^{I_i} \rho_{i_k} \mathbb{E}\{\|\mathbf{f}_{i_k}\|^2\} = \sum_{k=1}^{I_i} \rho_{i_k}$. Accordingly, the received signal vector $\mathbf{y}_{i_k} \in \mathbb{C}^{N_r \times 1}$ at receiver $i_k \in \mathscr{I}$ can be written as:

$$\mathbf{y}_{i_k} = \underbrace{\sqrt{\rho_{i_k}}\mathbf{H}_{i_k,i}\mathbf{f}_{i_k}s_{i_k}}_{\text{desired signal}} + \underbrace{\sum_{m=1,m\neq k}^{I_i} \sqrt{\rho_{i_m}}\mathbf{H}_{i_k,i}\mathbf{f}_{i_m}s_{i_m}}_{\text{intra-cell interference}} + \underbrace{\sum_{j=1,j\neq i}^{G}\sum_{\ell=1}^{I_j} \sqrt{\rho_{j_\ell}}\mathbf{H}_{i_k,j}\mathbf{f}_{j_\ell}s_{j_\ell}}_{\text{inter-cell interference}} + \mathbf{n}_{i_k}, \tag{2.2}$$

where $\mathbf{n}_{i_k} \in \mathbb{C}^{N_r \times 1}$ represents the additive white Gaussian noise (AWGN) at receiver i_k with $\mathbb{E}\{\mathbf{n}_{i_k}\mathbf{n}_{i_k}^H\} = \sigma_{i_k}^2\mathbf{I}_{N_r}$. We assume that the signals for different users are independent from each other. In this chapter, we treat interference as noise and consider linear receive beam-forming strategy so that the estimated signal is given by $\hat{s}_{i_k} = \mathbf{u}_{i_k}^H\mathbf{y}_{i_k}$. Indeed, each receiver $i_k \in \mathscr{I}$, linearly processes the received signal to obtain $\mathbf{u}_{i_k}^H\mathbf{y}_{i_k}$ where $\mathbf{u}_{i_k} \in \mathbb{C}^{N_r \times 1}$ denotes the unit-norm post processing filter at receiver i_k, i.e., $\|\mathbf{u}_{i_k}\|^2 = 1$. Thus, after receive beam-forming the received signal at receiver i_k, $\forall i_k \in \mathscr{I}$, can be expressed as

$$\hat{s}_{i_k} = \mathbf{u}_{i_k}^H\mathbf{H}_{i_k,i}\mathbf{F}_i\mathbf{P}_i^{\frac{1}{2}}\mathbf{s}_i + \sum_{j=1,j\neq i}^{G} \mathbf{u}_{i_k}^H\mathbf{H}_{i_k,j}\mathbf{F}_j\mathbf{P}_j^{\frac{1}{2}}\mathbf{s}_j + \mathbf{u}_{i_k}^H\mathbf{n}_{i_k}, \tag{2.3}$$

where $\mathbf{P}_i$ is an $I_i \times I_i$ matrix with main diagonal $(\rho_{i_1}, \rho_{i_2}, \ldots, \rho_{i_{I_i}})$ and zeros on its off-diagonal.

2.2.2 Constructing Transmit And Receive Beam-forming

Inspired by our previous work [16], we use two cascaded beam-forming matrices to construct our precoder at each BS. The precoding matrix at each BS, is denoted by $\mathbf{F}_i$ for BS i, is composed of the product of a cell-specific beam-forming matrix, represented by $\mathbf{T}_i = [\mathbf{t}_{i_1}, \ldots, \mathbf{t}_{i_S}] \in \mathbb{C}^{N_t \times S}$, and a user-specific beam-forming matrix, denoted by $\mathbf{V}_i = [\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_{I_i}}] \in \mathbb{C}^{S \times I_i}$, where S is the number of virtual streams which is also equal to the number of co-scheduled users in each cell. Furthermore, S is upper bounded by $\min(N_t, N_r)$. $\mathbf{T}_i$ is used to mitigate the inter-cell interference while $\mathbf{V}_i$ will be utilized to mitigate the intra-cell interference. It is worth mentioning that the

$$\mathrm{SLNR}_{i_k} = \frac{\mathrm{Tr}\left(\mathbf{v}_{i_k}^H \mathbf{T}_i^H \mathbf{H}_{i_k,i}^H \mathbf{H}_{i_k,i} \mathbf{T}_i \mathbf{v}_{i_k}\right)}{\frac{\sigma_{i_k}^2}{\rho_{i_k}} + \mathrm{Tr}\left(\mathbf{v}_{i_k}^H \mathbf{T}_i^H \left\{ \sum_{m=1,m\neq k}^{I_i} \mathbf{H}_{i_m,i}^H \mathbf{H}_{i_m,i} + \sum_{j=1,j\neq i}^{G} \sum_{\ell=1}^{I_j} \mathbf{H}_{j_\ell,i}^H \mathbf{H}_{j_\ell,i} \right\} \mathbf{T}_i \mathbf{v}_{i_k}\right)}. \tag{2.5}$$

cell-specific beam-forming matrix $\mathbf{T}_i$ is independent of channel gains. Moreover, we use two cascaded receive beam-forming matrices to construct our post-precoder at each receiver. The post-precoding vector $\mathbf{u}_{i_k} \in \mathbb{C}^{N_r \times 1}$ at receiver i_k is composed of $\mathbf{Q}_i \mathbf{w}_{i_k}$, where $\mathbf{Q}_i \in \mathbb{C}^{N_r \times (N_r - S)}$ and $\mathbf{w}_{i_k} \in \mathbb{C}^{(N_r - S) \times 1}$. Similarly, the receive beam-forming matrix $\mathbf{Q}_i$ will mitigate the inter-cell interference. It is the role of the receive beam-forming vector $\mathbf{w}_{i_k}$ to alleviate the intra-cell interference. It is worth mentioning that, when a cascaded precoder is utilized, since both $\mathbf{T}_i$ and $\mathbf{Q}_i$ are cell-specific and fixed during the transmission, the additional complexity introduced at both transmitters and receivers is rather small. Note that, when designing the cell-specific IA-based precoding matrices $\mathbf{T}_i$ and $\mathbf{Q}_i$, the condition $S \leq \min(N_t, N_r)$ should be satisfied. It is important to note that N_t and N_r in modern cellular networks can be quite large. For example, Rel-10 LTE-Advanced networks support $N_t = N_r = 8$ while 64 antenna elements are being considered at both base stations and mobile stations for 5G networks [28].

Remark 1. *Since the user specific transmit beam-forming $\mathbf{V}_i$ and receive beam-forming $\mathbf{w}_{i_k}$ are utilized only to mitigate the intra-cell interference, $\mathbf{V}_i$ and $\mathbf{w}_{i_k}$ do not change the inter-cell interference level at each user.*

The received signal at user k in cell i can be rewritten as

$$\hat{s}_{i_k} = \mathbf{w}_{i_k}^H \mathbf{H}_{i_k,i}^{\mathrm{eff}} \mathbf{V}_i \mathbf{P}_i^{\frac{1}{2}} \mathbf{s}_i + \sum_{j=1,j\neq i}^{G} \mathbf{w}_{i_k}^H \mathbf{Q}_i^H \mathbf{H}_{i_k,j} \mathbf{T}_j \mathbf{V}_j \mathbf{P}_j^{\frac{1}{2}} \mathbf{s}_j + \hat{n}_{i_k},\ \forall i_k \in \mathscr{I}, \tag{2.4}$$

where $\mathbf{H}_{i_k,i}^{\mathrm{eff}} \in \mathbb{C}^{(N_r - S) \times S}$ is the effective channel matrix defined as $\mathbf{H}_{i_k,i}^{\mathrm{eff}} \triangleq \mathbf{Q}_i^H \mathbf{H}_{i_k,i} \mathbf{T}_i$, and $\hat{n}_{i_k} \triangleq \mathbf{u}_{i_k}^H \mathbf{n}_{i_k}$. Recall that the SLNR is defined as the ratio of the received signal power at the desired user to the leakage power plus the noise power, the SLNR for serving user i_k can be written as (2.5).

In what follows, we assume that each receiver $i_k \in \mathscr{I}$ knows its channels $\mathbf{H}_{i_k,j}$ perfectly based

on pilot signals transmitted by each of G cells. In Section 2.3.1, error-free dedicated broadcast links are assumed from each receiver to the other transmitters in the network. Hence, each transmitter knows the perfect CSI. In Section 2.3.2 we generalize results to a more realistic case where each transmitter does not have the perfect CSI, rather, each transmitter must rely on the quantized feedback from the receivers to obtain the CSI. To be specific, during the channel feedback phase, receiver i_k feeds back its CSI using B bits based codebook quantization.

2.2.3 Synthesized Channel Model And Feedback Framework

In FDD LTE/LTE-Advanced systems, each MS measures the downlink MIMO channel through the reference signals/pilots, and then feeds back the CSI to the BS using codebook-based channel feedback [29]. The CSI feedback in LTE/LTE-Advanced systems usually contains two information: the channel direction information (CDI) and the channel quality indicator (CQI). The CDI is related to the eigen-directions of the underlying MIMO channel which contains both rank indicator (RI) and precoding matrix indicator (PMI). Effectively, RI and PMI jointly tells eigen-directions of the underlying MIMO channel. The CQI is related to the strength of the corresponding spatial directions. Effectively, CQI tells the singular values of the underlying channel. Codebook-based quantization is adopted in LTE/LTE-Advanced systems for the feedback of CDI to minimize the feedback overhead. For example, in Rel-10 LTE-Advanced systems, 4-bit PMI and 3-bit RI are used for feeding back the 8×8 MIMO channel [4].

In limited feedback systems such as LTE/LTE-Advanced systems, each user feeds back its CQI and CDI to the BS. CQI feedback is relatively straightforward where user i_k feeds back the quantized version of the singular values using scalar or vector quantization methods. CDI feedback is generally more involved relying on codebook-based feedback. After obtaining each of the eigen-direction of the MIMO channel, say $\tilde{\mathbf{h}}_{i_k,j}^{(\ell)} = \mathbf{h}_{i_k,j}^{(\ell)} / \|\mathbf{h}_{i_k,j}^{(\ell)}\|$ for the ℓth eigen-direction, $\forall \ell \in \{1,\ldots,N_t\}$, user i_k quantizes it to $\hat{\mathbf{h}}_{i_k,j}^{(\ell)}$ using a random vector quantization codebook $\mathscr{C}$ which is known at both the BS and the MS. In general, there will be multiple CDIs of the underlying MIMO channel, the exact number of the CDIs to be fed back will impact the system performance

and the feedback overhead. The indices of the quantized CDIs will be sent to the BS through a feedback link. Accordingly, the BS obtains $\hat{\mathbf{h}}_{i_k,j}^{(\ell)}$s and then uses these information for downlink MIMO precoding. A quantization codebook $\mathscr{C}$ consist of 2^β N_t-dimensional unit norm vectors is given by $\mathscr{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{2^\beta}\}$, where 2^β is the codebook size, β is the number of feedback bits per user, and $\mathbf{c}_m \in \mathbb{C}^{N_t \times 1}$ is a unit norm codeword, i.e., $\|\mathbf{c}_m\|^2 = 1$.

Each user chooses the CDI in the codebook that is closet to its eigen-direction where closeness is usually measured in terms of the angle between the eigen-direction and the codeword in the codebook or equivalently the inner product. Hence, user k in the i-th cell computes quantization index $q_{i_k,j}^{(\ell)}$ according to

$$q_{i_k,j}^{(\ell)} = \underset{m=1,\ldots,2^\beta}{\arg\max} |\langle \tilde{\mathbf{h}}_{i_k,j}^{(\ell)}, \mathbf{c}_m \rangle| = \underset{m=1,\ldots,2^\beta}{\arg\min} \sin^2(\angle(\tilde{\mathbf{h}}_{i_k,j}^{(\ell)}, \mathbf{c}_m)), \tag{2.6}$$

and feeds the index back to the BS. Upon reception of the index $q_{i_k,j}^{(\ell)}$, the BS can recover the CDI by searching in the corresponding entry in the codebook. In this way, the feedback overhead is significantly reduced with the penalty of quantization error in the finite rate feedback systems.

As in [25], the relationship between the full CDI, $\tilde{\mathbf{h}}_{i_k,j}^{(\ell)}$, and the quantized CDI, $\hat{\mathbf{h}}_{i_k,j}^{(\ell)}$, can be expressed by

$$\tilde{\mathbf{h}}_{i_k,j}^{(\ell)} = \sqrt{1 - z_{i_k,j}^{(\ell)}} \hat{\mathbf{h}}_{i_k,j}^{(\ell)} + \sqrt{z_{i_k,j}^{(\ell)}} \mathbf{r}_{i_k,j}^{(\ell)}, \tag{2.7}$$

where $z_{i_k,j}^{(\ell)} = 1 - |\tilde{\mathbf{h}}_{i_k,j}^{(\ell)^H} \hat{\mathbf{h}}_{i_k,j}^{(\ell)}|^2$ is distributed according to the quantization error distribution and is independent of $\mathbf{r}_{i_k,j}^{(\ell)} \in \mathbb{C}^{N_t \times 1}$ which represents a unit norm vector isotropically distributed in the null space of $\hat{\mathbf{h}}_{i_k,j}^{(\ell)}$. It is shown in [30] that $\mathbb{E}\{\mathbf{r}_{i_k,j}^{(\ell)}\} = \mathbf{0}$, $\mathbb{E}\{\mathbf{r}_{i_k,j}^{(\ell)} \mathbf{r}_{i_k,j}^{(\ell)^H}\} = 1/(N_t - 1)(\mathbf{I}_{N_t} - \hat{\mathbf{h}}_{i_k,j}^{(\ell)} \hat{\mathbf{h}}_{i_k}^{(\ell)^H})$, and $\mathbb{E}\{z_{i_k,j}^{(\ell)}\} = \delta(N_t - 1)/N_t$ where $\delta \triangleq 2^{-\beta/(N_t-1)}$. It is worth noting that feedback overhead of the system depends on 1) the number of CDIs to be fed back to each BS, and 2) the number of BSs in the cooperating set.

2.3 A Novel Algorithm For Combating Both Inter-cell and Intra-cell Interferences

This section describes how our introduced techniques cancel out the inter-cell and intra-cell interferences for the downlink multi-cell MU-MIMO network.

2.3.1 Transceiver Design With Perfect CSI

2.3.1.1 IA technique and combating inter-cell interference

IA serves as means for obtaining as many interference free dimensions for communication as possible, and in practice stands for designing the transmit and receive strategies for each transmitter-receiver pair of a wireless network. Under the assumption of perfect CSI at the BSs, one can achieve maximum multiplexing gain, or the maximum DoF by utilizing the IA techniques. We begin by reviewing the concept of IA by adapting the main results of [14] to our setup. This framework aims for the maximum inter-cell interference alignment, which stands for the maximum inter-cell interference suppression possible, when the signal space of each BS is required to span $d = \min\{S, N_r - S\}$ spatial dimensions worth of communication. The suggested insight for IA effectiveness is that the interference leakage from all other cells will be zero and such an algorithm will obtain the optimal interference cancellation. This implies that the receive beamforing matrix $\mathbf{Q}_i$ is chosen in the null space of $\sum_{j=1, j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j$ such that $\mathbf{Q}_i^H \sum_{j=1, j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j = \mathbf{0}$. Hence, the inter-cell interference signals will be nulled out using linear processing of the received signal at each receiver, when the useful signal space spans d dimensions, so that IA is perfectly attained. Recalling (2.4), the perfect IA requires

$$\mathbf{Q}_i^H \sum_{j=1, j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j = \mathbf{0}, \tag{2.8}$$

$$\text{Rank}\left(\mathbf{Q}_i^H \mathbf{H}_{i_k,i}\mathbf{T}_i\right) = d. \tag{2.9}$$

The first equation guarantees that all the interfering signals in cell i lie in the subspace orthogonal to $\mathbf{Q}_i$, while the second one ensures that the signal subspace has a full rank dimension and is linearly independent of the interference subspace. We assume that all the elements of the channel matrices are randomly and independently generated from continuous distributions. So if $\mathbf{Q}_i$ and $\mathbf{T}_j$ can be found to satisfy condition (2.8), condition (9) will also be satisfied with probability 1 [31].

Suppose that each BS independently generates $\mathbf{t}_{i_\ell}$ according to the isotropic distribution over the N_t-dimensional unit sphere. Specifically, $\mathbf{t}_{i_\ell} \in \mathbb{C}^{N_t \times 1}$ is an orthonormal basis for $i \in \mathscr{G}$ and $\ell \in \{1,...,S\}$. If the reference beam-forming matrix is generated in a pseudo-random fashion, BSs do not need to broadcast them to users. Then, user i_k obtains $\mathbf{T}_j$, $j \in \mathscr{G}$. In order to cancel the inter-cell interference, we need to design the receive matrix $\mathbf{Q}_i$ such that $\mathbf{Q}_i$ is in the null space of $\sum_{j=1,j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j$. Now this raises the question of how we can find the null space of the above interference channel. In this regards, the k-th user in the i-th cell estimates the inter-cell interference $\sum_{j=1,j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j$ using pilots or a preamble. It then generates a null matrix $\mathbf{Q}_i$ such that (2.8) is satisfied. It is worth mentioning that since the $\sum_{j=1,j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j$ is of dimension $N_r \times S$, the dimension of the null space will be $N_r - S$, thus a matrix $\mathbf{Q}_i$ always exists, in which columns of this matrix are the orthonormal basis of the null signal spaces. Consequently, $\mathbf{T}_j$ causes no interference to receiver i_k by completely removing the inter-cell interference term.

Determining the feasibility of a linear IA solution is a key step. The following theorem provides such a condition.

Theorem 1. *Consider the G-cell network where the i-th BS is equipped with $N_t^{(i)}$ transmit antennas, serves I_i users, and the k-th user in this cell is equipped with $N_r^{(i_k)}$ receive antennas. The IA is feasible if $\sum_{i=1}^{G} N_t^{(i)} > GS$ where S is the number of co-schedule users. In the special case where $N_t^{(i)} = N_t$ and $N_r^{(i_k)} = N_r$, IA is feasible if $N_t > S$.*

Proof. See Appendix A.0.1. □

Applying the optimal cell-specific matrices $\mathbf{T}_i$ and $\mathbf{Q}_i$, the inter-cell interference at cell i can be

eliminated completely. Accordingly, the received symbol at the i_k user in (2.4) can be expressed as

$$\hat{s}_{i_k} = \sqrt{\rho_{i_k}}\mathbf{w}_{i_k}^H \mathbf{H}_{i_k,i}^{\text{eff}} \mathbf{v}_{i_k} s_{i_k} + \sum_{m=1,m\neq k}^{I_i} \sqrt{\rho_{i_m}}\mathbf{w}_{i_k}^H \mathbf{H}_{i_k,i}^{\text{eff}} \mathbf{v}_{i_m} s_{i_m} + \hat{n}_{i_k}, \tag{2.10}$$

where the first term represents the desired signal at the i_k user and the second term is the intra-cell interference. In the following, we will discuss how to cancel out the intra-cell interference by designing the user-specific beam-forming matrix $\mathbf{V}_i$ and receive beam-forming vectors $\mathbf{w}_{i_k}$.

2.3.1.2 Leakage-based iterative CB scheme and intra-cell interference mitigation

The SLNR expression for serving the i_k user, given in Eq. (2.5), can be written as

$$\text{SLNR}_{i_k} = \frac{\text{Tr}(\mathbf{v}_{i_k}^H \mathbf{H}_{i_k,i}^{\text{eff}^H} \mathbf{H}_{i_k,i}^{\text{eff}} \mathbf{v}_{i_k})}{\text{Tr}(\mathbf{v}_{i_k}^H \left(\frac{\sigma_{i_k}^2}{\rho_{i_k}} \mathbf{I}_S + \bar{\mathbf{H}}_{i_k,i}^H \bar{\mathbf{H}}_{i_k,i} \right) \mathbf{v}_{i_k})}, \tag{2.11}$$

where $\bar{\mathbf{H}}_{i_k,i} \triangleq [\mathbf{H}_{i_1,i}^{\text{eff}^H}, \ldots, \mathbf{H}_{i_{k-1},i}^{\text{eff}^H}, \mathbf{H}_{i_{k+1},i}^{\text{eff}^H}, \ldots, \mathbf{H}_{i_{I_i},i}^{\text{eff}^H}]^H$, $i_k \in \mathscr{I}$, represents the corresponding concatenated leakage channel. Using the concept of leakage, we can formulate an optimization problem which mitigates the total intra-cell interfering power that user i_k generates in cell i. The optimization problem can be formulated as

$$\mathbf{v}_{i_k}^{\text{opt}} = \underset{\mathbf{v}_{i_k} \in \mathbb{C}^{\text{S}\times 1}}{\arg\max} \text{SLNR}_{i_k}, \ \forall i_k \in \mathscr{I} \tag{2.12}$$

subject to $\|\mathbf{v}_{i_k}\|^2 = 1$. It was shown in [9] that the solution is given by

$$\mathbf{v}_{i_k}^{\text{opt}} = \nu_{\max} \left\{ \left(\frac{\sigma_{i_k}^2}{\rho_{i_k}} \mathbf{I}_S + \bar{\mathbf{H}}_{i_k,i}^H \bar{\mathbf{H}}_{i_k,i} \right)^{-1} \mathbf{H}_{i_k,i}^{\text{eff}^H} \mathbf{H}_{i_k,i}^{\text{eff}} \right\} \tag{2.13}$$

Algorithm 1: Introduced Algorithm with Perfect CSI

(S.1) : Each BS independently generates the columns of $\mathbf{T}_i,\ i \in \mathscr{G}$ according to the isotropic distribution over the N_t-dimensional unit sphere.
(S.2) : Design the receive matrix $\mathbf{Q}_i$ in such a way that it lies in the null space of $\sum_{j=1, j\neq i}^{G} \mathbf{H}_{i_k,j}\mathbf{T}_j$.
(S.3) : Compute the optimal transmit beam-forming vector $\mathbf{v}_{i_k}^{\text{opt}}$ based on Eq. (2.13).
(S.4) : Update the receive beam-forming vector $\mathbf{w}_{i_k}$ based on Eq. (2.14).
(S.5) : If $\mathbf{V}_i$ satisfy a termination criterion: STOP.

where $v_{\max}(\mathbf{A})$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{A}$. The optimal minimum mean square error (MMSE) receive beam-forming vector becomes

$$\mathbf{w}_{i_k}^{\text{opt}} = \mathbf{J}_{i_k}^{-1}\mathbf{H}_{i_k,i}^{\text{eff}}\mathbf{v}_{i_k}^{\text{opt}} \tag{2.14}$$

where $\mathbf{J}_{i_k} = \sum_{\ell=1}^{I_i} \rho_{i_\ell}\mathbf{H}_{i_k,i}^{\text{eff}}\mathbf{v}_{i_\ell}^{\text{opt}}\mathbf{v}_{i_\ell}^{\text{opt}^H}\mathbf{H}_{i_k,i}^{\text{eff}^H} + \sigma_{i_k}^2\mathbf{I}_{N_r-S}$ is the covariance matrix of the received signal at receiver i_k. Algorithm 1 summarized the introduced scheme in the present of perfect CSI for a multi-cell MU-MIMO network.

The sum rate can be expressed as

$$R_{\text{sum}} = \sum_{i=1}^{G}\sum_{k=1}^{I_i} \log_2\left(1 + \frac{\rho_{i_k}|\mathbf{w}_{i_k}^{\text{opt}^H}\mathbf{H}_{i_k,i}^{\text{eff}}\mathbf{v}_{i_k}^{\text{opt}}|^2}{\sigma_{i_k}^2 + \hat{\mathscr{I}}_{i_k}}\right) \tag{2.15}$$

where $\hat{\mathscr{I}}_{i_k}$ is the total multi-user interference given by $\hat{\mathscr{I}}_{i_k} = \sum_{m=1, m\neq k}^{I_i} \rho_{i_m}|\mathbf{w}_{i_k}^{\text{opt}^H}\mathbf{H}_{i_k,i}^{\text{eff}}\mathbf{v}_{i_m}^{\text{opt}}|^2$.

It is important to note that the above introduced method requires perfect and global CSI, an assumption that is clearly unrealistic: obtaining full CSI for all links would inevitably introduce a large amount of control/feedback overhead. In the following, we consider a more realistic scenario where the CSI is fed back using limited feedback overhead. Since, it is impossible to achieve zero interference with limited CSI feedback, a performance loss is inevitable. To address this issue we introduce a novel transceiver design scheme which complies with the limited feedback mechanism introduced in LTE-Advanced systems.

2.3.2 Transceiver Design With Limited Feedback

2.3.2.1 Limited feedback IA Scheme and combating inter-cell interference

As mentioned in the previous section, IA can only be achieved with perfect CSI at the BSs. In the case of limited feedback, only a quantized version of channel matrices, namely $\hat{\mathbf{H}}_{i_k,j}$, are available at the BSs, where each user uses B bits to quantize $\mathbf{H}_{i_k,j}$ to $\hat{\mathbf{H}}_{i_k,j}$. We can then perform naive IA where $\hat{\mathbf{H}}_{i_k,j}$ are treated as the true channels while performing IA. To distinguish these beam-forming matrices from those selected with perfect CSI at the BSs, we denote these cell-specific beam-forming matrices as $\hat{\mathbf{T}}_j$ and $\hat{\mathbf{Q}}_i$ so that each of these matrices are chosen in such a way that $\hat{\mathbf{Q}}_i^H \sum_{j=1,j\neq i}^{G} \hat{\mathbf{H}}_{i_k,j} \hat{\mathbf{T}}_j = \mathbf{0}$. Clearly, $\hat{\mathbf{Q}}_i^H \sum_{j=1,j\neq i}^{G} \mathbf{H}_{i_k,j} \hat{\mathbf{T}}_j \neq \mathbf{0}$ leading to residual interference. That is, due to the limited CSI feedback, the IA transceiver cannot achieve perfect alignment. Thus, there will be some residual interference which may lead to a sum-rate loss.

Denoted the total interference leakage at the i_k user due to all undesired BSs as $\mathscr{L}_{i_k}$, we have

$$\mathscr{L}_{i_k}(\hat{\mathbf{Q}}_i, \hat{\mathbf{T}}_j) = \mathrm{Tr}\left\{\hat{\mathbf{Q}}_i^H \mathbf{R}_{i_k} \hat{\mathbf{Q}}_i\right\} \tag{2.16}$$

where $\mathbf{R}_{i_k} = \sum_{j=1,j\neq i}^{G} \mathbf{H}_{i_k,j} \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \mathbf{H}_{i_k,j}^H$ is the interference covariance matrix at the i_k user.

Given $\hat{\mathbf{T}}_j$, $\hat{\mathbf{Q}}_i$ can be optimized to minimize $\mathscr{L}_{i_k}(\hat{\mathbf{Q}}_i, \hat{\mathbf{T}}_j)$ for an improved system performance. Towards this end, let $\mathbf{H}_{i_k,j} = \mathbf{U}_{\mathbf{H}_{i_k,j}} \Lambda_{\mathbf{H}_{i_k,j}} \mathbf{V}_{\mathbf{H}_{i_k,j}}^H$ be a singular value decomposition (SVD) of the $N_r \times N_t$ channel matrix $\mathbf{H}_{i_k,j}$, with $\mathbf{U}_{\mathbf{H}_{i_k,j}}$ and $\mathbf{V}_{\mathbf{H}_{i_k,j}} \triangleq [\mathbf{v}_{i_k,j}^{(1)}, \ldots, \mathbf{v}_{i_k,j}^{(N_t)}]$ being two unitary matrices with dimension $N_r \times N_r$ and $N_t \times N_t$, respectively, and $\Lambda_{\mathbf{H}_{i_k,j}}$ being an $N_r \times N_t$ matrix with main diagonal $(\lambda_{\mathbf{H}_{i_k,j}}^{(1)}, \ldots, \lambda_{\mathbf{H}_{i_k,j}}^{(\tilde{d})})$ and zeros on its off-diagonal, where $\tilde{d} \triangleq \min(N_t, N_r)$. Since $\mathbf{U}_{\mathbf{H}_{i_k,j}}$ is unitary, without loss of generality, we can assume that the receive beam-forming matrix at user i_k is in the form of $\hat{\mathbf{Q}}_i = \mathbf{U}_{\mathbf{H}_{i_k,j}} \mathbf{W}_i$. Then, the residual interference at user i_k can be expressed as

$$\mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j) = \mathrm{Tr}\left\{\mathbf{W}_i^H \tilde{\mathbf{R}}_{i_k} \mathbf{W}_i\right\}, \tag{2.17}$$

where $\tilde{\mathbf{R}}_{i_k} = \sum_{j=1,j\neq i}^{G} \Lambda_{\mathbf{H}_{i_k,j}} \mathbf{V}_{\mathbf{H}_{i_k,j}}^H \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \mathbf{V}_{\mathbf{H}_{i_k,j}} \Lambda_{\mathbf{H}_{i_k,j}}^H$.

It is clear that the j-th BS only needs to know the quantized version of CDI, $\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}$, and CQI, $\hat{\Lambda}_{\mathbf{H}_{i_k,j}}$, to obtain the transmit and receive beam-formers. To this end, user i_k feedbacks the quantized version of singular values using scalar or vector quantization method while employs a codebook-based feedback to feed back the CDI to the BSs. In this regard, after obtaining each of the eigen-direction of the MIMO channel, say $\tilde{\mathbf{v}}_{i_k,j}^{(\ell)}$, user i_k quantizes it to $\hat{\mathbf{v}}_{i_k,j}^{(\ell)}$ using a random vector quantization codebook $\mathscr{C}$, as discussed in Section II-C. In the following, we show that the subspace of the perfect channel matrix can be decomposed as the weighted sum of two spaces.

Lemma 1. *The quantization $\hat{\Lambda}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H$ of the perfect channel $\Lambda_{\mathbf{H}_{i_k,j}}\mathbf{V}_{\mathbf{H}_{i_k,j}}^H$ follows the following decomposition:*

$$\tilde{\Lambda}_{\mathbf{H}_{i_k,j}}\tilde{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H \triangleq \mathbf{A}_{\mathbf{H}_{i_k,j}}\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,j}} + \mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}} \tag{2.18}$$

where $\tilde{\Lambda}_{\mathbf{H}_{i_k,j}}\tilde{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H \in \mathbb{C}^{N_r\times N_t}$ is an orthonormal basis for the subspace spanned by the columns of $\Lambda_{\mathbf{H}_{i_k,j}}\mathbf{V}_{\mathbf{H}_{i_k,j}}^H$; $\mathbf{A}_{\mathbf{H}_{i_k,j}} \triangleq \hat{\blacksquare}_{\mathbf{H}_{i_k,j}}(\mathbf{I}_{N_t} - \mathbf{Z}_{\mathbf{H}_{i_k,j}})^{1/2}$; $\mathbf{B}_{\mathbf{H}_{i_k,j}} \triangleq \hat{\blacksquare}_{\mathbf{H}_{i_k,j}}\mathbf{Z}_{\mathbf{H}_{i_k,j}}^{1/2}$; $\mathbf{Z}_{\mathbf{H}_{i_k,j}}$ is a $N_t \times N_t$ matrix with main diagonal $(z_{i_k,j}^{(1)},\ldots,z_{i_k,j}^{(N_t)})$ and represents the quantization error and satisfies $Tr\left\{\mathbf{Z}_{\mathbf{H}_{i_k,j}}^H\mathbf{Z}_{\mathbf{H}_{i_k,j}}\right\} \triangleq \mathrm{d}^2(\mathbf{H}_{\mathrm{i_k,j}},\hat{\mathbf{H}}_{\mathrm{i_k,j}})$; and $\mathbf{R}_{\mathbf{H}_{i_k,j}} = [\mathbf{r}_{i_k,j}^{(1)},\ldots,\mathbf{r}_{i_k,j}^{(N_t)}]^H \in \mathbb{C}^{N_t\times N_t}$ is an orthonormal basis for an isotropically distributed (complex) N_t-dimensional plane in the null space of $\hat{\Lambda}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H$. The quantities $\mathbf{A}_{\mathbf{H}_{i_k,j}}$, $\mathbf{B}_{\mathbf{H}_{i_k,j}}$ and $\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}$ are distributed independent of each other, as are the pair $\mathbf{R}_{\mathbf{H}_{i_k,j}}$ and $\mathbf{Z}_{\mathbf{H}_{i_k,j}}$. The random matrices $\mathbf{A}_{\mathbf{H}_{i_k,j}}$, $\mathbf{B}_{\mathbf{H}_{i_k,j}}$, and $\mathbf{R}_{\mathbf{H}_{i_k,j}}$ have the following properties: $\mathbb{E}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}^H\mathbf{A}_{\mathbf{H}_{i_k,j}}\} = (N_t - (N_t-1)\delta)\mathbf{I}_{N_t}$, $\mathbb{E}\{\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\mathbf{B}_{\mathbf{H}_{i_k,j}}\} = (N_t-1)\delta\mathbf{I}_{N_t}$ and $\mathbb{E}\{\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\mathbf{R}_{\mathbf{H}_{i_k,j}}\} = \frac{N_t}{N_t-1}\mathbf{I}_{N_t} - \frac{1}{N_t-1}\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}$.

Proof. The result is immediate upon generalizing of the decomposition in [25] by using the fact that $\tilde{\mathbf{v}}_{i_k,j}^{(\ell)} = \mathbf{v}_{i_k,j}^{(\ell)}/||\mathbf{v}_{i_k,j}^{(\ell)}||$ and $\tilde{\mathbf{v}}_{i_k,j}^{(\ell)} = \sqrt{1-z_{i_k,j}^{(\ell)}}\hat{\mathbf{v}}_{i_k,j}^{(\ell)} + \sqrt{z_{i_k,j}^{(\ell)}}\mathbf{r}_{i_k,j}^{(\ell)}$, $\ell \in \{1,\ldots,N_t\}$. □

Based on Lemma 1, we can solve for inter-cell IA along with two concepts: 1) the minimization of the signal power leakage towards unintended users; 2) the maximization of the signal power within the desired signal sub-space in each cell. Recalling Eq. (2.5) and assuming $\|\mathbf{v}_{i_k}\|^2 = 1$, the SLNR for serving the k-th user in the i-th cell can be expressed as

$$\mathrm{SLNR}_{i_k} = \frac{\mathrm{Tr}\left(\hat{\mathbf{T}}_i^H \mathbf{H}_{i_k,i}^H \mathbf{H}_{i_k,i} \hat{\mathbf{T}}_i\right)}{\mathrm{Tr}\left(\hat{\mathbf{T}}_i^H (\frac{\sigma_{i_k}^2}{\rho_{i_k} S}\mathbf{I}_S + \sum_{j=1,j\neq i}^{G}\sum_{m=1}^{I_j} \mathbf{H}_{j_m,i}^H \mathbf{H}_{j_m,i})\hat{\mathbf{T}}_i\right)}, \tag{2.19}$$

Using SLNR as the metric, the cell-specific precoding matrix $\hat{\mathbf{T}}_i$ can be designed based on the following

$$\max_{\hat{\mathbf{T}}_i} \quad \sum_{k=1}^{S} \mathrm{SLNR}_{i_k}, \tag{2.20}$$

subject to $\sum_{k=1}^{I_i} \rho_{i_k} = p_i$ (the power constraint). Due to the limited feedback framework, it is reasonable to consider the beam-forming design via maximizing an expected SLNR averaging over all possible channel realizations. Hence, the problem of interest can be formulated as

$$\hat{\mathbf{T}}_i^{\mathrm{opt}} = \arg\max_{\hat{\mathbf{T}}_i \in \mathbb{C}^{\mathrm{N_t}\times \mathrm{S}}} \quad \sum_{k=1}^{S} \mathbb{E}_{\mathbf{H}}\{\mathrm{SLNR}_{i_k}\}. \tag{2.21}$$

It is not only difficult to derive a closed-form expression of $\mathbb{E}\{\mathrm{SLNR}_{i_k}\}$, but also hard to obtain a low-complexity algorithm to obtain the beam-forming matrices. To tackle this problem, instead of maximizing the expected value of SLNR, we maximize the lower bound of $\mathbb{E}\{\mathrm{SLNR}_{i_k}\}$.

Theorem 2. *The optimal precoders which are able to maximize the lower bound of the objective function in (2.21) can be obtained by extracting the leading* S *columns of the generalized eigenmatrix of the pair* $(\sum_{k=1}^{S} \mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{i_k,i}^H \mathbf{H}_{i_k,i}\right\}, (\frac{\sigma_{i_k}^2}{\rho_{i_k} S}\mathbf{I}_S + \sum_{j=1,j\neq i}^{G}\sum_{m=1}^{I_j} \mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{j_m,i}^H \mathbf{H}_{j_m,i}\right\}))$*, namely* $\mathbf{J}_i$*, as*

$$\hat{\mathbf{T}}_i^{opt} = \mu \mathbf{J}_i[\mathbf{I}_{\mathrm{S}}; \mathbf{0}], \tag{2.22}$$

where μ *is a scaling factor so that* $\sum_{k=1}^{I_i} \rho_{i_k} = p_i$.

Proof. See Appendix A.0.4. □

Given the optimal cell-specific precoding matrices, minimization of the total interference leak-

$$\mathbb{E}\{\tilde{\mathbf{R}}_{i_k}\} = \begin{cases} \left(\frac{1}{N_t-1}\sum_{m=1}^{N_t} \mathbf{e}_m^T \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \mathbf{e}_m \left(\mathbf{I}_{N_t} - \hat{\mathbf{v}}_{i_k,j}^{(m)} \hat{\mathbf{v}}_{i_k,j}^{(m)H}\right)\right) \odot \mathbf{G} + \left(\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\right) \odot \mathbf{E} & , \quad N_r = N_t \\ \left(\mathbf{L}_s\left(\frac{1}{N_t-1}\sum_{m=1}^{N_t} \mathbf{e}_m^T \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \mathbf{e}_m \left(\mathbf{I}_{N_t} - \hat{\mathbf{v}}_{i_k,j}^{(m)} \hat{\mathbf{v}}_{i_k,j}^{(m)H}\right)\right)\mathbf{L}_s^T\right) \odot \mathbf{G} + \left(\mathbf{L}_s \left(\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\right)\mathbf{L}_s^T\right) \odot \mathbf{E} & , \quad N_r < N_t \\ \begin{bmatrix} \left(\frac{1}{N_t-1}\sum_{m=1}^{N_t} \mathbf{e}_m^T \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \mathbf{e}_m \left(\mathbf{I}_{N_t} - \hat{\mathbf{v}}_{i_k,j}^{(m)} \hat{\mathbf{v}}_{i_k,j}^{(m)H}\right)\right) \odot \mathbf{G} + \left(\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}^H \hat{\mathbf{T}}_j \hat{\mathbf{T}}_j^H \hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\right) \odot \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} & , \quad N_r > N_t \end{cases} \tag{2.25}$$

age at each receiver can be formulated as follows

$$\min_{\mathbf{W}_i} \quad \mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j^{\text{opt}}). \tag{2.23}$$

Considering the impact of quantization errors on the precoder design, we can use $\mathbb{E}\{\mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j)\}$ as the new objective function to optimize $\mathbf{W}_i$. Mathematically, the beam-forming design problem can be formulated as

$$\mathbf{W}_i^{\text{opt}} = \underset{\mathbf{W}_i \in \mathbb{C}^{\mathrm{N_r} \times (\mathrm{N_r} - \mathrm{S})}}{\arg\min} \quad \mathbb{E}\{\mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j^{\text{opt}})\}, \tag{2.24}$$

where $\mathbb{E}\{\mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j)\} = \mathrm{Tr}\left\{\mathbf{W}_i^H \hat{\mathbf{R}}_{i_k} \mathbf{W}_i\right\}$ and $\hat{\mathbf{R}}_{i_k} = \mathbb{E}\{\tilde{\mathbf{R}}_{i_k}\}$. $\hat{\mathbf{R}}_{i_k}$ can be calculated using the following theorem.

Theorem 3. *Let* $\mathbf{G}$ *and* $\mathbf{E}$ *be matrices of dimensions* $\tilde{d} \times \tilde{d}$ *with* $\hat{\lambda}_{\mathbf{H}_{i_k,j}}^{(\ell)} \hat{\lambda}_{\mathbf{H}_{i_k,j}}^{(m)} \left(1 - \frac{N_t-1}{8N_t^2(1+N_t)}\right)^2$ *and* $\hat{\lambda}_{\mathbf{H}_{i_k,j}}^{(\ell)} \hat{\lambda}_{\mathbf{H}_{i_k,j}}^{(m)} \left(1 - \frac{1}{2}\delta\frac{N_t-1}{N_t} - \frac{1}{8}\delta^2\frac{N_t-1}{N_t+1} - \frac{1}{16}\delta^3\frac{N_t-1}{N_t+2}\right)^2$ *in the* ℓm*-th position, respectively. Then for a given* $\hat{\mathbf{T}}_j$, $\mathbb{E}\{\tilde{\mathbf{R}}_{i_k}\}$ *can be expressed as (2.25), at the top of the next page, where* $\mathbf{L}_s$ *is defined as* $\mathbf{L}_s \triangleq [\mathbf{I}_{N_r \times N_r}\ \mathbf{0}] \in \mathbb{C}^{N_r \times N_t}$.

Proof. See Appendix **??**. □

The optimization problem (2.24) is solved by deriving the structure of the optimal $\mathbf{W}_i$ with an optimal given $\hat{\mathbf{T}}_j$. In other words, the i_k user chooses its interference suppression filter $\mathbf{W}_i$ to minimize the leakage interference due to all undesired BSs. The $N_r - S$ dimensional received signal subspace that contains the least interference, is the space spanned by the eigenvectors corresponding to the $N_r - S$ smallest eigenvalues of the interference covariance matrix $\hat{\mathbf{R}}_{i_k}$. Thus, the ℓ-th

column of $\mathbf{W}_i$ is given by

$$\mathbf{W}_i^{\text{opt}}(:,\ell) = \nu_\ell(\hat{\mathbf{R}}_{i_k}), \tag{2.26}$$

where $\nu_\ell(\mathbf{X})$ is the eigenvector corresponding to the ℓ-th smallest eigenvalue of $\mathbf{X}$. Invoking $\mathbf{W}_i^{\text{opt}}$, $\mathbb{E}\{\mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j)\}$ can be reformulated as

$$\mathbb{E}\{\mathscr{L}_{i_k}(\mathbf{W}_i, \hat{\mathbf{T}}_j)\} = \rho_{N_r-S}(\hat{\mathbf{R}}_{i_k}), \tag{2.27}$$

$\rho_m(\mathbf{X})$ is the sum of m smallest eigenvalues of $\mathbf{X}$.

Remark 2. *It is worth mentioning that residual interference does not always lead to a loss in sum-rate. Therefore, there is no need to force it to be zero, especially in the power limited regime. To design the optimal MIMO precoding scheme, it is often desirable to maximize the received SINR for each user. However, it is an extremely challenging job to do so which requires joint operation from almost all aspects of MIMO communication including user-grouping, user-scheduling, resource allocation, and MIMO precoding across all coordinating BSs. In this chapter, we adopted a more tractable but suboptimal design to reduce the residual interference. Using this design criterion, we are able to analytically characterize the optimal cell-specific MIMO precoding strategies to mitigate inter-cell interference. Furthermore, the design of cell-specific MIMO precoding decreases the complexity of multi-cell coordination for mitigating inter-cell interference.*

The inter-cell interference can be minimized by using the optimal cell-specific transmit precoding matrix $\hat{\mathbf{T}}_j$ (which can be obtained via Theorem 2) and the optimal cell-specific receive beam-forming matrix $\mathbf{W}_i$ according to Eq. (2.26). Once we obtain the optimum cell-specific precoding and postprocessing matrices $\hat{\mathbf{T}}_i$ and $\hat{\mathbf{Q}}_i$, the inter-cell interference in cell i will be minimized. However, these matrices do not guarantee the suppression of intra-cell interference intended for the other users in the same cell i. Applying the optimal receive beam-forming matrix $\mathbf{Q}_i$ to the received

signal, the received symbol at the i_k user in (2.4) can be reformulated as

$$\hat{s}_{i_k} = \sqrt{\rho_{i_k}}\mathbf{w}_{i_k}^H \mathbf{H}_{i_k,i}^{\text{eff}} \mathbf{v}_{i_k} s_{i_k} + \sum_{m=1, m\neq k}^{I_i} \sqrt{\rho_{i_m}}\mathbf{w}_{i_k}^H \mathbf{H}_{i_k,i}^{\text{eff}} \mathbf{v}_{i_m} s_{i_m} + \tilde{n}_{i_k}, \tag{2.28}$$

where the first term represents the desired signal at useri_k , the second term is the intra-cell interference, and the last term is $\tilde{n}_{i_k} = \hat{n}_{i_k} + \rho_{N_r-S}(\hat{\mathbf{R}}_{i_k})$. Let $\mathbf{H}_{i_k,i}^{\text{eff}} = \mathbf{U}_{\mathbf{H}_{i_k,i}^{\text{eff}}} \Lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}} \mathbf{V}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^H$ be the SVD of the $(N_r - S) \times S$ effective channel matrix $\mathbf{H}_{i_k,i}^{\text{eff}}$, with $\mathbf{U}_{\mathbf{H}_{i_k,i}^{\text{eff}}}$ and $\mathbf{V}_{\mathbf{H}_{i_k,i}^{\text{eff}}} \triangleq [\mathbf{e}_{i_k,i}^{(1)}, \ldots, \mathbf{e}_{i_k,i}^{(\mathrm{S})}]$, two unitary matrices with dimension $(N_r - \mathrm{S}) \times (\mathrm{N_r} - \mathrm{S})$ and $\mathrm{S} \times \mathrm{S}$, respectively, and $\Lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}$ be an $(N_r - \mathrm{S}) \times \mathrm{S}$ matrix with main diagonal $(\lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(1)}, \ldots, \lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(d)})$ and zeros on its off-diagonal. Since $\mathbf{U}_{\mathbf{H}_{i_k,i}^{\text{eff}}}$ is unitary, without loss of generality, we can assume that the optimal receive beam-forming vector at the i_k user has the form of $\mathbf{w}_{i_k} = \mathbf{U}_{\mathbf{H}_{i_k,i}^{\text{eff}}} \mathbf{g}_{i_k}$. The received signals can then be expressed as

$$\hat{s}_{i_k} = \mathbf{g}_{i_k}^H \Lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}} \mathbf{V}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^H \sum_{m=1}^{I_i} \sqrt{\rho_{i_m}} \mathbf{v}_{i_m} s_{i_m} + \tilde{n}_{i_k}, \tag{2.29}$$

which contains the intra-cell interference caused by the multi-user nature of each cell as well as the inter-cell interference.

In the following, we will discuss how to mitigate the intra-cell interference by designing the user-specific beam-forming matrices $\mathbf{V}_i$ and user-specific receive beamforming vectors $\mathbf{g}_{i_k}$. To do so, the effective channel matrix should be fed back to the corresponding BS as discussed in Section II-C.

Remark 3. *It is worth mentioning that given the receive beam-forming vector at each receiver, the i-th BS only needs to know the quantized version of $\Lambda_{\mathbf{H}_{i_k,i}^{eff}}$ and $\mathbf{V}_{\mathbf{H}_{i_k,i}^{eff}}$ of the k-th user to obtain the transmit beamformer due to the fact that*

$$\mathbf{h}_{i_k} \triangleq \mathbf{g}_{i_k}^H \Lambda_{\mathbf{H}_{i_k,i}^{eff}} \mathbf{V}_{\mathbf{H}_{i_k,i}^{eff}}^H = \sum_{\ell=1}^{d} (g_{i_k}^{(\ell)})^* \lambda_{\mathbf{H}_{i_k,i}^{eff}}^{(\ell)} (\mathbf{e}_{i_k,i}^{(\ell)})^H. \tag{2.30}$$

2.3.2.2 Leakage-based iterative CB with feedback error and the intra-cell interference mitigation

This subsection focuses on optimizing the user-specific beam-forming matrix design based on the synthesized channel model presented in (2.7). As we discussed earlier, due to the finite-rate feedback mechanism in our system, the problem of interest can be formulated as

$$\mathbf{v}_{i_k}^{\text{opt}} = \underset{\mathbf{v}_{i_k} \in \mathbb{C}^{S\times 1}}{\arg\max} \quad \mathbb{E}\{\text{SLNR}_{i_k}\}, \tag{2.31}$$

which is subject to $\|\mathbf{v}_{i_k}\|^2 = 1$. As we discussed earlier, instead of maximizing the expected value of SLNR, we maximize the lower bound of the $\mathbb{E}\{\text{SLNR}_{i_k}\}$ as follows.

$$\mathbb{E}\{\text{SLNR}_{i_k}\} = \mathbb{E}\left\{ \frac{\mathbf{v}_{i_k}^H \mathbf{h}_{i_k}^H \mathbf{h}_{i_k} \mathbf{v}_{i_k}}{\mathbf{v}_{i_k}^H \left(\sum_{\substack{m=1 \\ m\neq k}}^{I_i} \mathbf{h}_{i_m}^H \mathbf{h}_{i_m} \right) \mathbf{v}_{i_k}^H + \frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}} \right\} \overset{(a)}{\geq} \frac{\mathbf{v}_{i_k}^H \mathbb{E}\left\{ \mathbf{h}_{i_k}^H \mathbf{h}_{i_k} \right\} \mathbf{v}_{i_k}}{\mathbf{v}_{i_k}^H \left(\sum_{\substack{m=1 \\ m\neq k}}^{I_i} \mathbb{E}\left\{ \mathbf{h}_{i_m}^H \mathbf{h}_{i_m} \right\} \right) \mathbf{v}_{i_k}^H + \frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}}$$

where $\tilde{\sigma}_{i_k}^2 = \sigma_{i_k}^2 + \rho_{N_r - S}(\hat{\mathbf{R}}_{i_k})$; and (a) comes from Jensen's inequality. Hence, in order to design the specific beam-forming vector $\mathbf{v}_{i_k}$, say $\mathbf{v}_{i_k}^{\text{opt}}$ for all $i_k \in \mathscr{I}$, we deal with the below optimization problem instead of solving (2.31)

$$\underset{\mathbf{v}_{i_k} \in \mathbb{C}^{S\times 1}, \|\mathbf{v}_{i_k}\|^2 = 1}{\arg\max} \quad \frac{\mathbf{v}_{i_k}^H \mathbb{E}\left\{ \mathbf{h}_{i_k}^H \mathbf{h}_{i_k} \right\} \mathbf{v}_{i_k}}{\mathbf{v}_{i_k}^H \left(\sum_{\substack{m=1 \\ m\neq k}}^{I_i} \mathbb{E}\left\{ \mathbf{h}_{i_m}^H \mathbf{h}_{i_m} \right\} \right) \mathbf{v}_{i_k} + \frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}}. \tag{2.32}$$

Theorem 4. *For the given receive beam-forming vectors $\mathbf{g}_{i_k}$, the optimization problem (2.32) is a generalized Rayleigh quotient problem [32, 33]. The optimal closed-form solution for $\mathbf{v}_{i_k}$ is available and can be expressed as follows:*

$$\mathbf{v}_{i_k}^{opt} = \nu_{\max}\left\{ \left(\frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}\mathbf{I}_S + \sum_{\substack{m=1 \\ m\neq k}}^{I_i} \left(\Psi_{i_m} + \sum_{\ell=1}^{d} \alpha_{i_m,\ell}\hat{\mathbf{e}}_{i_m,i}^{(\ell)}\hat{\mathbf{e}}_{i_m,i}^{(\ell)H} \right) \right)^{-1} \left(\Psi_{i_k} + \sum_{\ell=1}^{d} \alpha_{i_k,\ell}\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(\ell)H} \right) \right\}. \tag{2.33}$$

$$\alpha_{i_k,\ell} \triangleq |g_{i_k}^{(\ell)}|^2 |\hat{\lambda}_{\mathbf{H}_{i_k,i}^{eff}}^{(\ell)}|^2 (1-\delta),$$

$$\beta_{i_k,\ell} \triangleq g_{i_k}^{(\ell)} g_{i_k}^{(m)*} \hat{\lambda}_{\mathbf{H}_{i_k,i}^{eff}}^{(\ell)*} \hat{\lambda}_{\mathbf{H}_{i_k,i}^{eff}}^{(m)} (1 - \frac{\mathrm{S}-1}{2\mathrm{S}}\delta - \frac{\mathrm{S}-1}{8(\mathrm{S}+1)}\delta^2)^2,$$

$$\Psi_{i_k} \triangleq \sum_{\ell=1}^{d} \sum_{\substack{m=1 \\ m\neq \ell}}^{d} \beta_{i_k,\ell}\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(m)H} + \sum_{\ell=1}^{d} \frac{\delta}{\mathrm{S}} |g_{i_k}^{(\ell)}|^2 |\hat{\lambda}_{\mathbf{H}_{i_k,i}^{eff}}^{(\ell)}|^2 \mathbf{I}_{\mathrm{S}}.$$

where $\nu_{\max}(\mathbf{A})$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{A}$; $\hat{\mathbf{e}}_{i_k,i}^{(\ell)}$ is the quantized of channel direction information; $\hat{\lambda}_{\mathbf{H}_{i_k,i}^{eff}}^{(\ell)}$ is the quantized of channel quality information; and δ is the quantization error. The norm of $\mathbf{v}_{i_k}^{opt}$ is adjusted to $\|\mathbf{v}_{i_k}^{opt}\|^2 = 1$.

Proof. See Appendix A.0.4. □

Remark 4. *In our scheme, BS i only needs to know $\hat{\Lambda}_{\mathbf{H}_{i_k,i}^{eff}} \hat{\mathbf{V}}_{\mathbf{H}_{i_k,i}^{eff}}^H$ from user i_k. Comparing to conventional SLNR-based approaches [9] and [12], our scheme has similar operations, resulting in similar complexity. Moreover, the introduced scheme took the feedback error into account for designing transmit precoding matrices, thus, it is more robust to the feedback error as opposed to the conventional approaches. It is worth mentioning that the quantization error δ decreases as the feedback payload size β increases. Therefore, the introduced scheme will converge to the traditional beam-forming scheme when β is considerably large.*

Given user-specific transmit beam-forming vectors, $\mathbf{v}_{i_k}^{\mathrm{opt}}$, optimal MMSE receive beam-forming vectors become

$$\mathbf{g}_{i_k}^{\mathrm{opt}} = \left(\frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}\mathbf{I}_{N_r-\mathrm{S}} + \hat{\Lambda}_{\mathbf{H}_{i_k,i}^{\mathrm{eff}}} \hat{\mathbf{V}}_{\mathbf{H}_{i_k,i}^{\mathrm{eff}}}^H \left(\sum_{m=1,m\neq k}^{I_i} \rho_{i_m} \mathbf{v}_{i_m}^{\mathrm{opt}} \mathbf{v}_{i_m}^{\mathrm{opt}^H} \right) \hat{\mathbf{V}}_{\mathbf{H}_{i_k,i}^{\mathrm{eff}}} \hat{\Lambda}_{\mathbf{H}_{i_k,i}^{\mathrm{eff}}}^H \right)^{-1} \hat{\Lambda}_{\mathbf{H}_{i_k,i}^{\mathrm{eff}}} \hat{\mathbf{V}}_{\mathbf{H}_{i_k,i}^{\mathrm{eff}}}^H \mathbf{v}_{i_k}^{\mathrm{opt}}.$$

Table 2 summarizes the introduced algorithm.

Algorithm 2: Introduced Transceiver Design with Limited Feedback

(S.1) : At each user $i_k \in \mathscr{I}$, the CSI $\mathbf{H}_{i_k,j}$ are quantized to be $\hat{\mathbf{H}}_{i_k,j}$ using the codebook $\mathscr{C}$.
(S.2) : The quantized codeword indexes are then fed back to the j-th BS using feedback link.
(S.3) : Each BS receives the codebook indexes and then reconstructs the CSIs to be $\hat{\mathbf{H}}_{i_k,j}$.
(S.4) : Compute $\hat{\mathbf{T}}_i$ for all BSs [cf. (2.22)] based on the collected quantized CSI;
(S.5) : For all $i_k \in \mathscr{I}$ compute $\mathbf{W}_i$ [cf. (2.26)];
(S.6) : If $\hat{\mathbf{T}}_i$ satisfy a termination criterion: STOP.
(S.7) : Initialize the receive beam-forming vector $\mathbf{g}_{i_k}$ for $i_k \in \mathscr{I}$.
(S.8) : Compute the optimal transmit beam-forming vector $\mathbf{v}_{i_k}^{\text{opt}}$ based on Eq. (2.33) for different feedback strategies.
(S.9) : Update the receive beam-forming vector $\mathbf{g}_{i_k}$ based on Eq. (**??**) corresponding to the quantized channel matrix.
(S.10) : If $\mathbf{V}_i$ satisfy a termination criterion: STOP.

$$\mathbf{v}_{i_k}^{\text{opt}} = \nu_{\max}\left\{ \Big(\frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}\mathbf{I}_{\text{S}} + \sum_{\substack{m=1\\ m\neq k}}^{I_i}\left(\alpha_{i_m,1}\hat{\mathbf{e}}_{i_m,i}^{(1)}\hat{\mathbf{e}}_{i_m,i}^{(1)H} + \frac{\delta}{1-\delta}\frac{1}{\text{S}}\alpha_{i_m,1}\mathbf{I}_{\text{S}}\right)\Big)^{-1}\left(\alpha_{i_k,1}\hat{\mathbf{e}}_{i_k,i}^{(1)}\hat{\mathbf{e}}_{i_k,i}^{(1)H} + \frac{\delta}{1-\delta}\frac{1}{\text{S}}\alpha_{i_k,1}\mathbf{I}_{\text{S}}\right)\right\}, \tag{2.34}$$

$$\mathbf{v}_{i_k}^{\text{opt}} = \nu_{\max}\left\{ \Big(\frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}\mathbf{I} + \sum_{\substack{m=1\\ m\neq k}}^{I_i}\left(\Psi_{i_m} + \alpha_{i_m,1}\hat{\mathbf{e}}_{i_m,i}^{(1)}\hat{\mathbf{e}}_{i_m,i}^{(1)H} + \alpha_{i_m,2}\hat{\mathbf{e}}_{i_m,i}^{(2)}\hat{\mathbf{e}}_{i_m,i}^{(2)H}\right)\Big)^{-1}\left(\Psi_{i_k} + \alpha_{i_k,1}\hat{\mathbf{e}}_{i_k,i}^{(1)}\hat{\mathbf{e}}_{i_k,i}^{(1)H} + \alpha_{i_k,2}\hat{\mathbf{e}}_{i_k,i}^{(2)}\hat{\mathbf{e}}_{i_k,i}^{(2)H}\right)\right\}, \tag{2.35}$$

Feedback strategies: In the following, we introduce two feedback strategies: "*Rank 1 feedback*" and "*Rank 2 feedback*". Rank 1 feedback: In a typical cellular network, usually the largest singular value is dominant over other singular values. In order to reduce the feedback overhead, each user only need to feedback the quantized version of the dominant singular value and the corresponding eigen-direction, which is usually termed as "*Rank* 1 *feedback*". Without loss of generality, we assume that the first singular value is the largest one. For a given $\mathbf{g}_{i_k}$, the optimal transmit beam-forming vectors $\mathbf{v}_{i_k}$ calculated by BS i can be expressed as (2.34), and therefore, the optimal MMSE receive beam-forming vector can be obtained via (**??**) where $\hat{\Lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}\hat{\mathbf{V}}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{H} = [\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(1)}\hat{\mathbf{e}}_{i_k,i}^{(1)H}\ ,\ \mathbf{0}_{(N_r-S-1)\times S}]^T$ is the quantized channel matrix.

Rank 2 feedback: Alternatively, each user can feed back two dominant directions: both singular values and their corresponding eigen-directions are fed back to the BS. Without loss of generality,

we assume that the first two singular values are the dominant ones. Similar to Rank 1 feedback, the optimal solutions for the transmit beam-forming vectors can be expressed as (2.35), and the optimal receive beam-forming vector $\mathbf{g}_{i_k}$ in this case, has the same expression as (**??**) where the quantized channel matrix is in the form of $[\hat{\lambda}^{(1)}_{\mathbf{H}^{\text{eff}}_{i_k,i}} \hat{\mathbf{e}}^{(1)H}_{i_k,i},\ \hat{\lambda}^{(2)}_{\mathbf{H}^{\text{eff}}_{i_k,i}} \hat{\mathbf{e}}^{(2)H}_{i_k,i}\ , \mathbf{0}_{(N_r-\text{S}-2)\times \text{S}}]^T$. Note that Rank 2 feedback doubles the overhead as opposed to Rank 1 feedback.

Performance analysis: To characterize the performance of the introduced algorithm with limited feedback, we examine the rate loss [25] incurred by naive IA and leakage-based iterative coordinated beam-forming algorithm, where channel estimates are used to calculate the columns of the precoders, $\hat{\mathbf{f}}_{i_k}$ and combiners $\hat{\mathbf{u}}_{i_k}$ for $i_k \in \mathscr{I}$. The mean loss in sum-rate is then defined as $\triangle R_{sum} = \mathbb{E}_{\mathbf{H}}\{R_{sum}\} - \mathbb{E}_{\mathbf{H}}\{\hat{R}_{sum}\}$ where $\mathbb{E}_{\mathbf{H}}\{R_{sum}\}$ is the average sum rate from the introduced algorithm with perfect CSI (Algorithm 1); with the sum-rate given in (2.15), and $\mathbb{E}_{\mathbf{H}}\{\hat{R}_{sum}\}$ is the rate achieved with imperfect CSI using Algorithm 2.

Using the received signal, the instantaneous rate expression in (2.15), and defining the leakage interference as $\hat{\mathscr{I}}_{i_k} = \sum_{(j,m)\neq(i,k)} \rho_{j_m} |\hat{\mathbf{u}}^H_{i_k} \mathbf{H}_{i_k,j} \hat{\mathbf{f}}_{j_m}|^2$; the following upper bound on mean loss in sum-

rate will be achieved

$$
\begin{aligned}
\triangle R_{sum} &= \mathbb{E}_{\mathbf{H}}\left\{ \sum_{(i,k)} \log_2\left(1+\frac{\rho_{i_k}|\mathbf{u}_{i_k}^H \mathbf{H}_{i_k,i}\mathbf{f}_{i_k}|^2}{\hat{\sigma}_{i_k}^2}\right)\right\} \\
&- \mathbb{E}_{\mathbf{H},\hat{\mathbf{H}}}\left\{ \sum_{(i,k)} \log_2\left(1+\frac{\rho_{i_k}|\hat{\mathbf{u}}_{i_k}^H \mathbf{H}_{i_k,i}\hat{\mathbf{f}_{i_k}}|^2}{\sigma_{i_k}^2+\hat{\mathscr{I}}_{i_k}}\right)\right\} \\
&= \mathbb{E}_{\mathbf{H}}\left\{ \sum_{(i,k)} \log_2\left(1+\frac{\rho_{i_k}|\mathbf{u}_{i_k}^H \mathbf{H}_{i_k,i}\mathbf{f}_{i_k}|^2}{\hat{\sigma}_{i_k}^2}\right)\right\} \\
&- \mathbb{E}_{\mathbf{H},\hat{\mathbf{H}}}\left\{ \sum_{(i,k)} \log_2\left(1+\frac{\hat{\mathscr{I}}_{i_k}+\rho_{i_k}|\hat{\mathbf{u}}_{i_k}^H \mathbf{H}_{i_k,i}\hat{\mathbf{f}_{i_k}}|^2}{\sigma_{i_k}^2}\right)\right\} \\
&+ \mathbb{E}_{\mathbf{H},\hat{\mathbf{H}}}\left\{ \sum_{(i,k)} \log_2(1+\frac{\hat{\mathscr{I}}_{i_k}}{\sigma_{i_k}})\right\} \\
&\overset{(a)}{\leq} \mathbb{E}_{\mathbf{H},\hat{\mathbf{H}}}\left\{ \sum_{(i,k)} \log_2(1+\frac{\hat{\mathscr{I}}_{i_k}}{\sigma_{i_k}})\right\} \overset{(b)}{\leq} \sum_{(i,k)} \log_2\left(1+\frac{\mathbb{E}_{\mathbf{H},\hat{\mathbf{H}}}\{\hat{\mathscr{I}}_{i_k}\}}{\sigma_{i_k}}\right)
\end{aligned}
\tag{2.36}
$$

where $\hat{\sigma}_{i_k}^2 = \sigma_{i_k}^2 + \sum_{m=1,m\neq k}^{I_i} \rho_{i_m} |\mathbf{w}_{i_k}^{\text{opt}^H} \mathbf{H}_{i_k,i}^{\text{eff}} \mathbf{v}_{i_m}^{\text{opt}}|^2$, (a) holds due to the fact that the desired signal powers $|\mathbf{u}_{i_k}^H \mathbf{H}_{i_k,i}\mathbf{f}_{i_k}|^2$ and $|\hat{\mathbf{u}}_{i_k}^H \mathbf{H}_{i_k,i}\hat{\mathbf{f}}_{i_k}|^2$, resulting from introduced algorithm with perfect and imperfect CSI respectively, are identically and exponentially distributed [19, Lemma 1] and (b) follows from Jensen's inequality. The total interference $\hat{\mathscr{I}}_{i_k}$ can be simplified to include the residual interference due to the channel estimation error, therefore by recalling Eq. (2.27), the mean loss in sum-rate can be further upper bounded by $\sum_{(i,k)} \log_2\left(1+\frac{\rho_{N_r-S}(\hat{\mathbf{R}}_{i_k})}{\sigma_{i_k}}\right)$.

2.4 Simulation Evaluations

In this preliminary experimental evaluation, we investigate the performance of the introduced schemes for a MU-MIMO broadcast cellular system composed of multiple cells, with one BS

Table 2.1: System parameters used in simulation.

Parameters	**Values**
Cell Layout	Hexagonal grid
Number of transmit antenna	4
Number of receive antenna	2
Number of MSs per cell	500
Inter-site distance	500 m
Minimum distance between MS and BS	> 35 m
MS distribution	Uniform random distribution
Bandwidth	5 MHz
Maximum Transmission Power	43 dBm
Thermal Noise Level	-174 dBm
Path Loss (dB)	$128.1 + 37.6 \log_{10}(d)$ with d in Km
Shadowing model	Log-normal shadowing with 4dB SD.

and multiple randomly generated MSs in each cell. The setup of our experiments is the following. We simulated seven cells with multiple users that randomly and uniformly dropped (at a distance > 35 m and < 275 m from the BS) in each cell. The transmission is subject to interference from 6 neighboring BSs. The transmit power at each BS is fixed to 46 dBm and the noise variance at the MS is fixed to -174 dBm. System bandwidth is taken as 5 MHz. We consider a possible antenna configuration in a typical deployment scenario for LTE/LTE-Advanced systems: 4 transmit and 2 receive antennas. Channels are Rayleigh fading, in which path-loss is generated using 3 GPP (TR 36.814) methodology. All deployments and channel model parameters are listed in Table I. Due to the fact that co-scheduled MSs usually have similar SNRs in MU-MIMO operation, we assume the MSs have the same SNR. We also assume white zero-mean Gaussian noise at each receiver, with variance of 1. Random vector quantization is used to generate the codebook which is revealed to all BSs and the mobile MSs. In our evaluation, for simplicity, the initial guess of the receive beam-forming vector, $\mathbf{g}_{i_k}$, is set to be either $[1\ 0]^H$ or $[0\ 1]^H$, and the introduced algorithm is terminated when the absolute value of the sum-rate error in two consecutive rounds becomes smaller than $1e-2$.

In Fig. 5.1, we plot the average spectral-efficiency of our scheme versus the SNR. Orthogonal

transmissions are imposed a priori among the BSs; therefore each MS only experiences intra-cell interference. We observed that, as expected, the spectral-efficiency performance of introduced algorithm with the rank 2 feedback outperforms the rank 1 feedback by 8.869%, at SNR $= 10$ dB. The benefits become more substantial in the high SNR regime which is due to the fact that the second largest singular value cannot be ignored in this regime. However, the performance gain is achieved at the cost of double feedback payload. The spectral-efficiency performance of our introduced scheme is also compared to the traditional one [12] which does not consider the quantization error of the channel feedback. As shown in Fig. 5.1, our strategy has a better performance than the traditional one. To be specific, the rank 1 and rank 2 proposed scheme outperform the traditional one at SNR $= 10$ dB by 5.75% and 8.03%, respectively. The benefits become more considerable in higher SNR regime. Moreover, we compare the introduced precoding scheme with the ZF beamforming in finite-rate feedback system which consists of quantization error. It is shown that our introduced scheme outperforms the ZF by 143.99% at SNR $= 10$ dB. Moreover, with the same feedback overheads, the performance of the introduced scheme always outperforms the ZF. Furthermore, in the low SNR regime and with the Rank 2 feedback, the introduced scheme almost achieves the same performance as the ZF gains with the perfect feedback. In addition, for a fixed payload size of the codebook, unlike the ZF beamforming where the spectral-efficiency is bounded as the SNR increases, the performance gap between the rank 2 feedback and the perfect feedback remains almost constant as SNR increases in our scheme. As illustrated in Fig. 5.1, the introduced scheme can significantly improve the received SINR of co-scheduled users over existing MIMO precoding strategies.

In Fig. 5.2, we plot the average sum-rate of our introduced schemes versus the SNR under different feedback strategies: "Rank 1 feedback", "Rank 2 feedback", "Rank S feedback", and "Perfect CSI". In the "Rank S feedback" strategy, each user feedbacks all S quantized version of the effective channel singular values and their corresponding eigen-directions. We observed that, the IA with perfect CSI clearly outperforms the other feedback strategies. Moreover, the spectral-efficiency performance of introduced algorithm with the rank 2 feedback outperforms

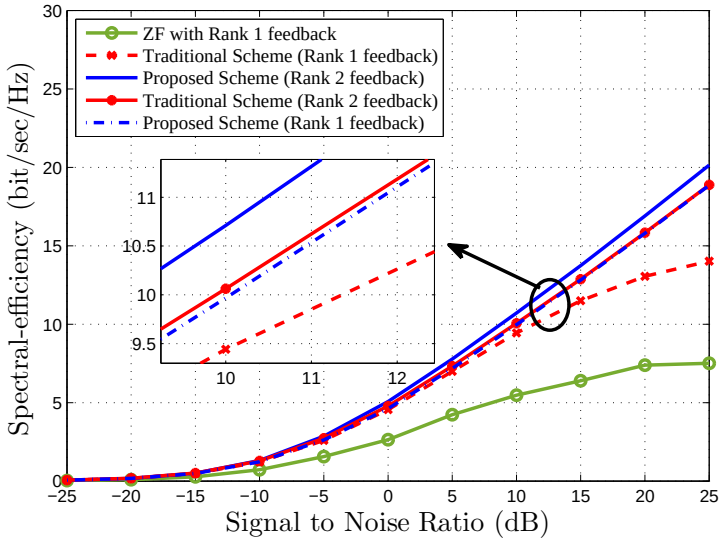

Figure 2.2. Spectral-efficiency vs. SNR ($N_t = 4$, $N_r = 2$, and $\beta = 8$)

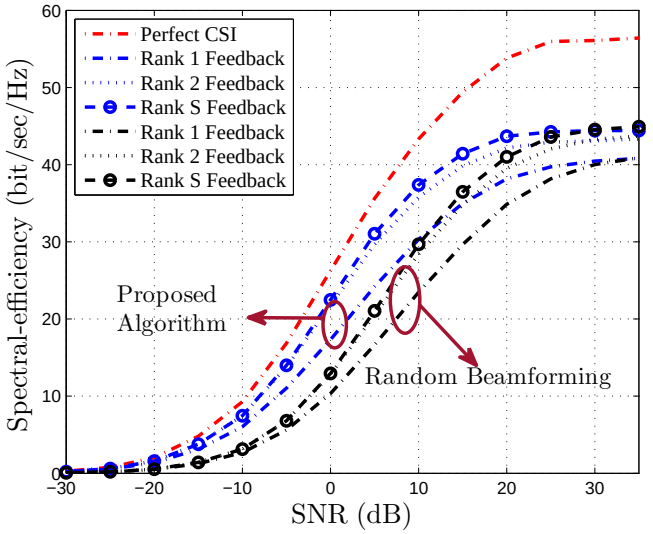

Figure 2.3. Spectral-efficiency vs. SNR ($N_t = N_r = 8$, and $\beta = 4$)

the rank 1 feedback. The benefits become more substantial in high SNR regime which is due to the fact that the second largest singular value cannot be ignored in this regime. However, the performance gain is achieved at the cost of double feedback payload. Interestingly, the Rank 2 feedback performs close to the Rank S feedback with much more less feedback payload. From Fig. 5.2, it is obvious that the introduced scheme with imperfect CSI demonstrates significant performance improvements over the random beamforming strategy [16] in which we did not design the beamforming matrices for combating inter-cell interference.

In Fig. 4.7, we plot the average sum-rate versus the number of co-scheduled users. The intuition behind this result is that the interference will never be aligned due to imperfect CSI and S is a very important system design parameter affecting the overall network performance. A larger S will lead to a larger signal space allowing more number of co-scheduled MSs, however, in this situation the residual inter-cell interference will be high due to the small dimension of the null-space $N_r - S$. On the other hand, even though a smaller S will significantly reduce the inter-cell interference, it sacrifices the DoF of the serving BS. Therefore, there exists an optimal S that maximizes the system throughput so that some tolerable inter-cell interference are allowed among cells. Moreover, we observed that, for $S = 1$ and $S = N_r - 1$, the performance of Rank 1 Rank 2, and Rank S feedback strategies are the same, which is due to the fact that in these cases the rank of the equivalent channel is equal to one. The same happens when we compare the performance of Rank 2 and Rank S feedback strategies for $S = 2$ and $S = N_r - 2$.

2.5 Conclusions

In this chapter, limited feedback-based co-channel interference mitigation of a multi-cell MU-MIMO is investigated. To be specific, limited feedback-based IA was introduced to eliminate the inter-cell interference while an improved low complexity iterative leakage-based coordinated beamforming strategy was introduced to mitigate the intra-cell interference. By jointly considering the transmit beamforming, receive beamforming, and the quantization error of the codebook-based

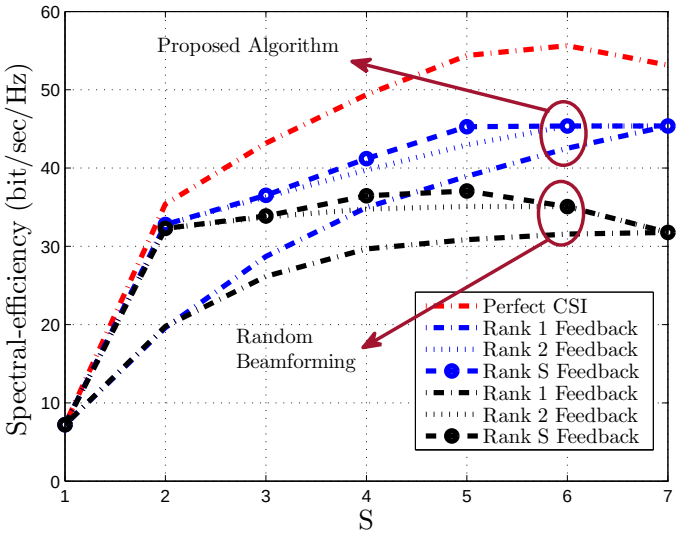

Figure 2.4. Spectral-efficiency vs. S ($N_t = N_r = 8$ and $\beta = 4$)

feedback, our algorithm greatly mitigates the CCI and achieves a better performance compared to the traditional ones. Through system level evaluations, it was shown that the introduced scheme significantly outperforms the conventional interference management schemes in practical environments.

Chapter 3

Coordinated Multi-Point Transmission and Reception: Resource Allocation

3.1 Introduction

Coordination among base stations (BSs) has been widely studied in recent years to tackle inter-cell interference that strongly limits the achievable rates of cellular networks [34]. Supported by the first results promising huge performance gains with respect to the baseline non-cooperative system [35], a lot of attention has been paid to the topic both in the academia [36–39] and in the industry [40–45]. This technique is usually called "Multi-cell MIMO" in academia and is also termed as "Coordinated Multi-point (CoMP)" in 3GPP LTE-Advanced systems. CoMP allows a group of BSs to cooperate and coordinate to improve the overall system performance and system efficiency. To be specific, these cooperating BSs communicate with one another through the inter-base station links such as X2 interfaces [46]. In general, depending on whether data is shared among multiple BSs, CoMP can be categorized into the following two operations [36, 40]: coordinated scheduling/coordinated beamforming (CS/CB) and joint transmission (JT). While the later one requires both channel state information (CSI) and data sharing among the BSs, the former one only requires CSI to be shared. Although the JT approach is attractive and can provide substantial

capacity gains for CoMP systems, it also bring a range of problems. For example, CoMP-JT will cause high signaling overhead through the backhaul incurred by distributing each user's data to multiple BSs in the cooperating set, lead to high computational complexity due to user scheduling and transmit precoding design, and require very tight synchronization among all BSs within the cooperating set [47–49]. On the other hand, in CoMP-CS/CB schemes BSs only need to share their associated control information (e.g., each BS's scheduling information as well as scheduled users' CSI) so that BSs in the cooperating set could coordinate their scheduling and/or beam-forming schemes so that the inter-cell interference is mitigated. Due to the fact that users' data does not need to be shared among BSs in the cooperating set and no coherent joint transmission is needed, CoMP-CS/CB puts a much lesser requirement on backhaul as well as synchronization. This is the main reason CoMP-JT is not adopted in Rel-10 LTE-Advanced systems while CoMP-CS/CB are [40, 41]. In this chapter, we focus on downlink CoMP-CB where each BS serves its associated mobile stations (MSs) upon designing its power allocation and transmitter precoder matrices to reduce the inter-cell interference. BSs cooperate to share CSI to reduce the interference toward other cell's MSs while maintaining the power of its own desired signal. Consequently, this coordination improves the overall network spectral-efficiency.

In general, two performance measures are critical for a cellular network: cell-average spectral-efficiency and cell-edge user spectral-efficiency. The cell-average spectral-efficiency specifies the average spectral-efficiency over all active MSs present in a network and the cell-edge user spectral-efficiency is defined to be the 5%-tile of the spectral-efficiency of the corresponding MSs. In order to achieve good spectral-efficiency performance of a cellular network, the cell-edge user spectral-efficiency target and the cell-average spectral-efficiency target have to be met simultaneously. This is a challenging task because there is usually a clear trade-off between cell-edge user performance and cell-average performance [37, 43]. To deal with this issue, some literatures have proposed different kind of fairness resource allocation such as maximum throughput and max-min fairness. When maximum throughput scheduler is adopted, the BS will always schedule the user with the largest throughput. This will definitely maximize the cell-average user spectral-efficiency (or the

sum throughput of the network), however, the cell-edge user spectral-efficiency will be zero since cell-edge users will never get the scheduling opportunity. On the other hand, max-min fairness tries to ensure a uniform user experience across the network. This will certainly benefit the cell-edge user spectral-efficiency, however, it will cause a hit on the cell-average user spectral-efficiency. A good compromise between max-min fairness and maximum throughput scheduling is proportional fairness since it can provide a good tradeoff between cell-edge user spectral-efficiency and cell-average user spectral-efficiency [37, 43]. This is the main reason why proportional-fair scheduling is widely used in 3GPP LTE-Advanced networks.

For an LTE-Advanced CoMP-CB system, each BS is equipped with multiple transmit antennas and each MS is equipped with multiple receive antennas. The multiple antennas at each BS are exploited to serve a dual objective: avoid interference and enhance throughput via multiple-input-multiple-output (MIMO) transmission toward the intended receivers. Since these two objectives may conflict with each other and to ensure the fairness across the network to achieve good balance between the two importance performance measures, MIMO transmit precoding has to be coupled with network scheduling. To be specific, in this chapter, a network-wide proportional-fair resource allocation problem is formulated where we optimize MIMO transmit precodings as well as power allocations for BSs involved in CoMP-CB to maximize the network proportional-fair metric. The network proportional-fair metric is essentially a weighted sum-rate of users presented in the network where the weights are dependent on the underlying proportional-fair scheduling and quality of service (QoS) constraints. By conducting the proportional-fair resource allocation (combining MIMO transmit precoding and power allocation), we will be able to achieve a good trade-off between the two spectral-efficiency performance measures to optimize network performance.

Efficiently obtaining the optimal resource allocation (power allocation and precoding matrices) to achieve the maximum network proportional-fair metric is a challenging task. The challenge is that the relevant optimization problem often turns out to be nonconvex due to the coupled interference among different receivers. The problem of weighted sum-rate maximization with linear precoding is claimed to be solved by the algorithm of [50]. However, convergence to the optimum

solution is not guaranteed and it strongly depends on the initialization. The algorithm works iteratively, requires formulating and solving a geometric program in each step and is based on a repeated transformation from the dual multiple access channel to the broadcast channel and back enabled by the single data stream duality of [51]. Successive solution of geometric programming problems was proposed in [52] with promising results. In the special case of two interfering links, a simple on/off power control strategy was shown to be optimal [53]. The study of distributed methods for finding the precoding matrices in a general nonconvex weighted sum-rate maximization problem is also considered in [54,55]. Besides the application of the classical gradient projection algorithm to the sum-rate maximization problem over MIMO interference channels [54], parallel iterative algorithm for MIMO broadcast interfering channels was proposed in [55]. Unfortunately, the gradient schemes [54] suffer from slow convergence and do not exploit any degree of convexity that might be present in the objective function; and [55] is based on the connection with a weighted minimum mean-square error (WMMSE) problem. Moreover, the authors did not consider either the cooperation between BSs (for mitigation of strong multi-cell interference caused by aggressive/universal frequency reuse in the network) or the proportional-fair resource allocation (the weights in weighted sum-rate are set equally for all users). The problem of weighted sum-rate maximization in LTE-Advanced CoMP-CB system is addressed in [56–58]. These papers take the sub-optimal concept referred to as signal-to-generating-interference-plus-noise ratio (SGINR). This metric reflects the covariance matrix of the interference which each BS generates as well as the covariance matrix of the desired channel. By using the eigen matrix of this metric as a beamforming matrix of each BS, they tried to reduce the interference among cooperating BSs.

This work differs from previously mentioned studies particularly in its aim to maximize the throughput in a proportionally fair way, taking into account a detailed resource allocation design for a CB-based CoMP LTE-Advanced scheme. The main contributions of this chapter can be summarized as follows:

- for the first time, we introduce the proportional-fair resource allocation by combining MIMO transmit precoding and power allocation to improve both the cell-edge as well as the cell-

average user spectral-efficiency and identify the interaction between resource allocation and MIMO transmission for a CoMP-CB LTE-Advanced network;

- we introduce a parallel scheme based on the successive convex approximation (SCA) algorithm for a nonconvex network proportional-fair optimization problem in a CoMP-CB LTE-Advanced systems. The decomposition enables all BSs to update their optimization variables in parallel by solving a sequence of strongly convex subproblems, one for each BS;

- to the best of our knowledge, there is no existing closed-form expressions of the locally optimal solution for the aforementioned problem. In this chapter, we characterize the closed-form expressions for some special cases as well as for both high and low signal-to-noise (SNR) regimes;

- we introduce an iterative algorithm to characterize proportional-fair weights for the underlying resource allocation problem;

- we show that the proposed scheme is a promising candidate for improving both the cell-edge and cell-average user spectral-efficiency in the CoMP-CB LTE-Advanced systems and the benefits of the proposed proportional-fair resource allocation algorithm are illustrated in terms of performance and computational complexity.

3.2 Problem Formulation

3.2.1 System Model and Assumptions

In this chapter, we focus on the scenario where an LTE-Advanced network consists of L cooperating BSs, each equipped with N_t transmit antennas, is supporting L MSs, each equipped with N_r antennas. That is, each cooperating BS is supporting a single MS. However, it is important to note that the framework can be easily extended to the case where more than one MSs are supported

simultaneously by each cooperating BS. Each BS in the cooperating set is assumed to conduct its own user scheduling individually. Once the scheduling decision is made, BSs will exchange coordination messages through the X2 interface [41,46] to enable CoMP-CB operation. It is important to note that the coordination message only contains information related to CSI. The delay required for exchanging coordination messages is assumed to be negligible [44].

It is further assumed that receivers perform single user detection; i.e., only the desired signal is detected at each receiver and the signals from co-channel users are treated as noise. Assuming Rayleigh flat fading channel model, the MIMO channel between ℓ-th BS and i-th MS is represented by the matrix $\mathbf{H}_{i\ell} = \mathbf{W}_{i,\ell}\mathbf{R}_{\ell}^{1/2}$, where $\mathbf{W}_{i,\ell} \in \mathbb{C}^{N_r \times N_t}$ is a random channel matrix whose elements are identically and independently distributed (i.i.d.) cyclic symmetric complex Gaussian with zero mean and unit variance, and $\mathbf{R}_{\ell} \in \mathbb{C}^{N_t \times N_t}$ represents the transmit correlation matrix of the ℓ-th BS. The channel transfer matrices $\mathbf{H}_{i\ell}$'s are assumed to be independent of each other. Moreover, all the channels in the network are assumed to be quasi-static block fading, meaning that the channel gains remain constant during one block and change independently from block to block.

At each transmission time slot, i-th MS is assumed to receive N_s parallel streams of information signal. It is supposed that ℓ-th BS simultaneously transmits the sequence $\mathbf{x}_{\ell} = (x_{\ell,1}, \ldots, x_{\ell,N_s})^T$ to its corresponding receiver, where the entries of vector $\mathbf{x}_{\ell}$ are independently Gaussian encoded symbols. In our model, ℓ-th BS processes its symbols using an $N_t \times N_s$ precoding matrix $\mathbf{V}_{\ell}$, whose columns constitute an orthonormal basis for the transmitted-signal space of user $\ell \in \mathscr{L}$, to form its transmitted signal vector $\mathbf{V}_{\ell}\mathbf{x}_{\ell}$.

Following a matrix notation, the received signals at i-th MS, represented by the length N_r column-vector

$$\mathbf{y}_i = \mathbf{H}_{ii}\mathbf{V}_i\mathbf{x}_i + \sum_{j \in \mathscr{L} \setminus i} \mathbf{H}_{ij}\mathbf{V}_j\mathbf{x}_j + \mathbf{n}_i, \tag{3.1}$$

In (3.1), the first term represents the desired signal from i-th BS, while the second term is inter-cell interference and $\mathbf{n}_i \sim \mathscr{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{N_i})$ is the complex normal thermal noise vector at receiver

$i \in \mathscr{L}$. The $N_s \times N_s$ power allocation matrix $\mathbf{P}_\ell = \mathbb{E}[\mathbf{x}_\ell \mathbf{x}_\ell^H]$ denotes the input covariance matrix at the transmitter $\ell \in \mathscr{L}$. Since the entries of $\mathbf{x}_\ell$, $\ell \in \mathscr{L}$, are zero-mean and mutually independent, covariance matrix $\mathbf{P}_\ell$ is diagonal, say equal to $\mathbf{P}_\ell = \text{diag}(p_{\ell,1}, \ldots, p_{\ell,N_s})$. For this case, the power constraint on the transmitted vector $\mathbf{V}_\ell \mathbf{x}_\ell$, $\ell \in \mathscr{L}$, can be written as

$$\text{Tr}(\mathbf{V}_\ell \mathbf{P}_\ell \mathbf{V}_\ell^H) = \text{Tr}(\mathbf{P}_\ell) \leq p_{\ell,\max}, \tag{3.2}$$

where $p_{\ell,max}$ is the maximum transmit-power of the ℓ-th BS and it is finite. The achievable rate of the i-th MS is calculated as:

$$R_i(\mathbf{P}_i, \mathbf{P}_{-i}) \triangleq \log_2 \det(\mathbf{I}_{N_r} + \mathbf{R}_i(\mathbf{P}_i)\mathbf{R}_i^{-1}(\mathbf{P}_{-i})) \tag{3.3}$$

where $\mathbf{P}_{-i} \stackrel{\text{def}}{=} (\mathbf{P}_j)_{j \neq i}$, $\mathbf{R}_i(\mathbf{P}_i)$ represent the covariance matrix of the desired signal, and $\mathbf{R}_i(\mathbf{P}_{-i}) = \sigma^2 \mathbf{I}_{N_r} + \sum_{j \neq i} \mathbf{H}_{ij} \mathbf{V}_j \mathbf{P}_j \mathbf{V}_j^H \mathbf{H}_{ij}^H$ which represents the covariance matrix of the thermal noise plus multiuser interference at i-th MS.

3.2.2 Proportional-Fair Resource Allocation

The problem of interest is to maximize (over the resource allocation: precoding matrices and power allocations) the network proportional-fair metric of an LTE-Advanced CoMP-CB system, which is essentially a weighted sum-rate of users presented in the network. The maximization of the proportional-fair metric allows to cover all the rate tuples on the rate region boundary. Mathematically, the proportional-fair metric maximization problem can be expressed as

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad \sum_{i=1}^{L} w_i R_i(\mathbf{P}_i, \mathbf{P}_{-i}) \tag{3.4}$$

where $\mathscr{K} \triangleq \{\mathbf{P}_i \in \mathbb{C}^{N_s \times N_s} \ : \ \mathbf{P}_i \succeq \mathbf{0}, \ \text{Tr}(\mathbf{P}_i) \leq p_{i,max}\}$ represents the per-node transmit power constraint and w_i denotes the nonnegative proportional-fair weight of the i-th MS. In each time slot t, the proportional-fair weight of MS i is defined as $w_i(t) \triangleq R_i^\alpha(t)/T_i^\beta(t)$, where α and β tune the fair-

ness of the algorithm and T_i is the accumulated throughput of MS i: $T_i(t+1) = T_i(t) + R_i(t+1) \times t_c$ [59] (t_c is the slot duration). Selecting $\alpha \approx 1$ and $\beta \approx 1$ will yield the proportional-fair scheduling used in 3GPP LTE-Advanced networks [60]. It is important to note that the proportional-fair weights are not fixed throughout the transmission and are adapted based on per-user average rate over time axis.

3.3 A Novel Algorithm For Coordinated Resource Allocation

Problem (3.4) has been shown to be NP hard [61]. Then, there is no hope to compute a globally optimal solution in polynomial time. Thus, we are interested in distributed solution methods for computing stationary solutions (possibly local optimum) of this problem. To solve the corresponding nonconvex problem efficiently, we develop a Successive Convex Approximation (SCA)-based method where (3.4) is replaced by a sequence of strongly convex problems. At the basis of the proposed technique, there is a suitable convex approximation of the nonconvex objective function $\sum_{i\in\mathscr{L}} w_i R_i(\mathbf{P}_i, \mathbf{P}_{-i})$. To be specific, given the strategy profile $\mathbf{P}_{-i}$, the aim of each BS is to choose a feasible power allocation $\mathbf{P}_i$ that maximizes the rate $R_i(\mathbf{P}_i, \mathbf{P}_{-i})$. Inspired by the algorithm proposed in [62], our method is based on solving a sequence of parallel convex problems, one for each BS, obtained by preserving the convex structure of the utility function network proportional-fair metric while linearizing the rest around $\bar{\mathbf{P}}_i$. To this end and in order to isolate the interference which makes (3.4) nonconvex, we define the network proportional-fair metric of the BSs other than the i-th as $f_i(\mathbf{P}_i, \mathbf{P}_{-i}) \triangleq \sum_{j\neq i} w_j R_j(\mathbf{P}_i, \mathbf{P}_{-i})$. By linearizing the nonconvex part of the objective function, i.e., $f_i(\mathbf{P}_i, \mathbf{P}_{-i})$, and keeping the convex part, i.e., $R_i(\mathbf{P}_i, \mathbf{P}_{-i})$, we can convexify the objective function. For this purpose, we use the first order Taylor series expansion of a continuously $\mathbb{R}$-differentiable function $f_i(\mathbf{P}_i, \mathbf{P}_{-i})$ that is given by:

$$f_i(\mathbf{P}_i, \mathbf{P}_{-i}) \approx f_i(\bar{\mathbf{P}}_i, \mathbf{P}_{-i}) + \langle \nabla_{\mathbf{P}_i} f_i|_{\mathbf{P}_i=\bar{\mathbf{P}}_i} \, , \, \mathbf{P}_i - \bar{\mathbf{P}}_i \rangle \tag{3.5}$$

where we introduced the trace inner product in a matrix space $\langle \cdot\ ,\ \cdot \rangle : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \to \mathbb{R}$, defined as $\langle \mathbf{A}, \mathbf{B} \rangle \triangleq \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. Note that the norm induced by the inner product $\langle \cdot\ ,\ \cdot \rangle$ is the Frobenius norm, i.e., $\langle \mathbf{A}, \mathbf{A} \rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_F^2$. Recalling that $d\ln(\det(\mathbf{Z})) = \mathrm{Tr}\left\{\mathbf{Z}^{-1}d\mathbf{Z}\right\}$ for all $\mathbf{Z}$ such that $\det\mathbf{Z} \neq 0$ [63, Prop. 3.14], with $d\ln(\det(\mathbf{Z}))$ being the differential of the $\ln(\det(\mathbf{Z}))$, the first-order differential (up to a constant positive factor) is given by Eqn. (3.6), at the top of the next page, where $\chi_j \triangleq \mathbf{H}_{jj}\mathbf{V}_j\mathbf{P}_j\mathbf{V}_j^H\mathbf{H}_{jj}^H$, $\mathbf{Z}_j \triangleq \mathbf{I}_{N_r} + \mathbf{R}_j^{-1}\chi_j$, $\mathbf{Y}_j \triangleq \mathbf{R}_j^{-1}\mathbf{H}_{ji}\mathbf{V}_i$, $\mathbf{M}_j = \mathbf{Y}_j^H\chi_j\mathbf{Z}_j^{-1}\mathbf{Y}_j$,

$$
\begin{aligned}
df_i(\mathbf{P}_i, \mathbf{P}_{-i}) &= \sum_{j\neq i} w_j \mathrm{Tr}\left(\mathbf{Z}_j^{-1}d\mathbf{Z}_j\right) = \sum_{j\neq i} w_j \mathrm{Tr}\left(\mathbf{Z}_j^{-1}d\mathbf{R}_j^{-1}\chi_j\right) \overset{(a)}{=} \sum_{j\neq i} -w_j \mathrm{Tr}\left(\mathbf{Z}_j^{-1}\mathbf{R}_j^{-1}d\mathbf{R}_j\mathbf{R}_j^{-1}\chi_j\right) \\
&= \sum_{j\neq i} -w_j \mathrm{Tr}\left(\mathbf{Z}_j^{-1}\mathbf{Y}_j d\mathbf{P}_i \mathbf{Y}_j^H\chi_j\right) \overset{(b)}{=} \sum_{j\neq i} w_j \mathrm{Tr}\left(-\mathbf{Y}_j^H\chi_j\mathbf{Z}_j^{-1}\mathbf{Y}_j d\mathbf{P}_i\right) = \mathrm{Tr}\left(\sum_{j\neq i} -w_j\mathbf{M}_j d\mathbf{P}_i\right) \\
&\overset{(c)}{=} \mathrm{vec}^T\left(\sum_{j\neq i} -w_j\mathbf{M}_j\right)^T \mathrm{vec}(d\mathbf{P}_i) \overset{(d)}{=} \mathrm{vec}^T\left(\sum_{j\neq i} -w_j\mathbf{M}_j\right)^T d\mathrm{vec}(\mathbf{P}_i)
\end{aligned}
\tag{3.6}
$$

and (a) comes from $d\mathbf{Z}^{-1} = -\mathbf{Z}^{-1}d\mathbf{Z}\mathbf{Z}^{-1}$ [63, Prop. 3.5], (b) comes from $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$ [63, cf. (2.96)], (c) comes from $\mathrm{Tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{vec}^T(\mathbf{A})\mathrm{vec}(\mathbf{B})$ where $\mathrm{vec}^T(\mathbf{A}) = (\mathrm{vec}(\mathbf{A}))^T$ [63, cf. (2.97)] and (d) comes from $\mathrm{vec}(d\mathbf{Z}) = d\mathrm{vec}(\mathbf{Z})$ [63, Prop. 3.9]. It is worthwhile noticing that, when f is a (complex-valued) scalar function of complex matrices, that is $f : \mathbb{C}^{n\times m} \longrightarrow \mathbb{C}$, we have $\frac{\partial f}{\partial(\mathbf{Z})_{ij}}$ and $\frac{\partial f}{\partial(\mathbf{Z}^*)_{ij}}$ which are $n\cdot m$ component-wise $\mathbb{R}$-derivatives and $n\cdot m$ conjugate $\mathbb{R}$-derivatives of the complex-valued function f with respect to $(\mathbf{Z}_{ij})$ and $(\mathbf{Z}_{ij}^*)$, respectively. Given f, the matrix gradient and conjugate-gradient of $\mathbf{Z}_0 \in \mathbb{C}^{n\times m}$ are defined as

$$
\nabla_{\mathbf{Z}} f(\mathbf{Z}_0) \triangleq \left.\frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}}\right|_{\mathbf{Z}=\mathbf{Z}_0}, \quad \nabla_{\mathbf{Z}^*} f(\mathbf{Z}_0) \triangleq \left.\frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}^*}\right|_{\mathbf{Z}=\mathbf{Z}_0}
$$

with $[\partial f/\partial \mathbf{Z}]_{ij} = \partial f/\partial(\mathbf{Z})_{ij}$ and $[\partial f/\partial \mathbf{Z}^*]_{ij} = \partial f/\partial(\mathbf{Z}^*)_{ij}$ for $i \in \{1,...,n\}$ and $j \in \{1,...,m\}$. Note that, $\nabla_{\mathbf{Z}} f(\mathbf{Z}_0)$ and $\nabla_{\mathbf{Z}^*} f(\mathbf{Z}_0)$ are matrices having the same size of $\mathbf{Z}$. Alternatively, one can

arrange the elements $\frac{\partial f}{\partial (\mathbf{Z})_{ij}}$ and $\frac{\partial f}{\partial (\mathbf{Z}^*)_{ij}}$ in a row vector, and define $\mathscr{D}_{\mathbf{Z}} f(\mathbf{Z})$ and $\mathscr{D}_{\mathbf{Z}^*} f(\mathbf{Z})$ at $\mathbf{Z}_0$ as

$$\begin{aligned} \mathscr{D}_{\mathbf{Z}} f(\mathbf{Z}_0) &\triangleq \frac{\partial f(\mathbf{Z})}{\partial \mathrm{vec}^T(\mathbf{Z})} = \mathrm{vec}^T\left(\nabla_{\mathbf{Z}} f(\mathbf{Z}_0)\right), \\ \mathscr{D}_{\mathbf{Z}^*} f(\mathbf{Z}_0) &\triangleq \frac{\partial f(\mathbf{Z})}{\partial \mathrm{vec}^T(\mathbf{Z}^*)} = \mathrm{vec}^T\left(\nabla_{\mathbf{Z}^*} f(\mathbf{Z}_0)\right), \end{aligned} \tag{3.7}$$

and then we have $df = \mathscr{D}_{\mathbf{Z}} f(\mathbf{Z}_0) d\mathrm{vec}(\mathbf{Z}) + \mathscr{D}_{\mathbf{Z}^*} f(\mathbf{Z}_0) d\mathrm{vec}(\mathbf{Z}^*)$ which leads to the following Jacobian matrices of $f_i(\mathbf{P}_i, \mathbf{P}_{-i})$:

$$\mathscr{D}_{\mathbf{P_i}} f_i(\mathbf{P}_i, \mathbf{P}_{-i}) = \mathrm{vec}^T\left(\sum_{j \neq i} -w_j \mathbf{M}_j\right)^T, \tag{3.8}$$

recalling $\mathscr{D}_{\mathbf{P_i}} f_i(\mathbf{P}_i, \mathbf{P}_{-i}) = \mathrm{vec}^T(\nabla_{\mathbf{P}_i} f_i(\mathbf{P}_i, \mathbf{P}_{-i}))$, the partial derivative of $f_i(\mathbf{P}_i, \mathbf{P}_{-i})$ with respect to $\mathbf{P}_i$, evaluated at $\mathbf{P}_i = \bar{\mathbf{P}}_i$ is then given by

$$\nabla_{\mathbf{P}_i} f_i(\mathbf{P}_i, \mathbf{P}_{-i})|_{\mathbf{P}_i = \bar{\mathbf{P}}_i} = -\sum_{j \neq i} w_j \mathbf{M}_j^T|_{\mathbf{P}_i = \bar{\mathbf{P}}_i} \tag{3.9}$$

Retaining only the linear term in the Taylor's expansion of $f_i(\mathbf{P}_i, \mathbf{P}_{-i})$ around $\bar{\mathbf{P}}_i$ and adding a proximal like regularization term (in order to guarantee strong convexity and enhancement of the convergence speed), it is possible to approximate the objective function in (3.4) by

$$\begin{aligned} \widetilde{f}_i(\mathbf{P}_i; \bar{\mathbf{P}}_i) \triangleq\ & w_i \log_2 \det(\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i) \\ & - \mathrm{Tr}\left(\sum_{j \neq i} w_j \mathbf{M}_j \Big|_{\mathbf{P}_i = \bar{\mathbf{P}}_i} \mathbf{P}_i\right) \\ & - \frac{\tau_i}{2} \mathrm{vec}(\mathbf{P}_i - \bar{\mathbf{P}}_i)^T \mathbf{H}_i(\bar{\mathbf{P}}_i) \mathrm{vec}(\mathbf{P}_i - \bar{\mathbf{P}}_i) \end{aligned} \tag{3.10}$$

where $\mathbf{C}_i \triangleq \mathbf{V}_i^H \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{V}_i$, τ_i is a given nonnegative constant and $\mathbf{H}_i(\bar{\mathbf{P}}_i)$ is an $n_i \times n_i$ uniformly positive definite matrix (possibly dependent on $\bar{\mathbf{P}}_i$), i.e., $\mathbf{H}_i(\bar{\mathbf{P}}_i) - c_H \mathbf{I} \succeq \mathbf{0}$, for all $\bar{\mathbf{P}}_i \in \mathscr{K}$ and some positive c_H. It is worth mentioning that $\mathbf{H}_i(\cdot)$ can play the role of the Hessian matrix of

function $f_i(\cdot)$, defined as $\mathscr{H}_{\mathbf{X}} f(\mathbf{X}) \triangleq \frac{\partial^2 f(\mathbf{X})}{\partial \mathrm{vec}^T(\mathbf{X}) \partial \mathrm{vec}(\mathbf{X})}$ and $\tau_i = 0$ if the convex part of the original objective function is uniformly strongly convex on $\mathscr{K}$. For calculating the Hessian matrix, we need to calculate the $d^2 f_i(\mathbf{P}_i, \bar{\mathbf{P}}_i)$ which can be expressed as follow

$$
\begin{aligned}
d^2 f_i(\mathbf{P}_i, \bar{\mathbf{P}}_i) &= \mathrm{Tr}\{-(d\mathbf{Z}_j^{-1})\mathbf{R}_j^{-1}\mathbf{H}_{ji}\mathbf{V}_i(d\mathbf{P}_i)\mathbf{V}_i^H\mathbf{H}_{ji}^H\mathbf{R}_j^{-1}\chi_j \\
&- \mathbf{Z}_j^{-1}(d\mathbf{R}_j^{-1})\mathbf{H}_{ji}\mathbf{V}_i(d\mathbf{P}_i)\mathbf{V}_i^H\mathbf{H}_{ji}^H\mathbf{R}_j^{-1}\chi_j \\
&- \mathbf{Z}_j^{-1}\mathbf{R}_j^{-1}\mathbf{H}_{ji}\mathbf{V}_i(d\mathbf{P}_i)\mathbf{V}_i^H\mathbf{H}_{ji}^H(d\mathbf{R}_j^{-1})\chi_j\} \\
&= \mathrm{Tr}\{-\mathbf{Z}_j^{-1}\mathbf{Y}_j(d\mathbf{P}_i)\mathbf{Y}_j^H\chi_j\mathbf{Z}_j^{-1}\mathbf{Y}_j(d\mathbf{P}_i)\mathbf{Y}_j^H\chi_j \\
&+ \mathbf{Z}_j^{-1}\mathbf{Y}_j(d\mathbf{P}_i)\mathbf{Y}_j^H\mathbf{H}_{ji}^H\mathbf{V}_i(d\mathbf{P}_i)\mathbf{Y}_j^H\chi_j \\
&+ \mathbf{Z}_j^{-1}\mathbf{Y}_j(d\mathbf{P}_i)\mathbf{Y}_j^H\mathbf{H}_{ji}\mathbf{V}_i(d\mathbf{P}_i)\mathbf{Y}_j^H\chi_j\} \\
&= \mathrm{Tr}\left(\mathbf{M}_j(d\mathbf{P}_i)(2\mathbf{S}_j - \mathbf{M}_j)(d\mathbf{P}_i)\right) \\
&= \mathrm{Tr}\left((d\mathbf{P}_i)(2\mathbf{S}_j - \mathbf{M}_j)(d\mathbf{P}_i)\mathbf{M}_j\right) \\
&= \mathrm{vec}^T(d\mathbf{P}_\mathbf{i}^\mathbf{T})\mathrm{vec}((2\mathbf{S}_j - \mathbf{M}_j)(d\mathbf{P}_i)\mathbf{M}_j) \\
&\overset{(a)}{=} \mathrm{vec}^T(d\mathbf{P}_\mathbf{i}^\mathbf{T})\left[\mathbf{M}_j^T \otimes (2\mathbf{S}_j - \mathbf{M}_j)\right]\mathrm{vec}(d\mathbf{P}_i) \\
&= [d\mathrm{vec}(\mathbf{P}_\mathbf{i})]^T K_{N_tN_t}\left(\mathbf{M}_j^T \otimes (2\mathbf{S}_j - \mathbf{M}_j)\right)[d\mathrm{vec}(\mathbf{P}_\mathbf{i})]
\end{aligned}
$$

where $\mathbf{S}_j \triangleq \mathbf{Y}_j^H\mathbf{H}_{ji}\mathbf{V}_i$, (a) follows from the property $\mathrm{vec}(\mathbf{XYZ}) = (\mathbf{Z}^T \otimes \mathbf{X})\mathrm{vec}(\mathbf{Y})$ [63, Lemma 2.11], and we introduced the commutation matrix $\mathbf{K}_{N_t^2}$ which is the $N_t^2 \times N_t^2$ permutation matrix such that $\mathrm{vec}(\mathbf{X}^T) = \mathbf{K}_{N_t^2}\mathrm{vec}(\mathbf{X})$ [63, Def. 1.8]. Then, using the identification rule [63, Table 3.3], the Hessian matrix of function $f_i(\mathbf{P}_i, \bar{\mathbf{P}}_i)$ is given by

$$
\mathscr{H}_{\mathbf{P}_i} f_i(\mathbf{P}_i, \bar{\mathbf{P}}_i) = \frac{w_j}{2} K_{N_tN_t}\Bigg(\mathbf{M}_j^T \otimes (2\mathbf{S}_j - \mathbf{M}_j) + (2\mathbf{S}_j - \mathbf{M}_j)^T \otimes \mathbf{M}_j\Bigg). \tag{3.11}
$$

in which, we use the properties of the commutation matrix where $\mathbf{K}_{mn} = \mathbf{K}_{mn}^T$ and $\mathbf{K}_{pm}(\mathbf{A}_{m\times n} \otimes$

$\mathbf{B}_{p\times q}) = (\mathbf{B}_{p\times q} \otimes \mathbf{A}_{m\times n})\mathbf{K}_{qn}$. Now it is possible to approximate (3.4) by a set of L per link problems given for $i \in \mathscr{L}$ by

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad \widetilde{f}_i(\mathbf{P}_i; \bar{\mathbf{P}}_i) \tag{3.12}$$

Associated with each $\widetilde{f}(\mathbf{P}_i; \bar{\mathbf{P}}_i)$ we have the following best response mapping for each user using the proposed algorithm in [62], which consists then in solving iteratively (possibly with a memory) the following sequence of a (strongly) convex optimization problem

$$\widehat{\mathbf{P}}_i(\bar{\mathbf{P}}_i, \tau_i) \triangleq \arg\max_{\mathbf{P}_i \in \mathscr{K}} \widetilde{f}(\mathbf{P}_i; \bar{\mathbf{P}}_i). \tag{3.13}$$

Unlike (3.4), (3.13) is strongly convex in $\mathbf{P}_i$, then it has a unique solution and can be efficiently solved by numerical iterative algorithms in which each user updates its strategy based on the best-response $\widehat{\mathbf{P}}_i(\cdot, \tau)$.

3.3.1 Some Special Cases

In this subsection, we analyze the network proportional-fair metric expression in high and low SNR regimes and consider some special cases of $\widetilde{f}_i(\mathbf{P}_i; \bar{\mathbf{P}}_i)$.

1) $\tau_i = 0$**.**

If $\tau_i = 0$, then we can rewrite the optimization problem (3.12) as follow

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad w_i \log_2 \det(\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i) - \mathrm{Tr}\,(\mathbf{E}_i \mathbf{P}_i) \tag{3.14}$$

where $\mathbf{E}_i \triangleq \sum_{j\neq i} w_j \mathbf{M}_j|_{\mathbf{P}_i = \bar{\mathbf{P}}_i}$. Note that (3.14) essentially maximizes the same objective function as (3.4) with respect to $\mathbf{P}_i$, except that the proportional-fair metric of the other link is approximated to the first order at the point $\bar{\mathbf{P}}_i$. The trace term in (3.14) plays the role of interference tax, discouraging selfish behavior of link i, which would otherwise just want to maximize its own rate (if the trace term be equal to zero, each link selfishly maximizes its own rate). We can reconsider it as follows: for a given LTE-Advanced CoMP-CB system, in each iteration, i-th MS announces an

interference price to all BSs, which is the marginal decrease in utility for an increase in received interference. The transmitters update their power to maximize their own utility minus the cost of interference they produce.

Theorem 5. *In an LTE-Advanced CoMP-CB system in which the proportional-fair metric of all the interference links are approximated to the first order (at the point $\bar{\mathbf{P}}_i$), the closed-form solution of power allocation for CoMP transmissions is available and can be expressed as*

$$\mathbf{P}_i(k,k) = \left[\frac{w_i\mathbf{C}_i(k,k) - \mathbf{G}_i(k,k)}{\mathbf{C}_i(k,k)\mathbf{G}_i(k,k)}\right]^+ \tag{3.15}$$

where $\mathbf{C}_i \triangleq \mathbf{V}_i^H\mathbf{H}_{ii}^H\mathbf{R}_i^{-1}\mathbf{H}_{ii}\mathbf{V}_i$, $\mathbf{G}_i \triangleq \mathbf{V}_i^H(\tilde{\mathbf{E}}_i + \mu_i\mathbf{I})\mathbf{V}_i$, and $\tilde{\mathbf{E}}_i \triangleq \sum_{j\neq i} w_j\mathbf{H}_{ji}^H\mathbf{R}_j^{-1}\chi_j\mathbf{Z}_j^{-1}\mathbf{R}_j^{-1}\mathbf{H}_{ji}$ evaluated at $\mathbf{P}_i = \bar{\mathbf{P}}_i$. The remaining elements of $\mathbf{P}_i$ are zero.

Proof. See Appendix A.0.5. □

Theorem 6. *An optimal solution of downlink precoding matrices for the aforementioned system is the generalized eigenmatrix of $\mathbf{H}_{ii}^H\mathbf{R}_i^{-1}\mathbf{H}_{ii}$ and $\tilde{\mathbf{E}}_i + \mu_i\mathbf{I}$, with $\mathbf{H}_{ii}^H\mathbf{R}_i^{-1}\mathbf{H}_{ii}\tilde{\mathbf{V}}_i = (\tilde{\mathbf{E}}_i + \mu_i\mathbf{I})\tilde{\mathbf{V}}_i\Lambda_i$, where $\tilde{\mathbf{V}}_i = \mathbf{V}_i\mathbf{P}_i^{1/2}$ is an unnormalized transmit precoder of i-th BS and the elements of the diagonal matrix Λ_i are the generalized eigenvalues of $\mathbf{H}_{ii}^H\mathbf{R}_i^{-1}\mathbf{H}_{ii}$ and $\tilde{\mathbf{E}}_i + \mu_i\mathbf{I}$.*

Proof. See Appendix A.0.6. □

2) High SNR.

In the high SNR regime, the achievable rate $R_i(\mathbf{P}_i;\bar{\mathbf{P}}_i)$ can be approximated by $\tilde{R}_i(\mathbf{P}_i;\bar{\mathbf{P}}_i) = \log_2\det(\chi_i) - \log_2\det(\mathbf{R}_i)$. Then, one can formulate the high SNR proportional-fair maximization problem as follows

$$\begin{aligned} \underset{\mathbf{P}_i}{\text{maximize}} \quad & \sum_{i=1}^{K} w_i\tilde{R}_i(\mathbf{P}_i;\bar{\mathbf{P}}_i) \\ \text{subject to} \quad & \mathbf{P}_i \succ \mathbf{0},\ \mathrm{Tr}(\mathbf{P}_i) \leq p_{i,\max} \end{aligned} \tag{3.16}$$

It is worth mentioning that, if we focus on optimizing the interference subspace, thus, by dropping

the signal term in $\tilde{R}_i$, we can bound it as follows

$$\begin{aligned}\tilde{R}_i \geqslant -\log\det(\mathbf{R}_i) &\overset{(a)}{\geqslant} \sum_{k=1}^{n} -\log([\mathbf{R}_i]_{kk}) \\ &\overset{(b)}{>} -\sum_{k=1}^{n} [\mathbf{R}_i]_{kk} = -\mathrm{Tr}(\mathbf{R}_i)\end{aligned} \tag{3.17}$$

where (a) comes from applying Hadamard's inequality, i.e., $\det(\mathbf{M}) \leqslant \prod_j \mathbf{M}_{jj}$ for $\mathbf{M} \succeq \mathbf{0}$, (b) follows from the fact that $\forall x > 0,\ x > \log(x)$, and $n = \min(Nr, Nt)$. It shows that minimizing the interference leakage at each user results in optimizing a lower bound on the user's high SNR rate.

Back to the optimization problem at high SNR regimes, and by retaining only the linear term in the Taylor's expansion of nonconvex part of the above objective function around $\bar{\mathbf{P}}_i$, it is possible to approximate (3.16) by a set of L per-link problems given for $i \in \mathscr{L}$ by

$$\begin{aligned}\underset{\mathbf{P}_i}{\text{maximize}} \quad & w_i \log_2\det(\chi_i) - \mathrm{Tr}(\mathbf{A}_i\mathbf{P}_i) \\ \text{subject to} \quad & \mathrm{Tr}(\mathbf{P}_i) \leq p_{i,\max},\ \ \mathbf{P}_i \succ 0,\end{aligned} \tag{3.18}$$

where $\mathbf{A}_i \triangleq \sum_{j\neq i} w_j \mathbf{S}_j$ evaluated at $\mathbf{P}_i = \bar{\mathbf{P}}_i$. Note that unlike (3.16), (3.18) is convex in $\mathbf{P}_i$ and can be efficiently solved by numerical iterative algorithms.

Theorem 7. *In the high SNR regime, the power allocation of each BS in an LTE-Advanced CoMP-CB system can expressed as follows*

$$\mathbf{P}_i = \mathbf{U}^H \,\mathrm{diag}\left[\frac{w_i}{\sigma_k + \mu_i}\right]\mathbf{U} \tag{3.19}$$

where $\mathbf{U}$ *is a unitary matrix and can be achieved by eigenvector decomposition of* $\mathbf{A}_i$*.*

Proof. See Appendix A.0.7. □

3) Low SNR.

In the low SNR regime, we can replace $\log_2\det(\mathbf{I}_{N_r} + \mathbf{R}_i^{-1/2}\chi_i\mathbf{R}_i^{-1/2})$ with $\sum_{i=1}\log(1+\lambda_i)$, where λ_i is the $i-$th eigenvalues of $\mathbf{R}_i^{-1/2}\chi_i\mathbf{R}_i^{-1/2}$. Now, we use the fact that χ_i is small, it implies

λ_i are small, so to first order we have $\log(1+\lambda_i) \approx \lambda_i$. Using this first-order approximation in the expression above, we get $\sum_{i=1} \log(1+\lambda_i) = \sum_{i=1} \lambda_i = \mathrm{Tr}(\mathbf{R}_i^{-1/2} \chi_i \mathbf{R}_i^{-1/2}) = \mathrm{Tr}(\mathbf{R}_i^{-1} \chi_i)$. Thus, we have

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad \sum_{i=1}^{K} w_i \mathrm{Tr}(\mathbf{R}_i^{-1} \chi_i) \tag{3.20}$$

which is a nonconvex optimization problem. By retaining only the linear term in the Taylor's expansion of nonconvex part of the above objective function around $\bar{\mathbf{P}}_i$, it is possible to approximate (3.20) by a set of L per-link problems given for $i \in \mathscr{L}$ by

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad \mathrm{Tr}\left((w_i \mathbf{C}_i - \mathbf{D}_i)\mathbf{P}_i\right) \tag{3.21}$$

where $\mathbf{D}_i \triangleq \sum_{j \neq i} w_j \mathbf{Y}_j^H \chi_j \mathbf{Y}_j$ evaluated at $\mathbf{P}_i = \bar{\mathbf{P}}_i$. Note that unlike (3.20), (3.21) is convex (a Semidefinite Programming (SDP)) in $\mathbf{P}_i$ and is therefore amenable to a wide variety of optimization techniques.

Theorem 8. *In the low SNR regime, the power allocation of each BS in an LTE-Advanced system can expressed as follows*

$$\mathbf{P}_i = \mathbf{V}^H \,\mathrm{diag}\left[\frac{\varepsilon}{\gamma_k + \mu_i}\right] \mathbf{V} \tag{3.22}$$

where $\mathbf{V}$ *is a unitary matrix and can be achieved by eigenvector decomposition of matrix* $-w_i \mathbf{C}_i + \mathbf{D}_i$.

Proof. See Appendix A.0.8. □

Remark 5. *It is worth mentioning that in the low SNR regime, the interference due to other BSs is overwhelmed by the noise power seen at the MSs. The proportional-fair metric maximizing beamformers in this regime are simply the* N_s *dominant right singular vectors obtained from the singular value decomposition of the direct link* $\mathbf{H}_{ii}$ *of the i-th BS. The MSs are the corresponding* N_s *left singular vectors. The power allocation strategy reduces to that of single-user MIMO scenario, i.e., water filling on the corresponding* N_s *dominant singular values. Which is not surprising*

$$\widehat{\mathbf{P}}_i(\bar{\mathbf{P}}_i, \tau_i) \triangleq \arg\max_{\mathbf{P}_i \in \mathscr{K}} w_i \log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i\right) - \mathrm{Tr}\left(\mathbf{E}_i \mathbf{P}_i\right) - \tau_i \|\mathbf{P}_i - \bar{\mathbf{P}}_i\|_F^2. \tag{3.23}$$

because when the noise dominates the received signal, the benefit of self interference cancelation is marginal.

4) $\mathbf{H}_i(\bar{\mathbf{P}}_i) = \mathbf{I}$.

If $\mathbf{H}_i(\bar{\mathbf{P}}_i) = \mathbf{I}$, the quadratic term in (3.10) reduces to the standard proximal regularization $\tau_i \|\mathbf{P}_i - \bar{\mathbf{P}}_i\|_F^2$, and then the best response matrix function of each BS is given by Eqn. (3.23), at the top of the next page.

Theorem 9. *If $\mathbf{H}_i(\bar{\mathbf{P}}_i) = \mathbf{I}$, then the closed-form solution of the above-mentioned optimization problem can be expressed as*

$$\mathbf{P}_i = \left[\bar{\mathbf{P}}_i - \frac{1}{2\tau_i}(\mu^* \mathbf{I} + \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix})\right]^+ \tag{3.24}$$

where $\mathbf{Z}$ is the matrix of lagrangian multipliers associated to the linear constraints, $[\mathbf{X}]^+$ denotes the projection of $\mathbf{X}$ onto the cone of positive semidefinite matrices, and μ^* is the multiplier which can be found by bisection.

Proof. See Appendix A.0.9. □

3.3.2 Complexity Analysis

In this section, we compare the computational complexity of our algorithm with that of the WMMSE algorithm for coordinated beamforming [55]. For simplicity, let us assume that all channel matrices of $\mathbf{H}_{ii}$ are full-column rank and set $\tau_i = 0$ in (3.13). Moreover, let L be the total number of users in the system and N_t and N_r denote the number of antennas at each transmitter and receiver, respectively. Since both the proposed proportional-fair resource allocation algorithm and the WMMSE

algorithm include a similar bisection step which generally takes few iterations, we ignore this bisection step in the computational complexity analysis. We compare the algorithms by given per iteration complexity, where an iteration of both algorithm means one round of updating all users' beamforming or covariance matrices. Under these conditions, each iteration of the WMMSE algorithm involves the computation of the MMSE receiver matrices and their corresponding MSE matrices in [55], i.e., $\mathbf{U}_{i_k}$ in (5) and $\mathbf{E}_{i_k}$ in (6) of [55]. To determine these matrices in the WMMSE algorithm, we need to first calculate the covariance matrix of the total received signal at each receiver and then compute their sum. Consequently, the per-iteration complexity of WMMSE algorithm is $\mathscr{O}(L^2N_tN_r^2+L^2N_t^2N_r+L^2N_t^3+LN_r^3)$. Using a similar analysis, the per-iteration complexity of the proposed algorithm can be shown to be $\mathscr{O}(L^2N_tN_r^2+L^2N_t^2N_r+LN_t^3+LN_r^3)$.

3.3.3 Convergence

The following propositions link the solution of the original problem (3.4) with that of its convexified approximation (3.13) and characterize the convergence of the proposed proportional-fair resource allocation algorithm.

Proposition 1. *A fixed point of the proposed algorithm exists, and it is a stationary solution of the original problem. Thus, if the algorithm converges, it converges to a stationary solution of (3.4).*

Proof. The proof is upon checking that the set of the first order conditions (Karush-Kuhn-Tucker (KKT) conditions) for (3.13) amount to precisely the KKT conditions for (3.4). □

Proposition 2. *The proposed proportional-fair resource allocation algorithm converges.*

Proof. The proof is inspired by [64] and follows from showing that after a link solves the convexified optimization problem, the objective function of the original nonconvex optimization problem is nondecreasing. Since the maximum network proportional-fair metric is bounded from above, value of the objective function converges. To this end, suppose that $\mathbf{P}_j = \bar{\mathbf{P}}_j$ and $\mathbf{V}_j = \bar{\mathbf{V}}_j$ for all $j \in \mathscr{L}$, from the previous iteration. Let $\mathbf{P}_i^*$ and $\mathbf{V}_i^*$ denote the optimal solution of the convexified problem for link i and define $\mathbf{Q}_i(\mathbf{V}_i, \mathbf{P}_i) = \mathbf{V}_i\mathbf{P}_i\mathbf{V}_i^H$. Then the original objective function can be

$$\sum_{j=1}^{L} w_j R_j(\mathbf{Q}_i^*, \bar{\mathbf{Q}}_{-i}) = w_i R_i(\mathbf{Q}_i^*, \bar{\mathbf{Q}}_{-i}) + f_i(\mathbf{Q}_i^*, \bar{\mathbf{Q}}_{-i}) \overset{(a)}{\geq} w_i R_i(\mathbf{Q}_i^*, \bar{\mathbf{Q}}_{-i}) + f_i(\bar{\mathbf{Q}}_i, \bar{\mathbf{Q}}_{-i}) \tag{3.25}$$
$$+ \mathrm{Tr}\left(\nabla f_i(\bar{\mathbf{Q}}_i, \bar{\mathbf{Q}}_{-i})^T \left(\mathbf{Q}_i^* - \bar{\mathbf{Q}}_i \right) \right) \overset{(b)}{\geq} w_i R_i(\bar{\mathbf{Q}}_i, \bar{\mathbf{Q}}_{-i}) + f_i(\bar{\mathbf{Q}}_i, \bar{\mathbf{Q}}_{-i}) = \sum_{j=1}^{L} w_j R_j(\bar{\mathbf{Q}}_i, \bar{\mathbf{Q}}_{-i})$$

rewritten as Eqn. (3.25), at the top of the next page, where (a) comes from the fact that $f_i(\mathbf{Q}_i, \mathbf{Q}_{-i})$ is a convex function with respect to $\mathbf{Q}_i$ (see Appendix F), and (b) holds since $\mathbf{Q}_i^*$ is the optimal solution of the convexified optimization problem. Thus, the original nonconvex objective function is nondecreasing after each link i updates $\mathbf{P}_i$ and $\mathbf{V}_i$. Since the objective is bounded from above, the algorithm must converge. □

3.3.4 System Design Issues

Channel state information. To enable the CoMP-CB operation within the cooperating set, perfect CSI is assumed to be available at each BS (similar to [55]). In modern communication systems, CSI can usually be obtained at the BS through the following methods. (a) In some contexts channel reciprocity can be exploited to acquire CSI at the transmitters. To be specific, in a time division duplex (TDD) system, the downlink channel state information can be directly obtained at the transmitter through uplink sounding due to channel reciprocity; for frequency division duplex (FDD) systems, there is also certain reciprocity relationship between uplink and downlink channels [65]. (b) Feedback channels are often available in wireless systems for the receiver to feedback the downlink channel to the transmitter. (c) Learning mechanisms can be exploited to iteratively learn the required CSI. In Rel-10 LTE-Advanced networks, channel reciprocity is utilized for TDD systems to work under perfect CSI while codebook based feedback is utilized for FDD systems to obtain the CSI at the transmitter. The perfect CSI assumption provides us with an achievable upper bound on the transmission rate for each user.

Proportional-Fair Weight. The proportional-fair weight is adopted in (3.4) to achieve a good trade-off between cell-average and cell-edge spectral-efficiency. As discussed in Section II-B, in

time slot t, $w_i = R_i^{\alpha}(t)/T_i^{\beta}(t)$. From (3), it is clear that $R_i(t)$ depends on $\mathbf{V}_i$ and $\mathbf{P}_i$ which in turn depend on w_i. Therefore, the proportional-fair weight for MS i, w_i, should be obtained by an iterative method. In each slot, w_i is initialized to some nonnegative value, and then is updated iteratively until it converges. We repeat the iteration until the distance between two consecutive rate is within a predefined threshold. It is worth to note that each iteration only takes a short time to complete. The iteration process is carried out according to Algorithm I.

3.4 Simulation Evaluations

In this preliminary experimental evaluation, the performance of the proposed schemes for downlink CoMP coordinated beamforming systems is evaluated. We consider a possible antenna configuration in a typical deployment scenario for LTE-Advanced: 4 transmit antennas at the base station and 2 receive antennas at each mobile station. We assume that the system consists of two cells which each cell serves its own MS. The simulation is run for 1000 channel realization where each channel element is drown i.i.d. from a real Gaussian distribution with zero mean and a variance of 1.

Fig. 1 shows the evolution of the network proportional-fair metric when optimization problem (14) is solved via the introduced algorithm using Jacobi successive convex approximation scheme. Two sets of weights $w = [0.9\ 0.1]$ and $\bar{w} = [0.65\ 0.35]$ were tested. The monotonic increase of the network proportional-fair metric can be verified.

In Fig. 2-a, we plot the network proportional-fair metric versus the number of iterations. Our experiment shows that for a two cooperating BS CoMP LTE-Advanced system, with $N_t = 4$ and $N_r = 2$, the proposed algorithm has the monotonic convergence behavior. In Fig. 2-b and c, we consider the case of $w_2 \gg w_1$, since $\mathbf{P}_i$ is a function of w_i the BS 2 has the maximum proportional-fair metric and the optimal power allocation becomes a water-filling one over the spatial directions associated with the singular values of BS 2's channel transfer matrix.

In Fig. 3-a we plot the network proportional-fair metric versus the number of iterations. As

expected, this figure shows that the proposed algorithm has the monotonic convergence behavior and the network proportional-fair metric is seen to converge to the near optimal values quite rapidly. Moreover, in Fig. 3, we consider the case of $w_1 \gg w_2$, the number of transmit and receive antennas are the same and equal to 8 and $p_{1,\max} = p_{2,\max} = 8W$, respectively. Since $w_1 \gg w_2$, the optimal power allocation becomes a water-filling one. Assuming CSI is shared between BSs, overall throughput is achieved by implementing a water-filling power allocation scheme over the spatial directions associated with the singular values of BS 1 channel transfer matrix.

The setup of our experiments for the rest of simulation is the following. We simulated seven cells with multiple users that randomly and uniformly dropped (at a distance $>$ 35 m and $<$ 275 m from the BS) in each cell. The transmission is subject to interference from 6 neighboring base stations. The transmit power at each base station is fixed to 46 dBm and the noise variance at the mobile station is fixed to -174 dBm. System bandwidth is taken as 5 MHz. The BSs are all equipped with $N_t = 4$ transmit antennas and the MSs are equipped with $N_r = 2$ receive antennas. All deployments and channel model parameters are listed in Table I.

The performance is measured in terms of the cumulative distribution function (CDF) of the user average rate (Mbits/sec). To do so, we generate 500 drops for each user. At each time instant, the instantaneous rate achieved by each user is recorded. At the end of the drop, the average rate achieved by each user is computed (as an average over all instantaneous rates). According to these

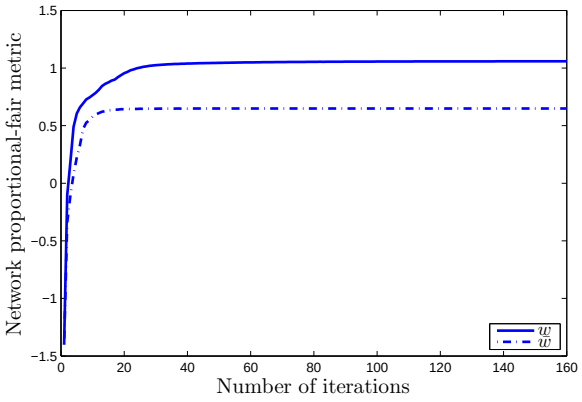

Figure 3.1. The Network proportional-fair metric versus Number of iterations.

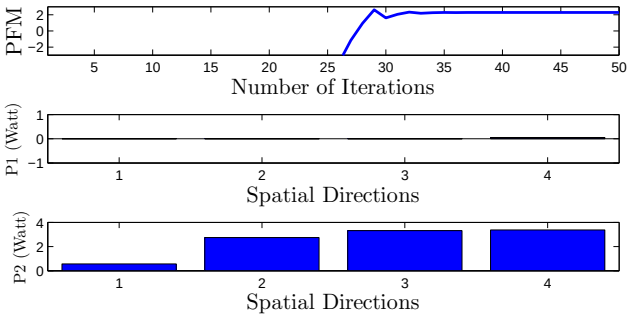

Figure 3.2. (a) The network proportional-fair metric (PFM) versus number of iterations in a 2 cooperating BS CoMP LTE-Advanced system, with $N_t = 4$ and $N_r = 2$, $w = [0.01\ 0.99]$, and $p_{1,\max} = p_{2,\max} = 12\ Watt$, (b) and (c) Optimal power allocation at BS 1 and BS 2 versus spatial directions (SD) associated with the singular values of their channel transfer matrix.

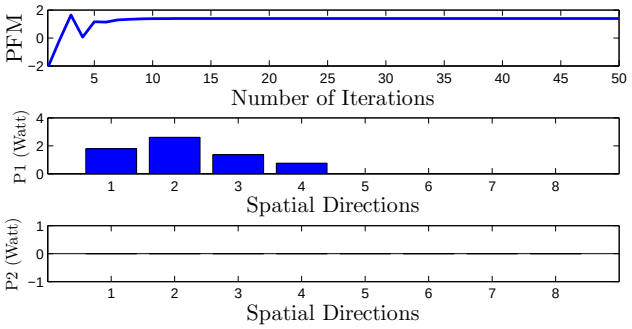

Figure 3.3. (a) The network proportional-fair metric (PFM) versus number of iterations in a 2 cooperating BS CoMP LTE-Advanced system, with $N_t = N_r = 8$, $w = [0.999\ 0.001]$, and $p_{1,\max} = p_{2,\max} = 8\ Watt$, (b) and (c) Optimal power allocation at BS 1 and BS 2 versus spatial directions (SD) associated with the singular values of their channel transfer matrix.

available values of user average rate, the CDF is computed. The CDF performance of the user average rate of the proposed scheme is compared with that of the non-cooperation scenario and that of WMMSE algorithm in Fig. 4. Proportional-fairness resource allocation algorithm is used in all the schemes. The non-cooperative algorithm is considered as a reference scheme. In this case, each BS does not try to reduce the interference that it induces to other MSs in determining its resource allocation matrices. To compare proposed algorithm with WMMSE, both algorithms are initialized by choosing the same feasible randomly generated point, and are terminated when (the absolute value) of the sum-rate error in two consecutive rounds becomes smaller than 10^{-2}. The accuracy in the bisection loops (required by both algorithms) is set to 10^{-3}.

From Fig. 4, it is clear that the proposed scheme demonstrates significant performance improvements over non-CoMP and WMMSE strategies. To be specific, users have much higher probability to achieve higher user average rate under the proposed resource allocation strategy. The corresponding cell-average spectral-efficiency and cell-edge spectral-efficiency are also obtained and summarized in Table II. From Table II, it is clear that both WMMSE and our scheme could provide significant performance improvements over non-CoMP scheme. To be specific, the WMMSE scheme provides a gain of 39.42% in cell-average spectral-efficiency, a gain of 241% in 10%-tile

Table 3.1: System parameters used in simulation.

Parameters	Values
Cell Layout	Hexagonal grid
Number of transmit antenna	4
Number of receive antenna	2
Number of MSs per cell	500
Inter-site distance	500 m
Minimum distance between UT and BS	$>$ 35 m
MS distribution	Uniform random distribution
Bandwidth	5 MHz
Maximum Transmission Power	43 dBm
Thermal Noise Level	-174 dBm
Doppler Spread	100 Hz
Coherence Time	10 msec
Path Loss (dB)	$128.1 + 37.6\log_{10}(d)$ with d in Km
Shadowing model	Log-normal shadowing with 4dB standard deviation.

cell-edge spectral-efficiency, and a gain of 330% in 5%-tile cell-edge spectral-efficiency over the non-CoMP scheme. Meanwhile, the proposed proportional-fair resource allocation strategy could even provide spectral-efficiency gains over WMMSE. To be specific, the proposed scheme provides a gain of 8.23% in cell-average spectral-efficiency, a gain of 35.1% in 10%-tile cell-edge spectral-efficiency, and a gain of 49.2% in 5%-tile cell-edge spectral-efficiency over the WMMSE algorithm. These results suggest that the proposed proportional-fair resource allocation scheme outperforms existing schemes and has the capability of sufficiently suppressing the inter-cell interference and simultaneously improving the cell-average and the cell-edge performance.

The cumulative distributive function of signal-to-noise-plus-interference ratios (SINRs) for different transmission schemes are shown in Fig. 5. The proposed proportional-fair resource allocation scheme shows significant performance gains over its counterparts in the low SINR regime where cell-edge users usually operate. Moreover, the proposed scheme also shows considerable performance gains over its counterparts in relatively high SINR regime as well.

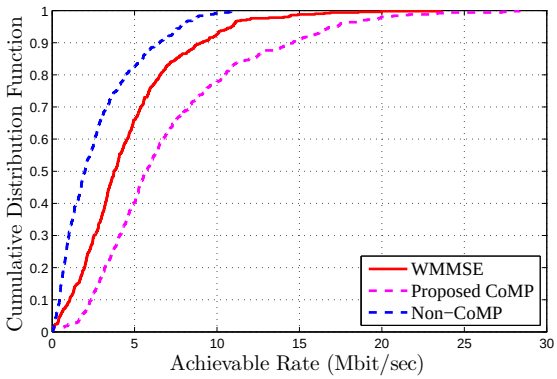

Figure 3.4. Cumulative distribution function (CDF) of achievable rate.

Table 3.2: Cell-average spectral-efficiency of different schemes

Transmission Scheme	Cell-average spectral-efficiency (bps/Hz/cell)	Cell-edge spectral-efficiency (5%-tile)	Cell-edge spectral-efficiency (10%-tile)
Non-CoMP	2.8570	0.2275	0.4075
WMMSE	3.9834	0.9782	1.3907
CoMP	4.3114	1.4593	1.8792

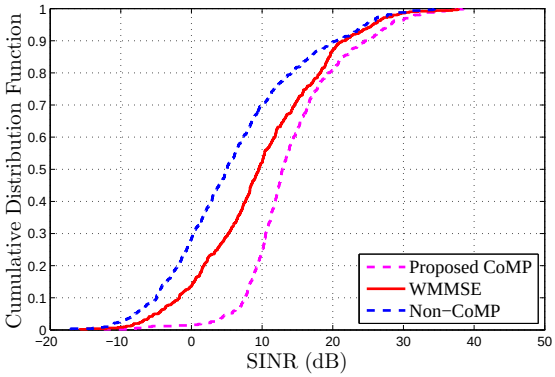

Figure 3.5. Cumulative distribution function (CDF) of SINR.

3.5 Conclusions

In this chapter, a resource allocation problem has been studied for downlink CoMP coordinated beamforming systems where each base station serves its own mobile stations. Due to the coupled interference among mobile stations the resulting optimization problem becomes nonconvex. To solve for optimal resource allocation strategies including downlink precoding and power allocation for CoMP-CB transmissions, we have introduced a stochastic parallel successive convex approximation-based algorithmic framework for a general nonconvex stochastic network proportional-fair metric optimization problem. The introduced novel decomposition enables all base stations to update their optimization variables in parallel by solving a sequence of strongly convex subproblems. Moreover, closed-form expressions of the locally optimal solution are characterized in some special cases as well as in both high and low SNR regimes. Numerical results show the introduced scheme significantly improves and optimizes the system performance by mitigating inter-cell interference.

Chapter 4

Cached Cloud-RAN: Content-Based User Association and MIMO Operation

4.1 Introduction

The big data era is being shaped with the ongoing growth of commercial data services, with mobile wireless network constituting a major data source contributor. Nowadays, wireless communication is becoming tightly integrated in our daily lives; especially with the global spread of laptops, tablets, smartphones, video streaming and online social networking applications. This globalization has paved the way to dramatically increase wireless network dimensions in terms of subscribers and amount of flowing data. Precisely, Cisco Systems forecasts that the number of mobile-connected devices per capita will reach 1.5 by 2021 and global mobile data traffic will increase sevenfold between 2016 and 2021 [1]. The volume, velocity, and variety of data from both mobile users and communication networks follow an exponential increase pattern. Consequently, big data will further be entrenched in the upcoming fifth-generation ($5G$) wireless networks.

The two important fundamental requirements for the future $5G$ wireless networks are abilities to support high data traffic and exceedingly low latency [2]. A likely candidate to fulfill these requirements is a Cloud Radio Access Network (Cloud-RAN). The network architecture of

Cloud-RAN meets the tremendous increase in data traffic while improving network throughput and energy efficiency for future networks [66–70]. This architecture spreads several low-cost low-power Base Stations (BSs) all over a small area as an alternative to a high-power BS [2]. In order to have an efficient resource allocation and interference management among multiple BSs, digital backhaul links connect all these low-power BSs to a central computing unit (cloud). Since one of the fundamental requirements for the $5G$ wireless networks is ability to support exceedingly low latency, establishing backhaul links with low latency is necessary [2]. High-speed fiber cables can achieve this requirement at the expense of an increase in infrastructure cost. Using the limited-capacity backhaul links can save cost; however, these links may result in higher latency and then, the overall performance would be much lower than that of network with high-capacity backhaul link. Therefore, the problem of interest in the Cloud-RANs is providing a new technique which can reduce the backhaul consumption.

More recently, researchers have investigated that the wireless caching is an effective way to address this issue. Caching capacity at the BSs is a new type of wireless networks' resource. The low-cost, low-complexity, and tight integration with big data analytical tools of the wireless caching will help shape future wireless big data processing. However, research on cache-enabled wireless networks is still in its infancy. The main idea behind wireless caching is equipping the BSs with inexpensive limited-size local storage units, and placing the most popular contents in them to create more Coordinated Multi-Point (CoMP) transmission opportunities while serving the users [71–76]. To do so, a cluster of BSs is assigned to each user to effectively relieve the backhaul capacity demand at the Cloud-RAN. The cutback in backhaul utilization is because of the reduced payload data transmission throughout the backhaul which is due to employing the caching capacity at each BS. Each cluster is formed by aggregating the BSs whose transmission strategies cooperatively serve the UEs within the cluster through joint proceding [71]. If the user's serving BSs cache the content that the user requests, it will be transmitted directly by the serving BSs, thereby reducing delivery latency as well as backhaul overhead. Otherwise, the content needs to be fetched into the serving BSs' caches from the cloud via the backhaul link. In such a model,

by downloading the most popular content during off-peak hours and serving users from the cache during peak hours, aside from reducing the capacity requirement of backhaul links as well as the delivery latency, CoMP opportunities also allow a reduction in the inter-cell interference [5]. Consequently, the users experience a substantial data rate improvement and the network can attain a high capacity gain. However, this wireless networks' resource is limited in comparison with the total amount of mobile data traffic.

In order to entirely utilize the benefit of wireless caching and to fully exploit the opportunity of serving users through a CoMP transmission, developing advanced caching placement strategies in Cloud-RANs is required. One way to increase the possibility for a user to access its desired content locally is designing the caching contents following some data popularity distribution, such as the Zipf distribution [77]. Moreover, user association can be regarded as an important consideration of whether a BS caches a content or not. To be specific, BSs might cache files which assigned channels to the particular user are not desirable in terms of the signal strength. In this case, although BSs are clustered to share the requested data and deliver service to a specific user, data transmission is not reliable or the high transmit power will be needed. Consequently, allocating the requested files via backhaul to BSs which have good channels to the served users will be unavoidable and such allocation via backhaul will increase the backhaul cost [2]. In a densely deployed wireless network, such as Cloud-RAN, each user can be associated with one or several BSs depending on both content availability and channel condition. As a result, jointly optimizing the user-association policy, caching placement strategy, and beamforming design can enhance the user experience.

A. Related Work

The importance of caching in the fifth-generation wireless networks was recognized in [78–86]. In [87–90], the authors considered the problem of jointly minimizing the total transmit power and backhaul traffic in wireless cooperative networks under the constraint of each user's SINR requirements and with respect to the beamforming vectors. Assuming there is a backhaul constraint per each BS, [91] considered a weighted sum rate optimization problem to design the beamforming vectors. Other than these works on unicast, [92,93] discussed the effect of caching on the multicast

beamforming in the Cloud-RAN. However, these works assumed that the cache placement is static, which means the caching placement matrix is fixed and known at the cloud.

In order to improve the efficiency of cache-enabled networks, [2, 94–96] conducted an investigation into the design of caching policy. In [94, 95], the caching problem at the small cell BSs was considered, and a caching policy was designed in such a way that the cache-hit-ratio is maximized. In order to minimize the downloading latency, [96] proposed a distributed caching algorithm. The line of works in [94–96] are further expanded in [2] for a cache-enabled Cloud-RAN system to take the tradeoff between the transmission power and the backhaul cost into consideration. In [2], the authors defined the network cost of the Cloud-RAN system as a normalized weighted sum and, minimized the network cost with respect to both the beamforming matrix and the cache placement matrix by considering the quality of service (QoS), peak transmission power, and cache capacity constraints. However, these works are only focused on designing the beamforming vectors and cache placement matrix while assuming the user association matrix is given and known at the cloud.

The user association problem that is mainly concerned with load balancing, was discussed in [97, 98]. The key point here is accounting for both wireless channels and the number of UEs connected to each BS. Based on a given caching policy, [99] designed the user association policy in a way that maximizes the average download rate. However, the aforementioned studies did not optimize the user association and cache placement jointly. As a result, the system was led to an inefficient operating point.

Designing jointly the caching strategy and the user association policy in cache-enabled wireless networks is considered in [100–105]. In order to minimize the number of requests performed by the macro BSs in a small-cell network, [100] designed a joint data caching and user association policy. To this end, [101] jointly designed user association and video caching policy by minimizing the user experienced delay while taking different quality requirements for each user into account. In order to obtain an optimal tradeoff between the content availability and the load balancing in the heterogeneous networks, an online algorithm was proposed in [102]. The complexity analysis of

joint user association and cache placement in heterogeneous networks was investigated in [103]. In order to maximize the system throughput in a coordinated small-cell cellular system, the problem of joint designing of caching, user association, and routing is discussed in [104]. Considering distinct users have different wireless channels, [105] jointly designed the caching and user association policy by minimizing the average delay of small cell UEs in a heterogeneous network. The line of these works was further expanded in [106] for a cache-enabled Cloud-RAN network to lessen the backhaul traffic by maximizing a proportional fairness network utility. However, [106] considered a single-input single output (SISO) case in which both UEs and BSs are equipped with a single antenna, and each user is only connected with one BS.

In practice, the cache placement and the content delivery (precoding and user association) usually happen in different timescales. Cache placement usually takes much longer (e.g., days or hours) than that of content delivery (e.g., seconds). Therefore, like [30], we study a mixed-timescale joint optimization, but in this case for content placement and content delivery in the cache-enabled cloud radio access networks to maximize a weighted backhaul-aware network utility function subject to the peak transmission power and cache capacity constraints at all BSs. The cache placement reduces the backhaul consumption and provides more CoMP opportunities. It is adaptive to the long-term popularity of data, therefore, the caching strategy should be adaptive to the channel statistics instead of the instantaneous channel realization in each channel coherent time. In contrast, the role of the content delivery is to guarantee to provide a better average throughput to each user and be adaptive to the instantaneous channel state information.

B. Main Contributions

This chapter differs from previously mentioned studies particularly in its aim to bring a consideration of caching along with the user association and resource allocation. We optimize the tradeoff between backhaul reduction and network throughput by maximizing a weighted backhaul-aware network utility function. Furthermore, we consider multiple real-word factors for effective content caching such as popularity distribution, caching placement from the user perspective, and temporal and spatial locality of the content demand, in order to accommodate challenging use cases with

strict quality of service requirement [107]. That's why the original problem and, therefore, the three sub-problems in our manuscript are completely different from the previous works. The main contributions of this chapter can be summarized as follows.

- For the first time, we define and maximize the network throughput as a function of caching placement strategy, user association policy, precoding vectors, probability that a file is requested by a specific user, and the distance from all connected BSs to this specific user.

- We introduce the tradeoff between network throughput and backhaul saving by combining content-based user association, MIMO transmit precoding, and cache placement to boost user experience, and identify the interaction between user association, resource allocation and cache placement for a multi-cluster multi-user cache-enabled Cloud-RAN network.

- To the best of our knowledge, there is no existing study on jointly optimizing the cache placement strategy, the user association policy, and the beamforming design. In this work, we consider a multi-cluster multi-user cache-enabled Cloud-RAN network consisting of different users with distinct file preferences, and jointly optimize the mixed-timescale optimization problem of cache placement, the user association, and the beamforming matrices which can result in significant benefit. Due to the coupled interference among UEs, the mixed-timescale joint optimization of content delivery and content placement is a non-convex optimization problem. Furthermore, the entries of cache placement and user association matrices take binary values making the optimization problem a mixed integer nonlinear programming, which is an NP-hard problem and non-tractable in practice. Since, it is highly unlikely to compute a globally optimal solution in polynomial time, our goal is to obtain a trackable near-optimal solution by developing effective suboptimal algorithms. As a consequence, the joint optimization problem decomposes into a short-term content delivery and a long-term content placement problem. Moreover, by making use of the fact that all constraints are separable, we propose an iterative algorithm which optimizes the cache placement, user association, and beamforming vectors.

- We propose an iterative novel algorithm for multi-cluster multi-user cache-enabled Cloud-RANs by leveraging the Successive Convex Approximation (SCA)-based method and the Alternating Direction Method of Multipliers (ADMM). The original non-convex optimization problem is essentially divided into three subproblems. In designing the optimal content placement, the SCA approximates this subproblem as a sequence of convex subproblems. The decomposition enables all BSs to update their optimization variables in parallel by solving a sequence of convex subproblems, one for each BS. In order to find the optimal beamformers, each subproblem can be replaced by a novel ADMM form. By solving multiple small-size subproblems, the proposed ADMM allows the updation of each step to take place in parallel. Finally, to design the user association strategy, each subproblem is reformulated as a partially dualized version of the original subproblem. By reducing the complexity, the proposed algorithm is feasible for future wireless big data processing systems.

4.2 System Model and assumptions

We consider a cache-enabled Cloud-RAN network consisting of one central computing unit (cloud), B base stations (BSs), and K user equipments (UEs) as depicted in figure 1. Table 4.1 summarized the major notations and symbols used in this chapter. The location of the BSs is modeled by a Poisson Point Process (PPP) with density λ_B while UEs are distributed around each BS independently and uniformly. We partition the area to M clusters. $|\mathscr{I}_i|$ and $|\mathscr{Q}_i|$ indicate the number of UEs and BSs in the i-th cluster, respectively, where $\mathscr{Q}_i \subseteq \{1,2,\ldots,B\}$, $\mathscr{I}_i \subseteq \{1,2,\ldots,K\}$, $\mathscr{Q}_i \cap \mathscr{Q}_j = \varnothing$, $\mathscr{I}_i \cap \mathscr{I}_j = \varnothing$, $\forall i \neq j, i, j \in \mathscr{M} \triangleq \{1,2,\ldots,M\}$. Let i_j, $i \in \mathscr{M}$ and $j \in \mathscr{Q}_i$, denote the j-th BS in the i-th cluster and i_k, $i \in \mathscr{M}$ and $k \in \mathscr{I}_i$, indicate the k-th user in the i-th cluster. BS i_j is equipped with $N_t^{i_j}$ transmit antennas and a cache that stores s_{i_j} bits of data whilst user i_k has $N_r^{i_k}$ receive antennas. The channel (propagation) coefficient between the i_j BS and the i_k user form channel matrix $\mathbf{G}_{i_k,i_j} = \sqrt{\beta_{i_k,i_j}}\mathbf{H}_{i_k,i_j} \in \mathbb{C}^{N_r^{i_k} \times N_t^{i_j}}$ where β_{i_k,i_j} is a large-scale fading coefficient that depends upon the shadowing and distance between the corresponding user and BS. The large-scale

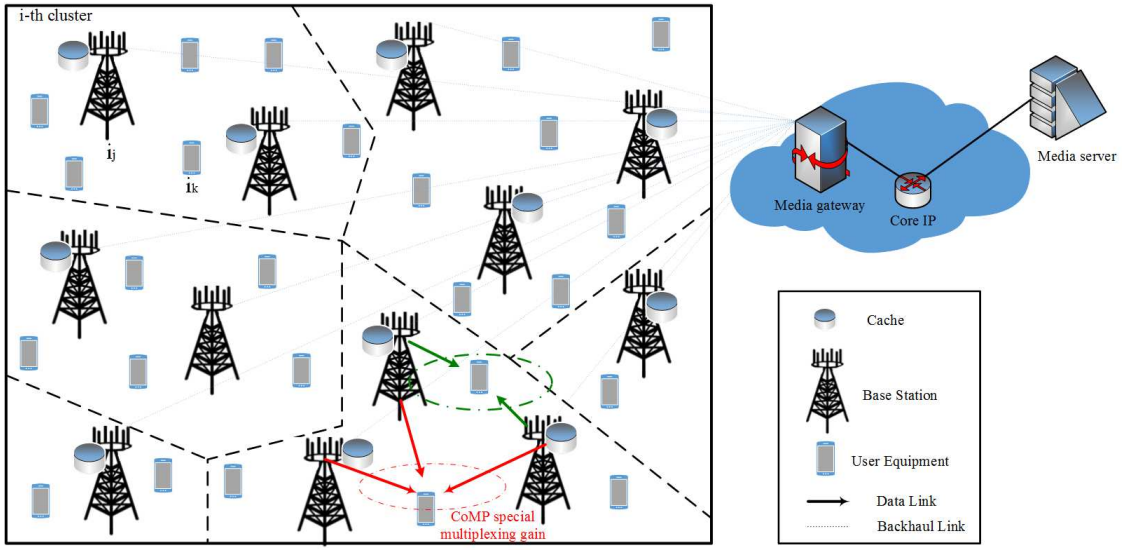

Figure 4.1. System architecture of a cache-enabled cloud radio access network.

fading coefficient denotes by $\beta_{i_k,i_j} = \psi_{i_k,i_j} d_{i_k,i_j}^{-\alpha}$, where d_{i_k,i_j} is the distance between the i_k user and the i_j BS; α is the path-loss exponent; and ψ_{i_k,i_j} is a log-normal random variable, i.e., the quantity $10\log_{10}(\psi_{i_k,i_j})$ is distributed zero-mean Gaussian with a standard deviation of $\sigma_{\text{shadowing}}$. The small-scale fading coefficients, i.e., elements of $\mathbf{H}_{i_k,i_j}$, are modeled as i.i.d. complex Gaussian variables with zero-mean and unit-variance. We further assume a block fading model, where small-scale channels are constant over a few time slots with respect to channel estimation and channel state information feedback procedures. Similarly, we assume that large-scale fading coefficients β_{i_k,i_j} stay constant during large-scale coherence blocks. The small-scale and large-scale fading coefficients in different coherence blocks are assumed to be independent.

All BSs are connected to the cloud through high-capacity backhaul links as in [74]. The cloud has access to the whole data library containing F files, where different contents are independent. We define $\Pi_i = \{\pi_{i_1}, \ldots, \pi_{i_{|\mathscr{I}_i|}}\}$ as the user request profile at the i-th cluster, where π_{i_k} denotes the index of the requested file by the i_k user. Users can make random requests from a directory of files $\mathscr{F} = \{f_1, f_2, \ldots, f_F\}$ where each file f_n has size ℓ_{f_n} bits. For the sake of simplicity, we assume that the cache size at any BS is at least large enough to cache any of the files, i.e., $\ell_{f_n} \leq s_{i_j}$ for all $f_n \in \mathscr{F}$, $i \in \mathscr{M}$, $j \in \mathscr{Q}_i$. Moreover, we assume that the i_k user makes q_{i_k} number of requests over a given time interval T. Therefore, $\mathbf{q}_i = [q_{i_1}, \ldots, q_{i_{|\mathscr{I}_i|}}]$ indicates the rates of requests that are made

by users in the i-th cluster. We also assume different users in the i-th cluster, $i \in \mathscr{M}$, may have different file preferences. Assuming $\bar{p}_{i_k,f_n}$ is the probability that i_k user, $i \in \mathscr{M}$, $k \in \mathscr{I}_i$, request file $f_n \in \mathscr{F}$, the discrete popularity distribution of files for the UEs in the i-th cluster can be indicated as

$$\overline{\mathbf{P}}_i = \begin{bmatrix} \overline{\mathbf{p}}_{i_1} \\ \vdots \\ \overline{\mathbf{p}}_{i_k} \\ \vdots \\ \overline{\mathbf{p}}_{i_{|\mathscr{I}_i|}} \end{bmatrix} = \begin{bmatrix} \bar{p}_{i_1,f_1} & \cdots & \bar{p}_{i_1,f_F} \\ \vdots & \ddots & \vdots \\ \bar{p}_{i_k,f_1} & \cdots & \bar{p}_{i_k,f_F} \\ \vdots & \ddots & \vdots \\ \bar{p}_{i_{|\mathscr{I}_i|},f_1} & \cdots & \bar{p}_{i_{|\mathscr{I}_i|},f_F} \end{bmatrix} \in [0,1]^{|\mathscr{I}_i| \times F}$$

where $\bar{p}_{m,n}$ represents the probability that the m-th UE in the i-th cluster requests the n-th file. It is worth noting that, the m-th row of matrix $\overline{\mathbf{P}}_i$ is a stochastic vector which indicates the discrete probability distribution of the m-th user.

Since the file popularity distributions seen at each BS depends on the local file popularities of all connected UEs to the BS [108], this matrix will be different from $\overline{\mathbf{P}}_i$. The popularity distributions at the BSs in the i-th cluster, namely $\mathbf{P}_i \in [0,1]^{|\mathscr{Q}_i| \times F}$, can be derived as

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{p}_{i_1} \\ \vdots \\ \mathbf{p}_{i_{|\mathscr{Q}_i|}} \end{bmatrix} = \begin{bmatrix} p_{i_1,f_1} & \cdots & p_{i_1,f_F} \\ p_{i_2,f_1} & \cdots & p_{i_2,f_F} \\ \vdots & \ddots & \vdots \\ p_{i_{|\mathscr{Q}_i|},f_1} & \cdots & p_{i_{|\mathscr{Q}_i|},f_F} \end{bmatrix} = \begin{bmatrix} \frac{1}{\mathbf{d}_{i_1}\mathbf{q}_i^T} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mathbf{d}_{i_2}\mathbf{q}_i^T} & \cdots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & \frac{1}{\mathbf{d}_{i_{|\mathscr{Q}_i|}}\mathbf{q}_i^T} \end{bmatrix} \mathbf{D}_i \begin{bmatrix} q_{i_1} & 0 & \cdots & 0 \\ 0 & q_{i_2} & \cdots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & q_{i_{|\mathscr{I}_i|}} \end{bmatrix} \overline{\mathbf{P}}_i$$

where

$$p_{i_j,f_n} = \frac{\sum_{k=1}^{|\mathscr{I}_i|} d_{i_j,i_k} q_{i_k} \bar{p}_{i_k,f_n}}{\sum_{k=1}^{|\mathscr{I}_i|} d_{i_j,i_k} q_{i_k}}$$

denotes the n-th file popularity distribution observed at the j-th BS in the i-th cluster and the

denominator is a normalization factor. In practice, by adding up the number of times that the f_n file is requested by users, p_{i_j,f_n} can be computed at the cloud. Moreover, since the user behavior is correlated with the previously requested data, p_{i_j,f_n} can provide the information regarding the file popularity of the future requests and so, it helps to efficiently store the files in caches before a request is made. $\mathbf{D}_i$ denotes the user association matrix in the i-th cluster which depicts the connection between the BSs and UEs in the i-th cluster. The user association matrix is structured as

$$\mathbf{D}_i = \begin{bmatrix} \mathbf{d}_{i_1} \\ \vdots \\ \mathbf{d}_{i_{|\mathscr{Q}_i|}} \end{bmatrix} = \begin{bmatrix} d_{i_1,i_1} & \dots & d_{i_1,i_k} & \dots & d_{i_1,i_{|\mathscr{I}_i|}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i_{|\mathscr{Q}_i|},i_1} & \dots & d_{i_{|\mathscr{Q}_i|},i_k} & \dots & d_{i_{|\mathscr{Q}_i|},i_{|\mathscr{I}_i|}} \end{bmatrix} \in \{0,1\}^{|\mathscr{Q}_i|\times|\mathscr{I}_i|}$$

where $d_{m,n} = \mathbb{1}(\tilde{r}_{m,n} \geq r)$. $\tilde{r}_{m,n}$ is the mn-th entry of the wireless downlink rate matrix $\tilde{\mathbf{R}}_i \in \mathbb{R}^{|\mathscr{Q}_i|\times|\mathscr{I}_i|}$ in the i-th cluster and represents the achievable data rate from BS m to user n while r guarantees a certain quality of service so that the n-th user would not connect to the m-th BS if the wireless link rate between them was below the threshold r.

Since the caching capacity is limited, the aim of designing a cache placement strategy is to store the most popular contents such that the BSs can directly serve the majority of UEs' demands. We define the content placement matrix at the i-th cluster as $\mathbf{C}_i \in \{0,1\}^{F\times|\mathscr{Q}_i|}$, where $\mathbf{C}_i(f_n,i_j) = 1$ means the f_n-th content is stored in the i_j BS and $\mathbf{C}_i(f_n,i_j) = 0$ represents the opposite. The caching matrix will be different for different user in one cluster. To be specific, the caching matrix corresponding to the BSs that are associated with UE i_k, $i \in \mathscr{M}$ and $k \in \mathscr{I}_i$, can be expressed as

$$\mathbf{C}_{i_k} = \mathbf{C}_i \blacksquare(\mathbf{D}_i(:,i_k)); \tag{4.1}$$

where $\blacksquare(\mathbf{x})$ is a diagonal matrix formed from vector $\mathbf{x}$ and $\mathbf{A}(:,n)$ indicates the n-th column of the matrix $\mathbf{A}$. Therefore, the sum of all elements in the n-th row of $\mathbf{C}_{i_k}$ represents the number of BSs whose serve the i_k user and cache the f_n-th file. Considering that the cache at i_j BS can only store

s_{i_j} bits of data[1], the following cache size constraint should be fulfilled at BS i_j

$$\sum_{n=1}^{F} \mathbf{C}_i(f_n, i_j)\ell_{f_n} \leq s_{i_j}, \quad \forall i \in \mathscr{M},\ j \in \mathscr{Q}_i. \tag{4.2}$$

Moreover, the i_k user is able to download the f_n-th file from the cache if the following condition is satisfied:

$$\sum_{j=1}^{|\mathscr{Q}_i|} d_{i_j,i_k}\mathbf{C}_i(f_n, i_j) > 0;\ \ \forall i \in \mathscr{M}, k \in \mathscr{I}_i, f_n \in \mathscr{F}, \tag{4.3}$$

otherwise, corresponding links may have backhaul cost to transfer the f_n-th file. As a result, by associating each user with the BSs that caches its requested content, the total backhaul reduction of the aforementioned cache-enabled Cloud-RAN can be represented as

$$\mathbb{E}\{\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{j=1}^{|\mathscr{Q}_i|} d_{i_j,i_k}\mathbf{C}_i(f_n, i_j)\} = \sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{n=1}^{F}\sum_{j=1}^{|\mathscr{Q}_i|} d_{i_j,i_k}\bar{p}_{i_k,f_n}\mathbf{C}_i(f_n, i_j) = \sum_{i=1}^{M}\mathrm{Tr}(\overline{\mathbf{P}}_i\mathbf{C}_i\mathbf{D}_i).$$

where expectation is with respect to the random user requests.

Furthermore, we assume that in the underlying Cloud-RAN, a cluster of cooperative BSs serves each UE. To be specific, a cooperating set $\mathscr{B}_{i_k}$ is assigned to i_k user where $\mathscr{B}_{i_k} = \{i_j | d_{i_j,i_k} = 1\} \subseteq \mathscr{Q}_i$ is formed by aggregating all the BSs that have knowledge of channels $\mathbf{H}_{i_k,i_j},\ i_j \in \mathscr{B}_{i_k}$ and, have access to i_k user's message. They may also jointly encode the message intended for this user in their transmission [109]. Note that, due to the shadowing effect, the $|\mathscr{B}_{i_k}|$ strongest BSs under coordination are not necessarily the $|\mathscr{B}_{i_k}|$ nearest BSs, where $|\mathscr{B}_{i_k}| = \|\mathbf{D}_i(:, i_k)\|_0$ denote the set's cardinality. Since each BS may get involved in transmission to more than one UE, the cooperating set of different users may overlap. We designate the set of users served by the i_j BS by $\mathscr{K}_{i_j} \triangleq \{i_k | i_j \in \mathscr{B}_{i_k}\} \subseteq \mathscr{I}_i$. Moreover, similar to [110], we assume each co-scheduled user operating in the MU-MIMO mode only receives one spatial stream (rank 1 transmission) as specified by the Rel-10 LTE-Advanced standard [27, Chapter 11]. BS i_j plans to communicate a symbol vector

[1]We assume that a file is either completely cached or not cached at all in a BS. Of course, partial/coded caching techniques can be envisaged, and these are left for future research.

$\mathbf{s}_{i_j} = [s_{i_1,i_j}, \ldots, s_{i_{|\mathscr{K}_{i_j}|},i_j}]^T \in \mathbb{C}^{|\mathscr{K}_{i_j}|}$ to its associated receivers, where s_{i_k,i_j} is the transmit symbol from the i_j BS to the $i_k \in \mathscr{K}_{i_j}$ receiver with unit power of $\mathbb{E}\{|s_{i_k,i_j}|^2\} = 1$ and $|\mathscr{K}_{i_j}|$ denote the set's cardinality.

Remark 6. *The main idea behind defining M and $\mathscr{B}_{i_k}$ is that the demand for content shows variations both across time and space and, in this chapter, we considered both temporal and spatial locality of the content demand. To be specific, it is crucial to consider the user location since the demand may differentiate from one geographical area to another. This space variation of the content request, referred to spatial locality, is why we define M clusters, which M can be any arbitrary number. On the other hand, the demand may differentiate from time to time. This temporal variation of the content request, referred to time locality, is the reason that we defined the $\mathscr{B}_{i_k}$ cluster associated with user i_k.*

Prior to transmitting, the i_j BS linearly precodes its symbol vector $\mathbf{x}_{i_j} = \sum_{i_k \in \mathscr{K}_{i_j}} \sqrt{\tilde{p}_{i_k,i_j}}\mathbf{v}_{i_k,i_j}s_{i_k,i_j}$ where $\tilde{p}_{i_k,i_j}$ stands for the transmit power allocated to user i_k from the i_j BS and $\mathbf{v}_{i_k,i_j}$ denotes the unit-norm beamformer that BS i_j uses to transmit the signal s_{i_k,i_j} to receiver i_k. Each BS i_j is under a transmit power constraint of $P_{i_j}^{\max}$ and so, the transmit power at the i_j BS is computed as $\tilde{P}_{i_j} = \mathbb{E}\{\|\mathbf{x}_{i_j}\|^2\} = \sum_{i_k \in \mathscr{K}_{i_j}} \tilde{p}_{i_k,i_j}\mathbf{v}_{i_k,i_j}^H\mathbf{v}_{i_k,i_j} = \sum_{i_k \in \mathscr{K}_{i_j}} \tilde{p}_{i_k,i_j} \leq P_{i_j}^{\max}$ [16, 110]. Under our assumptions, when user i_k request for file f_n which is available at the cache of the BSs in $\mathscr{B}_{i_k}$, the received signals from these BSs are combined coherently using the coordinated joint transmission. Since there are $|\mathscr{B}_{i_k}|$ BSs participating in the cooperating data transmission to user i_k, at the same frequency and time, we denote $s_{i_k} \in \mathbb{C}$ as the complex data symbol for the i_k UE, where $i \in \mathscr{M}$ and $k \in \mathscr{I}_i$. Consequently, throughout each symbol duration time, the cooperating BSs transmit the same symbol s_{i_k} and the received signal vector $\mathbf{y}_{i_k,f_n} \in \mathbb{C}^{N_r^{i_k} \times 1}$ at the k-th user in the i-th cluster, when user i_k requested file f_n, can be written as

$$\mathbf{y}_{i_k,f_n} = \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_k,f_n}\mathbf{C}_{i_k}(f_n,i_j)\sqrt{\tilde{p}_{i_k,i_j}}\sqrt{\beta_{i_k,i_j}}\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_k,i_j}s_{i_k} + \sum_{n=1}^{F}\sum_{\substack{m=1 \\ m\neq k}}^{|\mathscr{I}_i|}\sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_m,f_n}\mathbf{C}_{i_m}(f_n,i_j)\sqrt{\tilde{p}_{i_m,i_j}}\sqrt{\beta_{i_k,i_j}}$$
$$\times \mathbf{H}_{i_k,i_j}\mathbf{v}_{i_m,i_j}s_{i_m} + \sum_{n=1}^{F}\sum_{\substack{q=1 \\ q\neq i}}^{M}\sum_{\ell=1}^{|\mathscr{I}_q|}\sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}_{q_\ell,f_n}\mathbf{C}_{q_\ell}(f_n,q_j)\sqrt{\tilde{p}_{q_\ell,q_j}}\sqrt{\beta_{i_k,q_j}}\mathbf{H}_{i_k,q_j}\mathbf{v}_{q_\ell,q_j}s_{q_\ell} + \mathbf{n}_{i_k}, \tag{4.4}$$

where the first term on the right-hand side of (4.4) represents the received useful signal, the second and third terms represent the intra-cluster and inter-cluster interference respectively, and $\mathbf{n}_{i_k} \sim \mathscr{CN}(0, \sigma_{i_k}^2 \mathbf{I})$ is the additive white Gaussian noise (AWGN) at the i_k UE. We assumed that the signals for different users are independent from each other. In this chapter, we treat interference as noise and consider a linear receive beamforming strategy so that the estimated signal is given by $\hat{s}_{i_k} = \mathbf{u}_{i_k}^H \mathbf{y}_{i_k,f_n}$. Indeed, each receiver $i_k \in \mathscr{K}_{i_j}$, linearly processes the received signal to obtain $\mathbf{u}_{i_k}^H \mathbf{y}_{i_k,f_n}$ where $\mathbf{u}_{i_k} \in \mathbb{C}^{N_r^{i_k}}$ denotes the unit-norm post processing filter at receiver i_k, i.e., $\|\mathbf{u}_{i_k}\|^2 = 1$. Thus, after receive beamforming the received signal at receiver $i_k \in \mathscr{K}_{i_j}$, can be expressed as

$$\hat{s}_{i_k,f_n} = \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_k,f_n} \mathbf{C}_{i_k}(f_n, i_j) \sqrt{\tilde{p}_{i_k,i_j}} \sqrt{\beta_{i_k,i_j}} \mathbf{u}_{i_k}^H \mathbf{H}_{i_k,i_j} \mathbf{v}_{i_k,i_j} s_{i_k} + \sum_{n=1}^{F} \sum_{\substack{m=1 \\ m \neq k}}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_m,f_n} \mathbf{C}_{i_m}(f_n, i_j) \sqrt{\tilde{p}_{i_m,i_j}} \sqrt{\beta_{i_k,i_j}} \times$$
$$\mathbf{u}_{i_k}^H \mathbf{H}_{i_k,i_j} \mathbf{v}_{i_m,i_j} s_{i_m} + \sum_{n=1}^{F} \sum_{\substack{q=1 \\ q \neq i}}^{M} \sum_{\ell=1}^{|\mathscr{I}_q|} \sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}_{q_\ell,f_n} \mathbf{C}_{q_\ell}(f_n, q_j) \sqrt{\tilde{p}_{q_\ell,q_j}} \sqrt{\beta_{i_k,q_j}} \mathbf{u}_{i_k}^H \mathbf{H}_{i_k,q_j} \mathbf{v}_{q_\ell,q_j} s_{q_\ell} + \mathbf{u}_{i_k}^H \mathbf{n}_{i_k}. \tag{4.5}$$

As mentioned above, the received interference at the i_k-th UE is the summation of the intra-cluster and inter-cluster interferences. While the former is the interference experienced by the i_k UE from all BSs inside the i-th cluster, i.e.,

$$I_{i_k,f_n}^{\text{intra}} = \sum_{n=1}^{F} \sum_{\substack{m=1 \\ m \neq k}}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_m,f_n}^2 \mathbf{C}_{i_m}(f_n, i_j) \tilde{p}_{i_m,i_j} \beta_{i_k,i_j} \text{Tr}\left(\mathbf{v}_{i_m,i_j}^H \mathbf{H}_{i_k,i_j}^H \mathbf{u}_{i_k} \mathbf{u}_{i_k}^H \mathbf{H}_{i_k,i_j} \mathbf{v}_{i_m,i_j} \right),$$

the latter is the received interference from all the BSs outside the i-th cluster and can be represented as

$$I_{i_k,f_n}^{\text{inter}} = \sum_{n=1}^{F} \sum_{\substack{q=1 \\ q \neq i}}^{M} \sum_{\ell=1}^{|\mathscr{I}_q|} \sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}_{q_\ell,f_n}^2 \mathbf{C}_{q_\ell}(f_n, q_j) \tilde{p}_{q_\ell,q_j} \beta_{i_k,q_j} \text{Tr}\left(\mathbf{v}_{q_\ell,q_j}^H \mathbf{H}_{i_k,q_j}^H \mathbf{u}_{i_k} \mathbf{u}_{i_k}^H \mathbf{H}_{i_k,q_j} \mathbf{v}_{q_\ell,q_j} \right).$$

Therefore, the SINR at the k-th user in the i-th cluster, when user i_k requests file f_n, can be written

as

$$\begin{aligned}\mathrm{SINR}_{i_k,f_n} &= \frac{\sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}^2_{i_k,f_n} \mathbf{C}_{i_k}(f_n,i_j)\tilde{p}_{i_k,i_j}\beta_{i_k,i_j}\mathrm{Tr}\left(\mathbf{v}^H_{i_k,i_j}\mathbf{H}^H_{i_k,i_j}\mathbf{u}_{i_k}\mathbf{u}^H_{i_k}\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_k,i_j}\right)}{I^{\text{intra}}_{i_k,f_n} + I^{\text{inter}}_{i_k,f_n} + \sigma^2_{i_k}}, \\ &= \frac{\sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}^2_{i_k,f_n} \mathbf{C}_{i_k}(f_n,i_j)\tilde{p}_{i_k,i_j}\beta_{i_k,i_j}|\mathbf{u}^H_{i_k}\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_k,i_j}|^2}{\mathbf{u}^H_{i_k}\left(\sum_{n=1}^{F}\sum_{(\ell,q)\neq(k,i)}\sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}^2_{q_\ell,f_n}\mathbf{C}_{q_\ell}(f_n,q_j)\tilde{p}_{q_\ell,q_j}\beta_{i_k,q_j}\mathbf{H}_{i_k,q_j}\mathbf{v}_{q_\ell,q_j}\mathbf{v}^H_{q_\ell,q_j}\mathbf{H}^H_{i_k,q_j} + \sigma_{i_k}\mathbf{I}\right)\mathbf{u}_{i_k}},\end{aligned} \tag{4.6}$$

where $\mathbf{C}_{q_\ell}(f_n,q_j) = \mathbf{C}_q(f_n,q_j)\mathbf{D}_q(q_j,q_\ell)$. Therefore the transmission rate for user i_k, when requests file f_n, can be written as $R_{i_k,f_n} = \log(1+\mathrm{SINR}_{i_k,f_n})$. The SINR at the i_k user, when requests file f_n, is a function of transmit power, cache placement matrix, user association matrix, probability of file being requested, and distance from all connected BSs to i_k user. Therefore, the average throughput of i_k user can be formulated as

$$R_{i_k} = \mathbb{E}\{R_{i_k,f_n}\} = \sum_{n=1}^{F} \bar{p}_{i_k,f_n} R_{i_k,f_n}, \quad \forall i \in \mathscr{M},\ k \in \mathscr{I}_i \tag{4.7}$$

where expectation is with respect to both the channel realizations and the random user requests. The distribution of the UEs' content demands follows a Zipf-like distribution $P_{\mathscr{F}}(f)$ given as

$$P_{\mathscr{F}}(f) = \frac{1}{f^\gamma \sum_{f=1}^{F} f^{-\gamma}}, \ \forall f \in \mathscr{F} \tag{4.8}$$

where γ models the skewness of the popularity profile [111, 112]. Depending on both BSs deployment strategies and users' behavior, γ can take different values. The popularity is uniformly distributed over content files for lower values of γ meaning that users have more distinct interests. As γ grows, the popularity becomes more skewed towards the most popular files which means users have very similar file interests and a small subset of files are more desired than the rest.

4.3 Problem Formulation and Analysis

In this section, the problem of interest is a joint optimization of the content placement and content delivery (precoding and user association). We formulate a mixed-timescale optimization problem which maximizes the trade-off between the backhaul saving and the network throughput. For maximizing the backhaul reduction, each UE should be associated with a BS (cluster of BSs) that caches the largest amount of its desired contents. However, such a BS might be long way away. On the other hand, in order to maximize the network throughput, each UE should take the BSs' load into account and accordingly associate with a BS which provides it a reasonably high SINR. Nonetheless, such a BS may not store the user's desired contents. Considering such a tradeoff into account, in this chapter the problem of maximizing the user throughput and backhaul saving with respect to the precoding matrix, cache placement matrix, and user association matrix is formulated subject to the peak transmission power and cache capacity constraints.

Let $\mathbf{v} = \left\{ \mathbf{v}_{i_k,i_j}(\Pi_i, \{\mathbf{H}_{i_k,i_j}\}_{j=1}^{|\mathscr{Q}_i|}) : \forall i,k,j \right\}$ and $\mathbf{D} = \left\{ \mathbf{D}_i(\Pi_i, \{\mathbf{H}_{i_k,i_j}\}_{j=1}^{|\mathscr{Q}_i|}) : \forall i,k,j \right\}$ denote all the beamforming vectors and user-association matrices for all user request profile $\{\Pi_i\}_{i=1}^{M}$ and instantaneous channel state information matrices $\{\mathbf{H}_{i_k,i_j}\}_{j=1}^{|\mathscr{Q}_i|}$, respectively. Then for given set of optimization variables $(\mathbf{v}, \mathbf{D}, \mathbf{C}_i)$ and user request profile $\{\Pi_i\}_{i=1}^{M}$, the average utility function of our optimization problem can be express as $\sum_{i=1}^{M} \sum_{k=1}^{|\mathscr{I}_i|} \mathbb{E}\left[\lambda R_{i_k,f_n} + (1-\lambda) \sum_{j=1}^{|\mathscr{Q}_i|} d_{i_j,i_k} \mathbf{C}_i(f_n, i_j) \middle| \Pi_i \right]$ where the parameter controlling the tradeoff between network throughput and backhaul saving is denoted by $0 < \lambda < 1$ so that by adjusting λ we can emphasize one term over the other. The optimization variables are partitioned into short-term (user association and beamforming) and long-term (content placement) variables. While the latter one is adaptive to the popularity of data and the channel statistics, the former one is adaptive to the instantaneous channel state information. The following feasible sets are defined for the cache placement $\mathbf{C}_i$, the user association $\mathbf{D}_i$, and the beamforming

vectors $\mathbf{v}_{i_k,i_j}$ as follows

$$\mathscr{S}_{\mathbf{C}} = \{\mathbf{C}_i :\ \mathbf{C}_i(f_n,i_j) \in \{0,1\}\ ,\ \sum_{n=1}^{F} \mathbf{C}_i(f_n,i_j)\ell_{f_n} \leq s_{i_j},\ \forall i \in \mathscr{M},\ j \in \mathscr{Q}_i\},$$

$$\mathscr{S}_{\mathbf{D}} = \{\mathbf{D}_i :\ \mathbf{D}_i(i_j,i_k) \in \{0,1\},\ \forall i \in \mathscr{M}, j \in \mathscr{Q}_i, k \in \mathscr{I}_i\},$$

$$\mathscr{S}_{\mathbf{v}} = \{\mathbf{v}_{i_k,i_j} :\ \sum_{k=1}^{|\mathscr{I}_i|} \mathrm{Tr}(\mathbf{v}_{i_k,i_j}^H \tilde{p}_{i_k,i_j} \mathbf{v}_{i_k,i_j}) \leq P_{i_j},\ \forall i \in \mathscr{M}, j \in \mathscr{Q}_i\}.$$

Then the joint content placement and content delivery problem is formulated as follows[2]:

$$\begin{aligned} \mathscr{P} : \underset{\mathbf{C}_i \in \mathscr{S}_{\mathbf{C}}, \mathbf{D}, \mathbf{v}}{\text{maximize}} \quad & \sum_{i=1}^{M} \sum_{k=1}^{|\mathscr{I}_i|} \mathbb{E}\left[\lambda R_{i_k,f_n} + (1-\lambda) \sum_{j=1}^{|\mathscr{Q}_i|} d_{i_j,i_k} \mathbf{C}_i(f_n,i_j) \middle| \Pi_i \right] \\ \text{subject to} \quad & \mathbf{v} \in \mathscr{S}_{\mathbf{v}}\ ,\ \mathbf{D} \in \mathscr{S}_{\mathbf{D}} \end{aligned} \tag{4.9}$$

where the objective function can be expressed in a more compact form of $\sum_{i=1}^{M} \mathrm{Tr}\Big(\lambda \mathbf{E}_i(\overline{\mathbf{P}}_i \odot \mathbf{R}_i) + (1-\lambda)\overline{\mathbf{P}}_i \mathbf{C}_i \mathbf{D}_i \Big)$ in which $\mathbf{E}_i$ is a $F \times |\mathscr{I}_i|$ matrix full of 1's and $\mathbf{R}_i$ is a $|\mathscr{I}_i| \times F$ rate matrix so that its $(i_k - f_n)$-th element is equal to R_{i_k,f_n}.

Due to the coupled interference among mobile stations, the optimization problem (4.9) is non-convex. Moreover, the entries of user association and cache placement matrices take binary values 0 and 1, thus the optimization problem falls into a mixed integer nonlinear programming (MINLP) which is usually NP-hard in general [113] and non-tractable in practice. Since, it is highly unlikely to compute a globally optimal solution in polynomial time, our goal is to obtain a trackable near-optimal solution by developing effective suboptimal algorithms. By utilizing the timescale separations of the optimization variables and making use of the fact that all constraints are separable, we divide the original optimization problem into three subproblems and propose an iterative algorithm that at each time maximizes the objective function with respect to one variable while assuming the rest two variables are given. Therefore, each of these subproblems can be relaxed to a convex problem so that it can be solved efficiently. The mixed-timescale joint optimization of content delivery and content placement can be decomposed to the following sub-problems:

[2]Note that we are considering a particular realization of the PPP, i.e., B is a sample of a Poisson random variable.

(a) Short-term Content Delivery (For given Π and $\mathbf{H}$)

As mentioned before, the content placement usually takes much longer than that of the content delivery. Therefore, in this subsection, it is assumed that the content placement matrix $\mathbf{C}_i$ is fixed and given. Therefore, we pay attention to the joint optimization of beamforming design and user-association. We further decouple the joint optimization problem in two stages. At the first stage, we associated each user with a cluster of BSs and the second stage, assuming the user-association is fixed and given, we design the beamformers. The proposed algorithm is described as follows:

$\diamond$ **User-association Stage:** Assuming caching policy and beamforming vectors are given, the user association problem can be simplified as

$$\begin{aligned} \mathscr{P}_1 : \underset{\mathbf{D}_i}{\text{maximize}} \quad & \sum_{i=1}^{M} \sum_{k=1}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} d_{i_j,i_k} \mu_{i_j,i_k} \\ \text{subject to} \quad & d_{i_j,i_k}(1-d_{i_j,i_k}) = 0, \forall i \in \mathscr{M}, j \in \mathscr{Q}_i, k \in \mathscr{I}_i \end{aligned} \tag{4.10}$$

where $\mu_{i_j,i_k} = \sum_{n=1}^{F} \bar{p}_{i_k,f_n} \mathbf{C}_i(f_n, i_j)$ represents the amount of backhaul saving by associating i_k-th UE with the i_j-th BS. Due to the fact that the entries of $\mathbf{D}_i$ take binary values 0 and 1, the optimization problem is a mixed integer optimization over the user association. In order to solve this optimization problem and inspired by the idea used in [98], this chapter adopts a new approach called dual analysis. The main idea behind this method is to answer how the optimal value can be deduced from the constraints. This method is used in [98] and [106] to find a solution to the user association problem in heterogeneous cellular networks under the proportional fairness criterion. Using this method, the optimization problem can be easily decoupled among the clusters and the solution can be expressed as:

$$d^*_{i_j,i_k} = \begin{cases} 1, & \text{if } \mu_{i_j,i_k} > 0 \\ 0, & \text{Otherwise} \end{cases} \tag{4.11}$$

which shows that tacking caching placement and so backhaul reduction into account can be viewed

as an additional incentive for UEs to associate with a BS.

◇ **Beamforming Stage**: With fixed content placement and user association, the problem of designing beamforming vectors $\mathbf{v}_{i_k}$ and $\mathbf{u}_{i_k}$ can be written as

$$\begin{aligned} \mathscr{P}_2 : \underset{\substack{\mathbf{v}_{i_k,i_j} \\ \mathbf{u}_{i_k,i_j}}}{\text{maximize}} \quad & \sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{n=1}^{F} \bar{p}_{i_k,f_n}\log(1+\text{SINR}_{i_k,f_n}) \\ \text{subject to} \quad & \sum_{k=1}^{|\mathscr{I}_i|} \tilde{p}_{i_k,i_j} \parallel \mathbf{v}_{i_k,i_j} \parallel^2 \leq P_{i_j}, \forall i \in \mathscr{M}, j \in \mathscr{Q}_i \end{aligned} \tag{4.12}$$

In order to suit our system model we applied the weighted minimum mean-square error (WMMSE) framework [114] into the above optimization problem and modified the Proposition 3.2. in [115]. This way, each user can connect to a cluster of BSs instead of serving by only one BS. Thereby, problem (4.12) can be effectively rewritten as

$$\begin{aligned} \underset{\substack{\mathbf{v}_{i_k,i_j} \\ \mathbf{u}_{i_k,i_j} \\ w_{i_k}}}{\text{maximize}} \quad & \sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{n=1}^{F} \bar{p}_{i_k,f_n}\Big(\log(w_{i_k}) - w_{i_k}\varepsilon_{i_k} + 1\Big) \\ \text{subject to} \quad & \sum_{k=1}^{|\mathscr{I}_i|} \tilde{p}_{i_k,i_j} \parallel \mathbf{v}_{i_k,i_j} \parallel^2 \leq P_{i_j}, \forall i \in \mathscr{M}, j \in \mathscr{Q}_i \end{aligned} \tag{4.13}$$

where $\{w_{i_k}\}$ are the weights variable introduced by WMMSE framework and $\{\varepsilon_{i_k}\}$ are the mean square estimation errors which are defined by

$$\begin{aligned} \varepsilon_{i_k} \triangleq\ & |1 - \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}^2_{i_k,f_n}\mathbf{C}_{i_k}(f_n,i_j)\tilde{p}_{i_k,i_j}\beta_{i_k,i_j}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_k,i_j}|^2 \\ & + \sum_{n=1}^{F}\sum_{(\ell,q)\neq(k,i)}\sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}^2_{q_\ell,f_n}\mathbf{C}_{q_\ell}(f_n,q_j)\tilde{p}_{q_\ell,q_j}\beta_{i_k,q_j}|\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,q_j}\mathbf{v}_{q_\ell,q_j}|^2 + \sigma_{i_k}\|\mathbf{u}_{i_k}\|^2 \end{aligned}$$

The objective function of (4.13) is convex with respect to each of the optimization variables $\mathbf{v}_{i_k,i_j}$, $\mathbf{u}_{i_k,i_j}$, and w_{i_k}, which enables us to employ the block coordinate descent method to solve it [114]. To be specific, we maximize the cost function of (4.13) by updating one of three variables $\mathbf{v}_{i_k,i_j}$, $\mathbf{u}_{i_k,i_j}$, and w_{i_k}, while assuming the rest are given. In particular, we iteratively run the following steps.

• Initializing all the transmit beamformers $\mathbf{v}_{i_k,i_j}$'s, $\forall i,k,j$, and minimizing the weighted sum-MSE

leads us to the MMSE receiver $\mathbf{u}_{i_k}$ as follows

$$\mathbf{u}_{i_k}^{\text{mmse}} = \mathbf{J}_{i_k}^{-1}\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_k,i_j} \tag{4.14}$$

where

$$\mathbf{J}_{i_k} = \sum_{n=1}^{F}\sum_{q=1}^{M}\sum_{\ell=1}^{|\mathscr{I}_q|}\sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}_{q_\ell,f_n}^2 \mathbf{C}_{q_\ell}(f_n,q_j)\tilde{p}_{q_\ell,q_j}\beta_{i_k,q_j}\mathbf{H}_{i_k,q_j}\mathbf{v}_{q_\ell,q_j}\mathbf{v}_{q_\ell,q_j}^H\mathbf{H}_{i_k,q_j}^H + \sigma_{i_k}^2\mathbf{I}$$

is the covariance matrix of the total received signal at the i_k receiver.

- By fixing all $\mathbf{u}_{i_k}$'s and $\mathbf{v}_{i_k,i_j}$'s, $\forall i,j,k$, the weights, for all i and k, can be updated as follows

$$w_{i_k} = (1 - \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_k,f_n}^2 \mathbf{C}_{i_k}(f_n,i_j)\tilde{p}_{i_k,i_j}\beta_{i_k,i_j}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_k,i_j})^{-1}, \tag{4.15}$$

- By fixing all w_{i_k}'s and $\mathbf{u}_{i_k}$'s, the transmit beamformers can be calculated using the following optimization problem

$$\begin{aligned} \underset{\mathbf{v}_{i_k,i_j}}{\text{minimize}} \quad & \sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{n=1}^{F} \bar{p}_{i_k,f_n} w_{i_k}\varepsilon_{i_k} \\ \text{subject to} \quad & \sum_{k=1}^{|\mathscr{I}_i|} \tilde{p}_{i_k,i_j} \parallel \mathbf{v}_{i_k,i_j} \parallel^2 \leq P_{i_j}, \end{aligned} \tag{4.16}$$

Problem (4.16) is convex and (4.14) and (4.15) can be locally implemented at the users. Therefore, we solve the problem (4.16) in a distritbuted manner based on the alternating direction method of multipliers (ADMM) [116]. In what follows, it is shown that by exchanging a fair amount of information between UEs and BSs the ADMM can be applied in a distributed fashion to solve optimization problem (4.16). In order to achieve distributed implementation of the ADMM in the aforementioned cache-enabled C-RAN network, the following assumptions are made (similar to [114] and [115]). We assume that each BS $j \in Q_i$ knows $\mathbf{H}_{i_k,i_j}$ for all i_k user in cluster $\mathscr{B}_{i_k}$ and each user i_k can estimate the interference plus noise covariance matrix. Under these assumptions, the ADMM can be applied distributively. Note that, In order to identify the beamforming vectors in a distributed fashion, our aim here is to have a same form as the one in [116]. To do so, we

introduce auxiliary variables $\{\mathbf{x}_{i_k,i_j}\}$ and $\{X_{i_k}^{i_k,i_j}\}$ and rewrite problem (4.16) as

$$\begin{aligned}
\underset{\substack{\mathbf{v}_{i_k,i_j}\\ \mathbf{x}_{i_k,i_j}\\ X_{i_k}^{i_k,i_j}}}{\text{minimize}} \quad & \sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{n=1}^{F}\bar{p}_{i_k,f_n}\Bigg(|\sqrt{w_{i_k}}-\sum_{j=1}^{|\mathscr{Q}_i|}\bar{p}_{i_k,f_n}^2\mathbf{C}_{i_k}(f_n,i_j)\tilde{p}_{i_k,i_j}\beta_{i_k,i_j}X_{i_k}^{i_k,i_j}|^2 \\
& \qquad +\sum_{n=1}^{F}\sum_{(\ell,q)\neq(k,i)}\sum_{j=1}^{|\mathscr{Q}_q|}\bar{p}_{q_\ell,f_n}^2\mathbf{C}_{q_\ell}(f_n,q_j)\tilde{p}_{q_\ell,q_j}\beta_{i_k,q_j}|X_{i_k}^{q_\ell,q_j}|^2\Bigg) \\
\text{subject to} \quad & \sum_{k=1}^{|\mathscr{I}_i|}\tilde{p}_{i_k,i_j}\parallel\mathbf{x}_{i_k,i_j}\parallel^2\le P_{i_j}, \qquad \forall i\in\mathscr{M}, j\in\mathscr{Q}_i, \\
& \qquad \mathbf{v}_{i_k,i_j}=\mathbf{x}_{i_k,i_j}, \qquad \forall i\in\mathscr{M}, k\in\mathscr{I}_i, j\in\mathscr{Q}_i, \\
& \sqrt{w_{i_k}}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_m,i_j}=X_{i_k}^{i_m,i_j}, \quad \forall i\in\mathscr{M},\ m\in\mathscr{I}_i,\ m\neq k,\ j\in\mathscr{Q}_i, \\
& \sqrt{w_{i_k}}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,q_m}\mathbf{v}_{q_\ell,q_m}=X_{i_k}^{q_\ell,q_m}, \quad \forall q\in\mathscr{M},\ q\neq i,\ \ell\in\mathscr{I}_q,\ m\in\mathscr{Q}_q
\end{aligned} \tag{4.17}$$

Then, we form the augmented Lagrangian as follows

$$\begin{aligned}
\mathscr{L}_\rho(\mathbf{v},\mathbf{x},\mathbf{X};\lambda_{i_k},\mathbf{z}_{i_k}) = & \sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{n=1}^{F}\bar{p}_{i_k,f_n}\Bigg(|\sqrt{w_{i_k}}-\sum_{j=1}^{|\mathscr{Q}_i|}\bar{p}_{i_k,f_n}^2\mathbf{C}_{i_k}(f_n,i_j)\tilde{p}_{i_k,i_j}\beta_{i_k,i_j}X_{i_k}^{i_k,i_j}|^2 \\
& +\sum_{n=1}^{F}\sum_{(\ell,q)\neq(k,i)}\sum_{j=1}^{|\mathscr{Q}_q|}\bar{p}_{q_\ell,f_n}^2\mathbf{C}_{q_\ell}(f_n,q_j)\tilde{p}_{q_\ell,q_j}\beta_{i_k,q_j}|X_{i_k}^{q_\ell,q_j}|^2\Bigg) \\
& +\mathrm{Re}\Bigg(\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{\substack{m=1\\ m\neq k}}^{|\mathscr{I}_i|}\sum_{j=1}^{|\mathscr{Q}_i|}\langle\sqrt{w_{i_k}}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_m,i_j}-X_{i_k}^{i_m,i_j},\lambda_{i_k}^{i_m,i_j}\rangle\Bigg) \\
& +\frac{\rho}{2}\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{\substack{m=1\\ m\neq k}}^{|\mathscr{I}_i|}\sum_{j=1}^{|\mathscr{Q}_i|}|\sqrt{w_{i_k}}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,i_j}\mathbf{v}_{i_m,i_j}-X_{i_k}^{i_m,i_j}|^2 \\
& +\mathrm{Re}\Bigg(\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{\substack{q=1\\ q\neq i}}^{M}\sum_{\ell=1}^{|\mathscr{I}_q|}\sum_{m=1}^{|\mathscr{Q}_q|}\langle\sqrt{w_{i_k}}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,q_m}\mathbf{v}_{q_\ell,q_m}-X_{i_k}^{q_\ell,q_m},\lambda_{i_k}^{q_\ell,q_m}\rangle\Bigg) \\
& +\frac{\rho}{2}\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{\substack{q=1\\ q\neq i}}^{M}\sum_{\ell=1}^{|\mathscr{I}_q|}\sum_{m=1}^{|\mathscr{Q}_q|}|\sqrt{w_{i_k}}\mathbf{u}_{i_k}^H\mathbf{H}_{i_k,q_m}\mathbf{v}_{q_\ell,q_m}-X_{i_k}^{q_\ell,q_m}|^2 \\
& +\mathrm{Re}\Bigg(\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{j=1}^{|\mathscr{Q}_i|}\langle\mathbf{v}_{i_k,i_j}-\mathbf{x}_{i_k,i_j},\mathbf{z}_{i_k,i_j}\rangle\Bigg)+\frac{\rho}{2}\sum_{i=1}^{M}\sum_{k=1}^{|\mathscr{I}_i|}\sum_{j=1}^{|\mathscr{Q}_i|}\|\mathbf{v}_{i_k,i_j}-\mathbf{x}_{i_k,i_j}\|^2
\end{aligned} \tag{4.18}$$

where ρ is the penalty parameter, and $\lambda_{i_k}\triangleq\{\lambda_{i_k}^{i_m,i_j}|m,k\in\mathscr{I}_i,\ j\in\mathscr{Q}_i\}$ and $\mathbf{z}_{i_k}\triangleq\{\mathbf{z}_{i_k,i_j}\in\mathbb{C}^{N_t^{i_j}}|k\in\mathscr{I}_i,\ j\in\mathscr{Q}_i\}$ are the scaled dual variables for the last three sets of equality constraints.

The ADMM approach consists of three steps. First, minimizing the augmented Lagrangian (4.18) over the decision variables $\mathbf{v}$, while assuming all the other variables are given at their current values. Second, minimizing the augmented Lagrangian (4.18) over the decision variables $\{\mathbf{x},\mathbf{X}\}$, assuming the rest of variables are given and fixed. While the latter is a constrained optimization problem the former is an unconstrained one. The last step consists of a simple dual update. Therefore, assuming $\mathbf{x}$ and $\mathbf{X}$ are given, the optimization problem with respect to $\mathbf{v}$ can be expressed as

$$\underset{\mathbf{v}}{\text{minimize}} \quad \mathscr{L}_\rho(\mathbf{v},\mathbf{x},\mathbf{X};\boldsymbol{\lambda}_{i_k},\mathbf{z}_{i_k}), \tag{4.19}$$

which can be decomposed into $i=1,\ldots,M,\ k=1,\ldots,|\mathscr{I}_i|,\ j=1,\ldots,|\mathscr{Q}_i|$

$$\underset{\mathbf{v}_{i_k,i_j}}{\text{minimize}} \ f(\mathbf{v}_{i_k,i_j}) \tag{4.20}$$

where $f(\mathbf{v}_{i_k,i_j})$ is defined as

$$\begin{aligned} f(\mathbf{v}_{i_k,i_j}) \triangleq \mathrm{Re}\Bigg(\sum_{\substack{m=1\\m\neq k}}^{|\mathscr{I}_i|} \lambda_{i_m}^{i_k,i_j}\sqrt{w_{i_m}}\mathbf{u}_{i_m}^H\mathbf{H}_{i_m,i_j}\mathbf{v}_{i_k,i_j}\Bigg) + \frac{\rho}{2}\sum_{\substack{m=1\\m\neq k}}^{|\mathscr{I}_i|}|\sqrt{w_{i_k}}\mathbf{u}_{i_m}^H\mathbf{H}_{i_m,i_j}\mathbf{v}_{i_k,i_j} - X_{i_m}^{i_k,i_j}|^2 \\ + \mathrm{Re}\Bigg(\mathbf{z}_{i_k,i_j}^H(\mathbf{v}_{i_k,i_j}-\mathbf{x}_{i_k,i_j})\Bigg) + \frac{\rho}{2}\|\mathbf{v}_{i_k,i_j}-\mathbf{x}_{i_k,i_j}\|^2, \end{aligned}$$

Assuming $\mathbf{v}$ is given, the constrained optimization problem with respect to $\{\mathbf{x},\mathbf{X}\}$ can be expressed as

$$\begin{aligned} &\underset{\mathbf{x},\mathbf{X}}{\text{minimize}} && \mathscr{L}_\rho(\mathbf{v},\mathbf{x},\mathbf{X};\boldsymbol{\lambda}_{i_k},\mathbf{z}_{i_k}), \\ &\text{subject to} && \sum_{k=1}^{|\mathscr{I}_i|}\tilde{p}_{i_k,i_j}\,\|\,\mathbf{x}_{i_k,i_j}\,\|^2 \leq P_{i_j}, \quad \forall i\in\mathscr{M}, j\in\mathscr{Q}_i \end{aligned} \tag{4.21}$$

which can be decomposed into $i=1,\ldots,M,\ k=1,\ldots,|\mathscr{I}_i|$

$$\underset{X_{i_k}^{i_k,i_j}}{\text{minimize}} \ g(X_{i_k}^{i_k,i_j}) \tag{4.22}$$

and $i = 1, \ldots, M$

$$\begin{aligned} &\underset{\mathbf{x}_{i_k,i_j}}{\text{minimize}} \;\; h(\mathbf{x}_{i_k,i_j}) \\ &\text{subject to} \;\; \sum_{k=1}^{|\mathscr{I}_i|} \tilde{p}_{i_k,i_j} \parallel \mathbf{x}_{i_k,i_j} \parallel^2 \leq P_{i_j}, \end{aligned} \tag{4.23}$$

where $g(X_{i_k}^{i_k,i_j})$ and $h(\mathbf{x}_{i_k,i_j})$ are defined as follows

$$\begin{aligned} g(X_{i_k}^{i_k,i_j}) &\triangleq \sum_{n=1}^{F} \bar{p}_{i_k,f_n} \left(|\sqrt{w_{i_k}} - \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_k,f_n}^2 \mathbf{C}_{i_k}(f_n, i_j) \tilde{p}_{i_k,i_j} \beta_{i_k,i_j} X_{i_k}^{i_k,i_j} |^2 \right) - \sum_{\substack{m=1 \\ m \neq k}}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} X_{i_m}^{i_k,i_j} \lambda_{i_m}^{i_k,i_j} \\ &\quad + \frac{\rho}{2} \sum_{\substack{m=1 \\ m \neq k}}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} |\sqrt{w_{i_m}} \mathbf{u}_{i_m}^H \mathbf{H}_{i_m,i_j} \mathbf{v}_{i_k,i_j} - X_{i_m}^{i_k,i_j}|^2, \\ h(\mathbf{x}_{i_k,i_j}) &\triangleq \mathrm{Re}\left(\sum_{k=1}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} \langle \mathbf{v}_{i_k,i_j} - \mathbf{x}_{i_k,i_j}, \mathbf{z}_{i_k,i_j} \rangle \right) + \frac{\rho}{2} \sum_{k=1}^{|\mathscr{I}_i|} \sum_{j=1}^{|\mathscr{Q}_i|} \|\mathbf{v}_{i_k,i_j} - \mathbf{x}_{i_k,i_j}\|^2 \end{aligned}$$

Note that all the three subproblems (4.20) and (4.22-4.23) are convex problems and can be solved efficiently. After calculating each variable at the $m+1$ iteration, we will update the λ_{i_k} as discussed earlier.

(b) Long-term Caching Placement (For given **D** and **v**)

Assuming user-association strategy and beamforming vectors are given, the caching problem can be decoupled among all the clusters in the aforementioned cache-enabled Cloud-RAN network. Precisely, the caching problem for the i-th cluster can be expressed as

$$\begin{aligned} \mathscr{P}_3 : \underset{\mathbf{C}_i}{\text{maximize}} & \left(\lambda \sum_{i=1}^{M} \sum_{k=1}^{|\mathscr{I}_i|} \sum_{n=1}^{F} \bar{p}_{i_k,f_n} R_{i_k,f_n} + (1-\lambda) \sum_{i=1}^{M} \sum_{n=1}^{F} \sum_{j=1}^{|\mathscr{Q}_i|} \mu_{i_j,f_n} \mathbf{C}_i(f_n, i_j) \right) \\ \text{subject to} & \quad \sum_{n=1}^{F} \mathbf{C}_i(f_n, i_j) \ell_{f_n} \leq s_{i_j}, \; \forall i \in \mathscr{M}, \; j \in \mathscr{Q}_i \end{aligned} \tag{4.24}$$

where $\mu_{i_j,f_n} = \sum_{k=1}^{|\mathscr{I}_i|} \bar{p}_{i_k,f_n} d_{i_j,i_k}$ represents the amount of backhaul saving due to storing f_n-th file in the i_j-th BS's cache. Maximization of $\sum_{n=1}^{F} \mu_{i_j,f_n} \mathbf{C}_i(f_n, i_j)$ for a simple SISO case in which both BSs and UEs are equipped with a single antenna, and each UE is only connected with one BS

is considered in [106]. This maximization can be understood as follows: Which files should be cached in the i_j-th BS to achieve the backhaul reduction μ_{i_j,f_n}? Under the condition that all files have the same size, the optimal solution to maximize $\sum_{n=1}^{F} \mu_{i_j,f_n} \mathbf{C}_i(f_n, i_j)$ is caching the s_{i_j} files that make the largest backhaul reduction; i.e.,

$$\mathbf{C}_i^*(f_n, i_j) = \begin{cases} 1, & \text{if } \mu_{i_j,f_n} \in \{\mu_{i_j,f_1}, \ldots, \mu_{i_j,f_{s_{i_j}}}\} \\ 0, & \text{Otherwise} \end{cases}$$

where $\mu_{i_j,s}$ is the s'th item which is more requested in the list of μ_{i_j,f_n} [106]. If each content has different sizes, the aforementioned caching strategy at each BS becomes a knapsack problem [117] that can be solved using dynamic programming. Here, we consider a more general case than the one discussed in [106], by considering the multi-cluster multi-user MIMO network in which each UE can be associated with a cluster of BSs. Consequently, the optimization problem becomes more complicated and the objective function can be written as:

$$\begin{aligned} O(\mathbf{C}_i, \mathbf{C}_{-i}) &= \sum_{i=1}^{M} \left(\lambda \sum_{k=1}^{|\mathscr{I}_i|} \sum_{n=1}^{F} \bar{p}_{i_k,f_n} \log_2 \det \left[\mathbf{I} + \mathbf{M}_i(\mathbf{C}_i) \mathbf{N}_i^{-1}(\mathbf{C}_{-i}) \right] + (1-\lambda) \sum_{n=1}^{F} \sum_{j=1}^{|\mathscr{Q}_i|} \mu_{i_j,f_n} \mathbf{C}_i(f_n, i_j) \right) \\ &\triangleq \sum_{i=1}^{M} O_i(\mathbf{C}_i, \mathbf{C}_{-i}) \end{aligned}$$

where $\mathbf{C}_{-i} \triangleq (\mathbf{C}_q)_{q \neq i}$ and

$$\mathbf{M}_i(\mathbf{C}_i) = \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}_{i_k,f_n}^2 \mathbf{C}_i(f_n, i_j) d_{i_j,i_k} \tilde{p}_{i_k,i_j} \beta_{i_k,i_j} \mathbf{H}_{i_k,i_j} \mathbf{v}_{i_k,i_j} \mathbf{v}_{i_k,i_j}^H \mathbf{H}_{i_k,i_j}^H$$

and

$$\mathbf{N}_i(\mathbf{C}_{-i}) = \sum_{(\ell,q) \neq (k,i)} \sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}_{q_\ell,f_n}^2 \mathbf{C}_q(f_n, q_j) d_{q_j,q_\ell} \tilde{p}_{q_\ell,q_j} \beta_{i_k,q_j} \mathbf{H}_{i_k,q_j} \mathbf{v}_{q_\ell,q_j} \mathbf{v}_{q_\ell,q_j}^H \mathbf{H}_{i_k,q_j}^H + \sigma^2 \mathbf{I}.$$

Since the entries of $\mathbf{C}_i$ take binary values 0 and 1, the optimization problem falls into a mixed integer nonlinear programming which is usually NP-hard in general [113] and non-tractable in

practice. Therefore, we are interested in obtaining a near-optimal solution. Inspired by the proposed method in [118], we allow the binary variables to take real values in $[0,1]$, and hence the original MINLP can be relaxed to a non-linear programming problem. Our aim is introducing a distributed solution method that efficiently computes the stationary solutions of this problem. To do so, we develop a Successive Convex Approximation (SCA)-based method and substitute a series of strongly convex problems for the optimization problem (4.24). The main idea here is approximating the non-convex objective function $O(\mathbf{C}_i,\mathbf{C}_{-i})$ by a suitable convex approximation. To be specific, the aim of BSs in each cluster is to choose a feasible cache placement matrix $\mathbf{C}_i$ that maximizes the objective function $O(\mathbf{C}_i,\mathbf{C}_{-i})$ assuming the strategy profile $\mathbf{C}_{-i}$ is given. Inspired by the introduced scheme in [5,119], our method is based on solving a sequence of parallel convex problems, one for each cluster. Each of these convex problems is obtained by maintaining the convex structure of the utility function while linearizing the rest around $\bar{\mathbf{C}}_i$. To isolate the inter-cluster and intra-cluster interferences which make $O(\mathbf{C}_i,\mathbf{C}_{-i})$ nonconvex, we define the utility function of the clusters other than the i-th cluster as

$$f_i(\mathbf{C}_i,\mathbf{C}_{-i})=\sum_{s\neq i}\left(\lambda\sum_{k=1}^{|\mathscr{I}_s|}\sum_{n=1}^{F}\bar{p}_{s_k,f_n}\log_2\det\left[\mathbf{I}+\mathbf{M}_s(\mathbf{C}_s)\mathbf{N}_s^{-1}(\mathbf{C}_{-s})\right]+(1-\lambda)\sum_{n=1}^{F}\sum_{j=1}^{|\mathscr{Q}_s|}\mu_{s_j,f_n}\mathbf{C}_s(f_n,s_j)\right).$$

Convexifing the objective function can be done by keeping the convex part; i.e., $O_i(\mathbf{C}_i,\mathbf{C}_{-i})$ while linearizing the nonconvex part; i.e., $f_i(\mathbf{C}_i,\mathbf{C}_{-i})$. As a consequence, we use the first order Taylor series expansion of the function $f_i(\mathbf{C}_i,\mathbf{C}_{-i})$ that is given by:

$$f_i(\mathbf{C}_i,\mathbf{C}_{-i})\approx f_i(\bar{\mathbf{C}}_i,\mathbf{C}_{-i})+f_i^{'}(\mathbf{C}_i,\mathbf{C}_{-i})|_{\mathbf{C}_i=\bar{\mathbf{C}}_i}(\mathbf{C}_i-\bar{\mathbf{C}}_i).$$

Recalling that $d/dx(\log_a(f(x)))=f(x)^{'}/(f(x)\ln a)$, the first-order differential is given by

$$f_i^{'}(\mathbf{C}_i,\mathbf{C}_{-i})=\sum_{s\neq i}\left(\lambda\sum_{k=1}^{|\mathscr{I}_s|}\sum_{n=1}^{F}\frac{\xi\bar{p}_{s_k,f_n}}{h(\mathbf{C}_i)}\frac{-\sum_{j=1}^{|\mathscr{Q}_s|}\bar{p}^2_{s_k,f_n}\mathbf{C}_s(f_n,s_j)d_{s_j,s_k}\tilde{p}_{s_k,s_j}\beta_{s_k,s_j}|\mathbf{H}_{s_k,s_j}\mathbf{v}_{s_k,s_j}|^2}{\left(\sum_{j=1}^{|\mathscr{Q}_q|}\bar{p}^2_{q_\ell,f_n}\mathbf{C}_q(f_n,q_j)d_{q_j,q_\ell}\tilde{p}_{q_\ell,q_j}\beta_{s_k,q_j}|\mathbf{H}_{s_k,q_j}\mathbf{v}_{q_\ell,q_j}|^2+\sigma^2\mathbf{I}\right)^2}\right)$$

where $\xi = \sum_{j=1}^{|\mathscr{Q}_i|} \bar{p}^2_{i_\ell,f_n} d_{i_j,i_\ell} \tilde{p}_{i_\ell,i_j} \beta_{s_k,i_j} |\mathbf{H}_{s_k,i_j} \mathbf{v}_{iq_\ell,i_j}|^2$ and

$$h(\mathbf{C}_i) \triangleq 1 + \left(\sum_{j=1}^{|\mathscr{Q}_s|} \bar{p}^2_{s_k,f_n} \mathbf{C}_s(f_n,s_j) d_{s_j,s_k} \tilde{p}_{s_k,s_j} \beta_{s_k,s_j} |\mathbf{H}_{s_k,s_j} \mathbf{v}_{s_k,s_j}|^2 \right) \times \left(\sum_{(\ell,q)\neq(k,s)} \sum_{j=1}^{|\mathscr{Q}_q|} \bar{p}^2_{q_\ell,f_n} \mathbf{C}_q(f_n,q_j) d_{q_j,q_\ell} \tilde{p}_{q_\ell,q_j} \beta_{s_k,q_j} |\mathbf{H}_{s_k,q_j} \mathbf{v}_{q_\ell,q_j}|^2 + \sigma^2 \mathbf{I} \right)^{-1}$$

By keeping only the linear term in the TaylorâĂŹs expansion of $f_i(\mathbf{C}_i, \mathbf{C}_{-i})$ around $\bar{\mathbf{C}}_i$ and adding a proximal like regularization term, the objective function in (4.24) can be approximated as

$$\tilde{O}(\mathbf{C}_i, \mathbf{C}_{-i}) = O_i(\mathbf{C}_i, \mathbf{C}_{-i}) + \mathbf{C}_i(f_n, i_j) f_i^{'}(\mathbf{C}_i, \mathbf{C}_{-i})|_{\mathbf{C}_i = \bar{\mathbf{C}}_i} - \frac{\tau_i}{2} |\mathbf{C}_i(f_n, i_j) - \bar{\mathbf{C}}_i(f_n, i_j)|^2$$

where τ_i is a given nonnegative constant. Now, it is possible to approximate (4.24) by a set of $|\mathscr{Q}_i|$ per cluster problems given for $i \in \mathscr{M}$ by

$$\underset{\mathbf{C}_i \in \mathscr{K}}{\text{maximize}} \quad \tilde{O}(\mathbf{C}_i, \mathbf{C}_{-i}) \tag{4.25}$$

where $\mathscr{K} \triangleq \{\mathbf{C}_i(f_n, i_j) | \sum_{n=1}^{F} \mathbf{C}_i(f_n, i_j) \ell_{f_n} \leq s_{i_j}\}$. Using the proposed algorithm in [119], for each BS we have got the following best response mapping which consisting in solving iteratively the sequence of a (strongly) convex optimization problem

$$\mathbf{C}_i^*(f_n, i_j) = \underset{\mathbf{C}_i \in \mathscr{K}}{\arg\max} \quad \tilde{O}(\mathbf{C}_i, \mathbf{C}_{-i}) \tag{4.26}$$

Unlike (4.24), (4.26) is strongly convex and can be efficiently solved by numerical iterative algorithms.

Remark 7. *(A Summary of Overall Operation) Utilizing the timescale separation of the optimization variables, we divide the original solution into short-term content delivery and long-term content placement. While the short-term process consists of user-association and beamforming*

optimization, the long-term process is composed of cache content placement. The content placement and the content delivery optimizations are performed in the cloud. The caching placement strategy is adaptive to the channel statistic. As soon as the user request profile changes, the cloud computes the updated cache placement and passes it to the BSs. Then the BSs update their cache. In each channel coherence time, the channel state information is acquired from the users through feedback, and the user-association and beamforming vectors are determined based on the instantaneous channel realization.

Remark 8. *(Computational Complexity) Since we employ the WMMSE algorithm [114] for designing the beamformers, the computational complexity of the beamforming optimization problem is much like the WMMSE, with the difference that the introduced ADMM-based algorithm decomposes the original large-scale problem into parallel small-scale subproblems. As a result, it needs more complex calculations than the coordinated descent method which is more desirable when the network size is small. However, the computation complexity of the proposed algorithm increases at a slower linear rate with respect to the number of users. The computational complexity of the user-association algorithm is similar to [98] and is polynomial in relation to the network size. However, we further lower the complexity of associating a user with a BS by taking content placement into account and excluding the candidate users that increase the backhaul consumption from consideration. The computational complexity of the long-term caching placement is exceedingly low due to the fact that for each realization of the user request profile, the only thing the introduced algorithm needs to do is a simple Jacobi/Gauss-Seidel update.*

4.4 Simulation Evaluations

In this preliminary simulation evaluation, we evaluate the performance of the proposed schemes in cache-enabled Cloud-RAN networks. The setup of our experiments is the following: we simulated a multi-cluster multi-user cache-enabled Cloud-RAN network in which the locations of the BSs are modeled using a PPP with density $\lambda_B = 1/(\pi R_B^2) = 5$ BS/km^2 which corresponds to an av-

erage inter-site distance of 500 m. Multiple users are randomly and uniformly distributed around each BS, excluding an inner circle of 35 meters, as illustrated in figure 5.1. The transmission is subject to interference from all neighboring base stations that do not serve the specific user. The transmit antenna power gain and the transmit power at each BS is set to 10 dBi and 46 dBm, respectively. The noise variance at the mobile station is fixed to -174 dBm. System bandwidth is taken as 5 MHz. We consider a possible antenna configuration in a typical deployment scenario for LTE/LTE-Advanced systems: 4 transmit and 2 receive antennas. The simulation is run for 1000 channel realizations where each channel is an uncorrelated Rayleigh fading and each channel element is drown i.i.d from a complex Gaussian distribution with zero mean and a variance of 1, i.e., $\mathscr{CN}(0,1)$. The path-loss is generated using 3 GPP (TR 36.814) methodology; i.e., $PL(dB) = 148.1 + 37.6\log_{10}(d)$, where d is the distance in kilometers. The log-normal shadowing parameter is assumed to be 8 dB. The total number of files available in the cloud is considered as $F = 20$. For the sake of simplicity, it is assumed that all files have the same size of one while the caches of the BSs can be filled with the $s = \{1,2,4,8,10\}$ bits of the most popular files. A Zipf-like distribution with parameter 0.56 is considered for the file popularity.

Figure 4.3 plots the average network throughput versus the SNR in the cache-enabled Cloud-

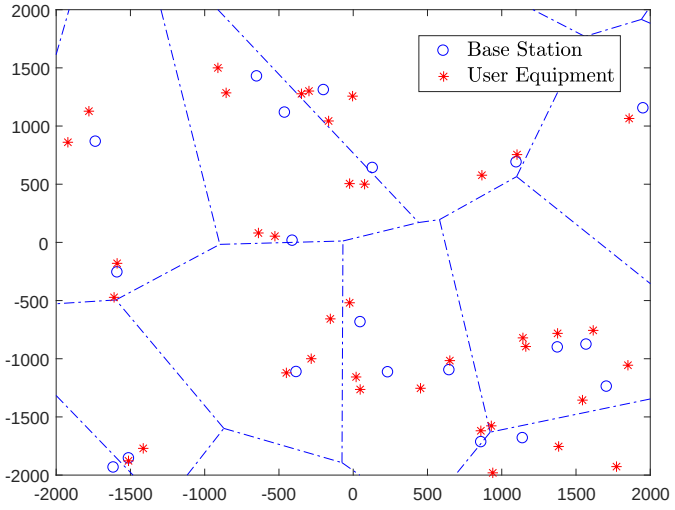

Figure 4.2. A realization of the Cloud-RAN network.

RAN containing of 10 user equipments. Here, we assume that each BS has the same cache size of 4. The algorithm is initialized by choosing a randomly generated feasible point. Moreover, the termination criterion is satisfied when the absolute value of the network throughput error in two consecutive rounds becomes smaller than $1e-2$.

Figure 4.4 shows the average network throughput versus the number of user equipments. Here we assumed that UEs have the same SNR, due to the fact that co-scheduled UEs usually have similar SNRs in multi-user MIMO operation. The average network throughput is plotted for two different SNR values. It is observed that the average sum rate gradually increases when the number of user equipments becomes larger. In addition, the average number of iterations versus the number of user equipments is plotted in figure 4.5. The average is taken over 1000 independent channel realizations. It is observed that the algorithm converges in several steps. Moreover, the average CPU time versus the total number of UEs is plotted in figure 5.3. Our experiments were run using Matlab R2016b on a 3.6 GHz Intel(R) Xeon(R) E51620 Processor Cores machine, equipped with 8 GB of memory. As can be expected, the average CPU time increases with the number of user equipments.

The influence of weighted coefficient on the introduced objective function in (10) is demonstrated in figure 4.7. The results can be interpreted as follows: when λ increases the backhaul saving plays a less critical role than the network throughput. Conversely, when λ approaches zero

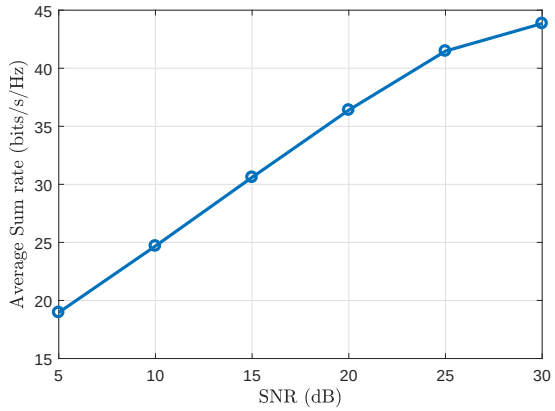

Figure 4.3. Average sum-rate versus SNR.

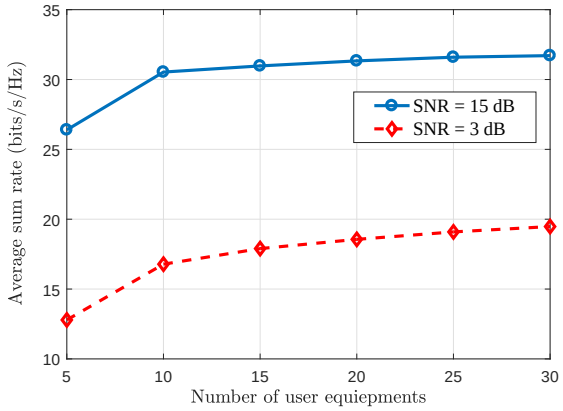

Figure 4.4. Average sum-rate versus number of UEs.

the backhaul reduction dominates the objective function. Figure 4.7 also investigates the influence of the cache size. It shows that the defined weighted objective function of the network with bigger cache size always performs better than the system in which the BSs have a smaller cache. The reason is, the smaller the cache sizes, the less the portions of contents cached. Therefore, the files need to fetch from cloud, which increases the backhaul usage. Moreover, when the size of caches grow from 2 to 8, an increase of approximately 76.42% in the backhaul saving is observed. Furthermore, 84% additional increase is acquired when the cache sizes enlarge from 1 to 2. It shows that even a small size of cache at each BS causes a substantial decrease in the backhaul usage.

Figure 4.8 demonstrate the performance comparison between our scheme and some benchmarks [90]. A different metric, normalized network cost is used for the comparison which is defined as a weighted sum of the backhaul cost and the transmit power cost. As shown in figure 4.8, compared to the full group sparse beamforming (F-GSBF) algorithm and the partial group sparse beamforming (F-GSBF) algorithm proposed in [90] which considered the caching matrix is given, our introduced algorithm can reduce the network cost which can interpreted as follows. Tacking caching placement into account can be viewed as an additional incentive for backhaul reduction.

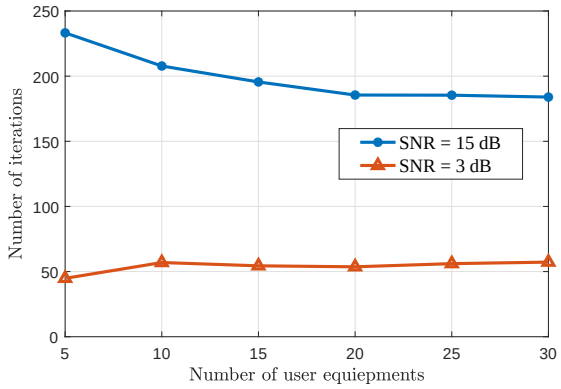

Figure 4.5. Average number of iterations vs number of UEs

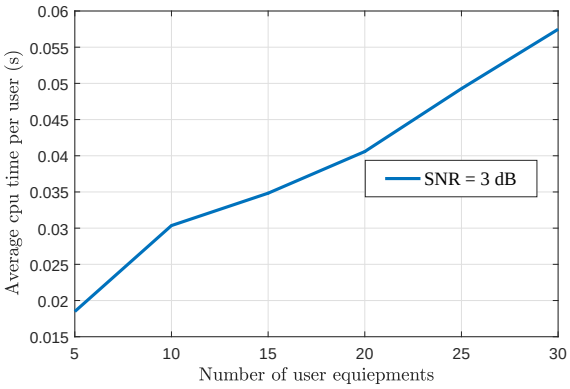

Figure 4.6. Average cpu time versus number of UEs

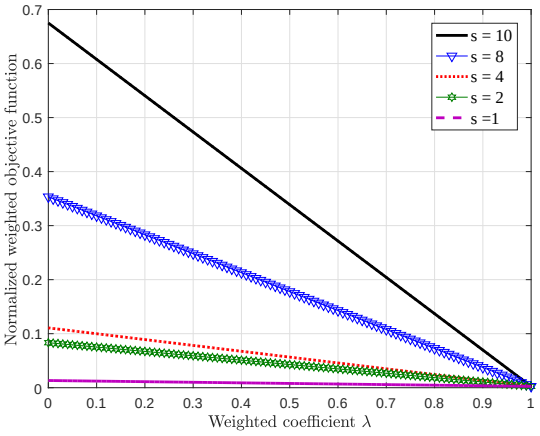

Figure 4.7. Normalized weighted objective function versus λ for different cache sizes.

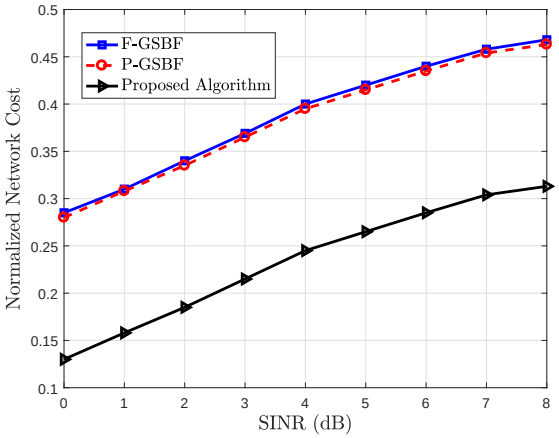

Figure 4.8. Normalized network cost versus SINR for different algorithms.

4.5 Conclusion

In this chapter we introduced a novel iterative algorithm to increase the network throughput and backhaul saving of a multi-cluster multi-user Cloud-RAN, by jointly optimizing the user association, caching placement, and beamforming design. The proposed algorithm utilizes the ADMM along with the SCA-based method and enables all base stations to update their optimization variables in parallel. Simulation results demonstrated that efficiently designing of the caching placement along with the user association and beamforming design will greatly influence the backhaul saving and network throughput.

Table 4.1: Nomenclatures and Notations Used

Notation	Description
B	Number of BSs
K	Number of UEs
M	Number of Clusters
$\lvert\mathscr{I}_i\rvert$	Number of UEs in the i^{th} cluster,
$\lvert\mathscr{Q}_i\rvert$	Number of BSs in the i^{th} cluster,
i_j	The j^{th} BS in the i^{th} cluster
i_k	The k^{th} UE in the i^{th} cluster
$N_t^{i_j}$	Number of transmit antennas at the i_j BS
$N_r^{i_k}$	Number of receive antennas at the i_k UE
$\mathbf{G}_{i_k,i_j}$	The channel coefficient between the i_j BS and the i_k UE
β_{i_k,i_j}	Large-scale fading coefficients
d_{i_k,i_j}	The distance between the i_k UE and the i_j BS
α	The path-loss exponent
ψ_{i_k,i_j}	A log-normal random variable
$\sigma_{\text{shadowing}}$	The standard deviation of shadowing
$\mathbf{H}_{i_k,i_j}$	The small-scale fading coefficients between the i_j BS and the i_k UE
F	Number of data files at the cloud
l_{f_n}	The size of the f_n^{th} file in bits
s_{i_j}	The capacity of the cache at the i_j BS
q_{i_k}	Number of requests made by the i_k UE over a given time interval
π_{i_k}	The index of the requested file by the i_k^{th} user
$\bar{p}_{i_k,f_n}$	Probability that the i_k UE request file f_n
p_{i_j,f_n}	The popularity distribution of file n observed at the i_j-th BS
$\mathbf{D}_i$	The user association matrix in the i^{th} cluster
$\mathbf{C}_i$	The cache placement matrix in the i^{th} cluster
$\mathbf{C}_{i_k}$	The caching matrix corresponding to the BSs that are associated with UEï£¡i_k
$\mathscr{B}_{i_k}$	A cooperating set of BSs assigned to i_k^{th} user
$\mathscr{K}_{i_j}$	A set of users served by i_j^{th} BS
s_{i_k,i_j}	The transmit symbol from the i_j BS and the i_k UE
$\tilde{p}_{i_k,i_j}$	The transmit power allocated to the i_k UE from the i_j BS
$\mathbf{v}_{i_k,i_j}$	The unit-norm beamformer from the i_j BS to the i_k UE
$P_{i_j}^{\max}$	The transmit power constraint at the i_j BS

Chapter 5

Cell-free Massive MIMO networks: Resource Allocation

5.1 Introduction

Multiple-input multiple-output (MIMO) systems with a large array antenna at the base stations (BSs), also known as Massive MIMO, have been widely studied recently to improve the spectral and energy efficiency of wireless systems with simple signal processing [120]. Due to the promising gains in [120], depicting high performance results with respect to the baseline MIMO system, more attention has been paid to the topic both in the academia [121, 122] and in the industry [123]. Massive MIMO allows a BS to simultaneously serve many number of user equipments (UEs) along with time-frequency resources to improve the overall system performance. In general, depending on the antenna arrays setup at the BSs, massive MIMO can be categorized into the following two architectures: distributed massive MIMO and co-located massive MIMO. While the latter one locates the service antennas in a compact area, the former spreads antennas all over a large area. Although the co-located architecture is attractive due to its low backhaual requirements, the distributed one offers higher coverage probability at the expense of increased infrastructure costs.

A cell-free Massive MIMO network, which has been introduced recently in [124], is a form of

distributed massive MIMO. It is considered a promising network architecture for the upcoming 5G wireless networks due to its ability to offer huge throughput and coverage probability to all users throughout a system. This architecture spreads a tremendous number of randomly-located access points (APs) over a large area, to simultaneously serve a much smaller number of single-antenna UEs, as an alternative to a small-cell network. In order to have an efficient resource allocation and interference management among multiple APs, unspecified backhaul links connect all these APs to a central processing unit (CPU) [124–126].

The performance of conjugate beamforming (CB) along with a pilot assignment algorithm has been investigated in [124, 125] to combat pilot contamination in a cell-free massive MIMO network. In order to boost system throughput, a max-min power allocation was considered. However, the optimal solution to the power allocation problem involves high computational complexity, due to the non-convexity of the optimization problem. To address this issue, [126] proposed a power allocation algorithm with a trade-off between low complexity and moderate decrease in performance. The authors in [126] further combined the max-min power allocation algorithm with a linear zero-forcing (ZF) precoder to tackle the high inter-user interference in CB technique. These studies restricted their discussion to the simplest linear precoding schemes which are CB and ZF, and did not design an optimal beamformer that ensures uniformly good service over all the coverage area.

In order to address this issue and improve the system performance along with per-user minimum level of service, we propose a design of the resource allocation in a way that the same quality of service can be provided to all UEs [127]. To enforce such fairness, we maximize (over the resource allocation: precoding vectors and power allocations) the minimum achievable rate among the users in a cell-free massive MIMO network. Due to the coupled interference among UEs, the resulting optimization problem is non-convex and difficult to solve. Therefore, we demonstrate the uplink-downlink duality and propose an iterative algorithm which solves the primal downlink problem efficiently by utilizing the result of the dual uplink. Exploiting proposed precoder and taking the channel estimation error into consideration, we derive the lower bound for the capacity

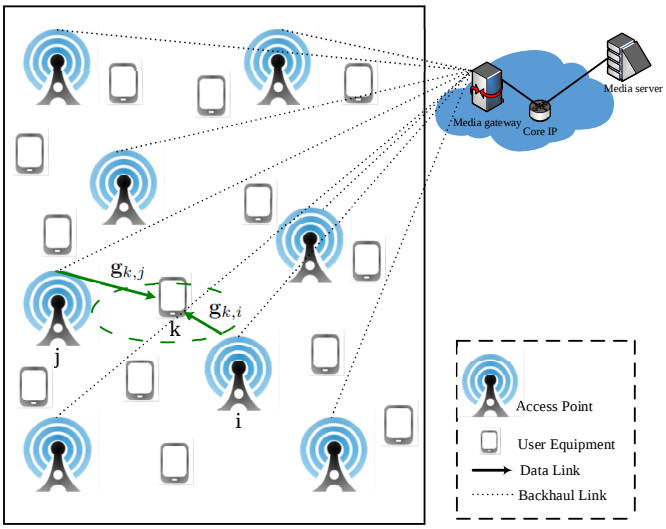

Figure 5.1. System architecture of a cell-free massive MIMO network.

of the underlying cell-free massive MIMO system. Furthermore, unlike [124–126] which assumed each single-antenna AP serves all UEs, we consider a cluster of multiple-antenna APs serving each UE. Note that, the strongest APs under coordination may not necessarily be the nearest ones to the UE, due to the shadowing effect.

5.2 System Model and Assumptions

We consider a downlink cell-free massive MIMO network consisting of one central processing unit, M access points, and K user equipments as depicted in Figure 5.1. The APs and UEs are randomly distributed over a large area. To exchange the network information, all APs are connected to the CPU through error-free infinite-capacity backhaul links.[1] Each AP is equipped with N_t transmit antennas, serves several single-antenna UEs. We assume that each UE is served by a cluster of cooperative APs. To be specific, a cooperating set $\mathscr{B}_k$ is assigned to the k^{th} UE, where $\mathscr{B}_k$ is formed by aggregating all the APs that have knowledge of channels to the k^{th} UE, access to its message, and may also jointly encode the intended message for the k^{th} UE in their transmission [110]. Note that, due to the shadowing effect, the $|\mathscr{B}_k|$ strongest APs under coordination are not necessarily the

[1] However, in a practical scenario the backhaul links are subject to practical constraints and investigating the effect of these constraints would be an important topic for future work.

$|\mathscr{B}_k|$ nearest APs, where $|\mathscr{B}_k|$ denote the setâĂŹs cardinality. Since each AP may get involved in transmission to more than one UE, the cooperating set of different users may overlap. We denote the set of users served by the i^{th} AP by $\mathscr{K}_i = \{k | i \in \mathscr{B}_k\}$.

The channel (propagation) coefficient between the i^{th} AP and the k^{th} UE form channel vector $\mathbf{g}_{k,i} = \sqrt{\beta_{k,i}}\mathbf{h}_{k,i} \in \mathbb{C}^{N_t}$ where $\beta_{k,i}$ and $\mathbf{h}_{k,i}$ indicate the large-scale and small-scale fading coefficients, respectively. The large-scale fading coefficient depends upon the shadowing and distance between the corresponding UE and AP, and is denoted by $\beta_{k,i} = \psi_{k,i} d_{k,i}^{-\alpha}$ where $d_{k,i}$ is the distance between the k^{th} user and the i^{th} AP; α is the path-loss exponent; and $\psi_{k,i}$ is a log-normal random variable, i.e., the quantity $10\log_{10}(\psi_{k,i})$ is distributed zero-mean Gaussian with a standard deviation of $\sigma_{\text{shadowing}}$. Here, we employ the same three slope path-loss model as in [125, 128, 129]. The path loss from the k^{th} user to the i^{th} AP is modeled as COST-231 Hata propagation model [130] and can be expressed in dB as

$$PL_{k,i} = \begin{cases} -L - 35\log_{10}(d_{k,i}), & \text{if } d_{k,i} > d_1 \\ -L - 10\log_{10}(d_{k,i}^2 d_1^{1.5}), & \text{if } d_0 < d_{k,i} \leq d_1 \\ -L - 10\log_{10}(d_0^2 d_1^{1.5}), & \text{if } d_{k,i} \leq d_0 \end{cases}$$

where

$$\begin{aligned} L = {} & 46.3 + 33.9\log_{10}(f) - 13.82\log_{10}(h_{\text{AP}}) \\ & - \Big(1.11\log_{10}(f) - 0.7\Big)h_{\text{UE}} + 1.56\log_{10}(f) - 0.8 \end{aligned}$$

where f is the carrier frequency in MHz, and h_{AP} and h_{UE} indicate the effective transmitter and receiver antenna heights in meters, respectively. Furthermore, we assume that the shadow fading random variables are correlated, since in a practical scenario the APs and UEs may be located close by, and therefore be all around the common obstacles. To model this correlation, we use the

two-component shadowing correlation model, as in [125, 128, 131],

$$z_{k,i} = \sqrt{\delta} a_i + \sqrt{1-\delta} b_k$$

where $0 \leq \delta \leq 1$ depicts the cross-correlation at the AP while $(1-\delta)$ denotes the cross-correlation at the UE, and $a_i \sim \mathcal{N}(0,1)\ \forall i = 1,\ldots,M$ and $b_k \sim \mathcal{N}(0,1)\ \forall k = 1,\ldots,K$ are independent random variables with $\mathbb{E}\{a_i a_j^*\} = 2^{-d_{\text{AP}}(i,j)/d_{\text{decorr}}}$ and $\mathbb{E}\{b_k b_\ell^*\} = 2^{-d_{\text{UE}}(k,\ell)/d_{\text{decorr}}}$ where $d_{\text{AP}}(i,j)(d_{\text{UE}}(k,\ell))$ denotes the geographical distance between the i^{th} and j^{th} (k^{th} and ℓ^{th}) APs (UEs), and d_{decorr} is the decorrelation distance. Therefore, the large-scale fading coefficients can be modified as $\beta_{k,i} = 10^{PL_{k,i}/10} \cdot 10^{z_{k,i}\sigma_{\text{shadowing}}/10}$. The small-scale fading coefficients, i.e., elements of $\mathbf{h}_{k,i}$, are modeled as i.i.d. complex Gaussian variables with zero-mean and unit-variance [124]. We further assume a block fading model, where small-scale channels are constant over a few time slots, with respect to channel estimation and channel state information (CSI) feedback procedures [132, 133]. Similarly, we assume that large-scale fading coefficients stay constant during large-scale coherence blocks. The small-scale and large-scale fading coefficients in different coherence blocks are assumed to be independent.

Time-division duplexing (TDD) mode is assumed in this system. In the uplink training phase, the users send pilot sequences to the AP synchronously. Each AP estimates the channel to all users based on the received pilot signals. Assuming $\beta_{k,i}$ is known, as in [124], the i^{th} AP computes the minimum mean squared error (MMSE) estimate of the channel vectore $\mathbf{g}_{k,i}$ as $\hat{\mathbf{g}}_{k,i}$. Considering the channel estimation error as $\tilde{\mathbf{g}}_{k,i} = \mathbf{g}_{k,i} - \hat{\mathbf{g}}_{k,i}$, it is easy to show that $\tilde{\mathbf{g}}_{k,i}$ and $\hat{\mathbf{g}}_{k,i}$ are uncorrelated [134]. In the downlink phase, the channel estimates $\hat{\mathbf{g}}_{k,i}$ are regarded as true channels by APs, i.e., APs rely on channel hardening, and use these channel estimates to precode data to UEs. Each AP, say $i \in \{1,2,\ldots,M\}$, plans to communicate a symbol vector $\mathbf{s}_i = [s_{i_1},\ldots,s_{i_{|\mathscr{K}_i|}}]^T \in \mathbb{C}^{|\mathscr{K}_i|}$ to its associated receivers, where s_{i_k} is the transmit symbol from the i^{th} AP to the k^{th} receiver with unit power of $E\{|s_{i_k}|^2\} = 1$. Prior to transmitting, the i^{th} AP linearly precodes its symbol vector $\mathbf{x}_i = \sum_{k=1}^{|\mathscr{K}_i|} \mathbf{f}_{i_k} s_{i_k}$ where $\mathbf{f}_{i_k}$ denotes the precoding vector that the i^{th} AP uses to transmit the

signal s_{i_k} to the k^{th} UE and $||\mathbf{f}_{i_k}||^2$ is the allocated downlink transmit power. Each AP i is under a transmit power constraint of $P_{i,\max}$ and so, the transmit power at the i^{th} AP is computed as $p_i = E\{||\mathbf{x}_i||^2\} \leq P_{i,\max}$ [**?**,**?**]. In order to transmit data to the k^{th} UE, coordinated joint transmission is used to coherently combine the received signals from APs that serve the k^{th} UE. Since there are $|\mathscr{B}_k|$ APs participating in the cooperating data transmission to user k, at the same frequency and time, we denote $s_k \in \mathbb{C}$ as the complex data symbol for the k^{th} UE. Hence, over each symbol duration time, the cooperating APs transmit the same symbol s_k. The received signal at the k^{th} receiver can be written as:

$$y_k = \sum_{i \in \mathscr{B}_k} \mathbf{g}_{k,i} \mathbf{f}_{i_k} s_k + \sum_{\substack{j=1 \\ j \notin \mathscr{B}_k}}^{M} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \mathbf{g}_{k,j} \mathbf{f}_{j_\ell} s_\ell + n_k \tag{5.1}$$

where the first term on the right-hand side of (5.1) depicts the received useful signal, the second term denotes the multi-user interference, and $n_k \sim \mathscr{CN}(0,1)$ is the additive white Gaussian noise (AWGN) at the k^{th} UE. Recall that the signal-to-interference-plus-noise ratio (SINR) is defined as the ratio of the received signal power at the desired user to the interference power plus the noise power, the SINR at the k^{th} UE can be expressed as

$$\begin{aligned} \text{SINR}_k &= \frac{\sum_{i \in \mathscr{B}_k} \beta_{k,i} \text{Tr}(\mathbf{f}_{i_k}^H \mathbf{h}_{k,i}^H \mathbf{h}_{k,i} \mathbf{f}_{i_k})}{\sigma_k^2 + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \sum_{j \in \mathscr{B}_\ell} \beta_{\ell,j} \text{Tr}(\mathbf{f}_{j_\ell}^H \mathbf{h}_{k,j}^H \mathbf{h}_{k,j} \mathbf{f}_{j_\ell})} \\ &= \frac{\sum_{i=1}^{M} \beta_{k,i} \text{Tr}(\mathbf{f}_{i_k}^H \mathbf{h}_{k,i}^H \mathbf{h}_{k,i} \mathbf{f}_{i_k})}{\sigma_k^2 + \sum_{\substack{j=1 \\ j \neq i}}^{M} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \beta_{\ell,j} \text{Tr}(\mathbf{f}_{j_\ell}^H \mathbf{h}_{k,j}^H \mathbf{h}_{k,j} \mathbf{f}_{j_\ell})} \end{aligned}$$

where the second equality holds under the assumption that $\mathbf{f}_{i_k} = \mathbf{0}$ for $i \notin \mathscr{B}_k$. Therefore, the downlink achievable transmission rate for the k^{th} UE can be expressed as $R_k = \log_2(1 + \text{SINR}_k)$.

5.3 A Novel Algorithm for Downlink Resource Allocation

In order to assure a uniform user experience across the network along with the improved system performance, the problem of interest in this chapter is to maximize (over the resource allocation: precoding vectors and power allocations) the minimal performance of each user in a cell-free massive MIMO network. Precisely, we are interested in the max-min SINR problem, over the transmit precoder and power control, given by

$$\max_{\mathbf{f}_{i_k}} \min_{k} \quad \text{SINR}_k \tag{5.2}$$

which is subject to $\sum_{k\in\mathscr{K}_i} \mathbf{f}_{i_k}^H \mathbf{f}_{i_k} < P_{i,\max}$.

Due to the interference present, the optimization problem (5.2) is non-convex and hence, finding the global optimum is challenging. Therefore, in order to efficiently solve this optimization problem we take advantage of the fact that in contrast to the downlink SINR which is coupled with both precoders and transmit powers; the uplink SINR is only accompanied by transmit power, thereby, it provides an easier optimization problem to solve. Consequently, to exploit network duality, we need to reformulate the optimization problem (5.2), in terms of precoders. We use the following theorem for this purpose, which establishes the uplink-downlink duality for the downlink beamforming in a cell-free massive MIMO network.

Theorem 10. *Let p_k be the uplink transmit power of user k and p_{i_k} stands for the allocated downlink transmit power to user k from the i-th AP. Then, the uplink and downlink SINR will be equal if the downlink power allocation satisfies the same power constraint as in the uplink, i.e.,* $\sum_{i=1}^{M}\sum_{k\in\mathscr{K}_i} p_{i_k} = \sum_{k=1}^{K} p_k$.

Proof. The proof can be achieved by following the procedures incorporated by Theorem 3 in [135]. □

Using Lagrangian duality presented in Theorem 10, the primal downlink problem (5.2) can be solved by exploiting the result of the dual uplink problem. Therefore, the optimal beamforming

vector in (5.2) can be acquired using the following theorem.

Theorem 11. *The optimal precoder $\mathbf{f}_{i_k}$, which the i-th AP uses to transmit the signal s_{i_k} to the k-th UE, is given by $\mathbf{f}_{i_k}^* = \sqrt{p_{i_k}}\mathbf{u}_k^H/||\mathbf{u}_k||^2$ where $\mathbf{u}_k$ is the optimal receive beamformer of the Lagrangian dual uplink problem of (5.2) and can be expressed as $\mathbf{u}_k = \mathbf{u}_{opt}[(k-1)N_t+1 : kN_t]$ where $\mathbf{u}_{opt} = \nu_{\max}(\mathbf{B}_i^{-1}\mathbf{A}_i)$ and $\mathbf{A}_i$ and $\mathbf{B}_i$ are defined as follows*

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{R}_{i,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{i,2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_{i,|\mathscr{K}_i|} \end{pmatrix}, \quad \mathbf{B}_i = \mathbf{I} + \sum_{j \neq i} \mathbf{A}_j$$

where $\mathbf{R}_{i,k} = p_k \mathbf{g}_{i,k}^H \mathbf{g}_{i,k}/\sigma_i^2$.

Proof. The Lagrangian dual uplink problem of the underlaying downlink precoding problem (5.2) can be written as

$$\begin{aligned} \max_{\mathbf{u}_k} \quad & \frac{\sum_{k\in\mathscr{K}_i} \mathbf{u}_k^H \mathbf{R}_{i,k} \mathbf{u}_k}{\mathbf{u}_k^H \left(\mathbf{I} + \sum_{j\neq i} \sum_{\ell \in \mathscr{K}_j} \mathbf{R}_{i,\ell} \right) \mathbf{u}_k} \\ \text{s.t.} \quad & \sum_{k\in\mathscr{K}_i} p_k < P_{\max} \end{aligned}$$

The above optimization problem can be rearranged as

$$\max_{\mathbf{u}} \quad \frac{\mathrm{Tr}(\mathbf{u}^H \mathbf{A}_i \mathbf{u})}{\mathrm{Tr}(\mathbf{u}^H \mathbf{B}_i \mathbf{u})} \tag{5.3}$$

where $\mathbf{u} \triangleq \mathrm{vec}(\mathbf{u}_1^T \ \mathbf{u}_2^T \ \ldots \ \mathbf{u}_{|\mathscr{K}_i|}^T)$ and the power constraint will be satisfied with $||\mathbf{u}||^2 \leq P_{\max}$. It can be shown that the optimum beamformer $\mathbf{u}$ at problem (5.3) can be obtained by generalized eigenvalue decomposition as $\mathbf{u}_{\mathrm{opt}} = \nu_{\max}(\mathbf{B}_i^{-1}\mathbf{A}_i)$. □

Utilizing the optimal precoder $\mathbf{f}_{i_k}^*$ as given in Theorem 11, it can be shown easily that the optimal downlink transmit power at each AP can be given using the following corollary.

Corollary 1. *Using the optimal beamformer $\mathbf{f}_{i_k}^*$ and assuming the perfect channel state information is available, the optimal downlink transmit power at the i^{th} AP can be expressed as follows:*

$$\begin{aligned}\mathbf{p}_i^* &= [p_{i_1}^*, \ldots, p_{i_k}^*, \ldots, p_{i_{|\mathscr{K}_i|}}^*]^T \\ &= \xi_i \sigma_{i_k}^2 (\mathbf{I}_{|\mathscr{K}_i|} - \xi_i \Gamma \Upsilon)^{-1} \Gamma \mathbf{1}_{|\mathscr{K}_i|}\end{aligned}$$

where $p_{i_k}^*$ denotes the optimal allocated downlink transmit power to the k^{th} user from the i^{th} AP, Γ is a diagonal matrix where the k^{th} nonzero elements in the diagonal is given by $\|\mathbf{f}_{i_k}^*\|^2/|\mathbf{g}_{i,k}^H \mathbf{f}_{i_k}^*|^2$ while all its non-diagonal elements are zero, the entries outside the main diagonal of matrix Υ are given by $[\Upsilon]_{k,i} = |\mathbf{g}_{i,k}^H \mathbf{f}_{i_k}^*|^2/\|\mathbf{f}_{i_k}^*\|^2$ while all its diagonal elements are zero, and ξ_i can be expressed as

$$\xi_i = \frac{P_{i,\max}}{\sum_{k=1}^{|\mathscr{K}_i|} (\mathbf{g}_{i,k}^H \mathbf{f}_{i_k}^*)^{-1}}.$$

Taking into consideration the existence of channel estimation error, the following theorem provides the lower bound on UE achievable rate in the underlying cell-free massive MIMO network.

Theorem 12. *Considering the channel estimation error and by utilizing max-min precoder and power control, the achievable rate lower bound of underlying cell-free massive MIMO network can be expressed as*

$$\log_2\Big(1 + \frac{\mathbb{E}\Big\{ |\sum\limits_{i \in \mathscr{B}_k} \hat{\mathbf{g}}_{k,i} \mathbf{f}_{i_k} s_k|^2 \Big\}}{\mathbb{E}\Big\{ |\sum\limits_{\substack{j=1 \\ j \notin \mathscr{B}_k}}^{M} \sum\limits_{\substack{\ell=1 \\ \ell \neq k}}^{K} \hat{\mathbf{g}}_{k,i} \mathbf{f}_{j_\ell} s_\ell|^2 + |\sum\limits_{j=1}^{M} \sum\limits_{\ell=1}^{K} \tilde{\mathbf{g}}_{k,j} \mathbf{f}_{j_\ell} s_\ell|^2 \Big\} + 1}\Big)$$

Proof. Taking the channel estimation error $\tilde{\mathbf{g}}_{k,i}$ into consideration, the received signal in (5.1) can

be re-evaluated as

$$y_k = \underbrace{\sum_{i\in\mathscr{B}_k} \hat{\mathbf{g}}_{k,i}\mathbf{f}_{i_k}s_k}_{\text{desired signal}} + \underbrace{\sum_{\substack{j=1\\ j\notin\mathscr{B}_k}}^{M}\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} \hat{\mathbf{g}}_{k,i}\mathbf{f}_{j_\ell}s_\ell}_{\text{multi-user interference}} + \underbrace{\sum_{j=1}^{M}\sum_{\ell=1}^{K} \tilde{\mathbf{g}}_{k,j}\mathbf{f}_{j_\ell}s_\ell}_{\text{channel estimation error}} + n_k$$

All the terms in the right hand side of the above equation are mutually uncorrelated due to the assumption that the signals are different for distinctive users, and n_k is independent from the channel coefficients and data symbols. In order to acquire the lower bound of the achievable rate, we take the worst case noise into account. It is shown in [136, Theorem 1] that the worst case noise is a Gaussian additive noise with the variance equal to the accumulation of variance of noise, multi-user interference, and channel estimation error. Therefore, the SINR can be written as

$$\text{SINR}_k = \frac{\mathbb{E}\left\{ |\sum_{i\in\mathscr{B}_k} \hat{\mathbf{g}}_{k,i}\mathbf{f}_{i_k}s_k|^2 \right\}}{\mathbb{E}\left\{ |\sum_{\substack{j=1\\ j\notin\mathscr{B}_k}}^{M}\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} \hat{\mathbf{g}}_{k,i}\mathbf{f}_{j_\ell}s_\ell|^2 + |\sum_{j=1}^{M}\sum_{\ell=1}^{K} \tilde{\mathbf{g}}_{k,j}\mathbf{f}_{j_\ell}s_\ell|^2 \right\} + 1}$$

and the achievable rate lower bound of underlying cell-free massive MIMO network can be expressed as $\log_2(1+\text{SINR}_k)$. □

5.4 Simulation Evaluations

In this preliminary experimental evaluation, we evaluate the performance of the proposed schemes for a downlink cell-free massive MIMO network. The setup of our experiments is in accordance with the ones in [124, 125, 128] and given as follows. We simulated a cell-free massive MIMO network in which M APs and K single-antenna UEs are randomly distributed over a square dense urban area of size 1000×1000 m^2. We employ a three-slope path-loss model along with an uncorrelated shadowing model as discussed in Section II. The transmission is subject to interference from all APs in the area. We assume that all the pilot sequences have the same length as the number

of UEs, i.e., $\tau = K$. All deployments and channel model parameters are listed in Table 5.1. The performance is measured in terms of the power consumption (dBm) and the rate (bit/s) per each UE. Moreover, we considered the following schemes for comparison under simulation: CB with optimal power allocation [124] and ZF with optimal power allocation [126].

The average total power consumption per UE in dBm is modeled as in [137, 138] and demonstrated in Figure 5.2. We compare the proposed resource allocation scheme with the co-located massive MIMO, the cell-free ZF with optimal power allocation, and the cell-free CB with optimal power allocation. It is shown that although the cell-free ZF algorithm with optimal power allocation causes an improved energy efficiency in compare with cell-free CB and co-located massive MIMO architecture, by designing the beamforming and power allocation vectors the proposed algorithm provides better improvements in terms of energy efficiency.

Figure 5.3 depicts the average rate per UE versus the number of UEs in the underlying network, for both cell-free and co-located massive MIMO architectures. It is demonstrated that the cell-free massive MIMO architecture outperforms the co-located one for different number of UEs as well as resource allocation schemes. The results indicate that for smaller number of UEs the proposed algorithm provides slightly higher achievable rate than other schemes. However, it is notable that

Table 5.1: System Parameters

Parameters	**Values**
Shadowing standard deviation	8 dB
Transmitted power of each UE	20 dBm
AP radiated power	23 dBm
Carrier frequency	1.9 GHz
Bandwidth	20 MHz
Noise figure	9 dB
Thermal noise level	-174 dBm/Hz
AP antenna height	15 m
UE antenna height	1.65 m
d_1	50 m
d_0	10 m
d_{decorr}	100 m
δ	0.5

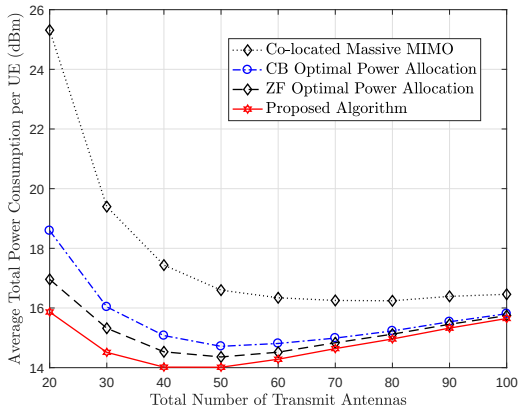

Figure 5.2. Average total power consumption versus total number of transmit antennas in a cell-free massive MIMO network.

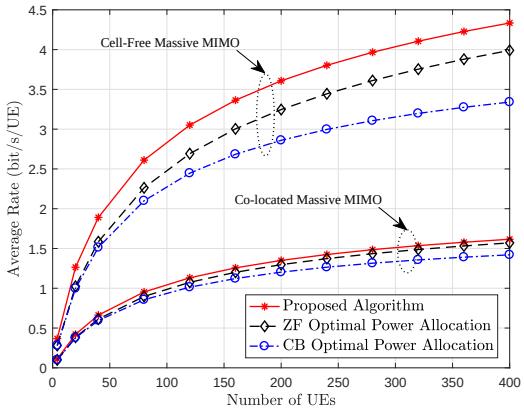

Figure 5.3. Downlink average rate per UE versus total number of UEs in both cell-free and co-located massive MIMO networks.

with an increase in the number of UEs, the achievable rate is significantly higher.

5.5 Conclusion

We have studied the downlink resource allocation in a cell-free massive MIMO network and proposed an algorithm for maximizing (over the resource allocation: precoding vectors and power allocations) the minimum achievable rate among the UEs in the network. The optimal solution of this problem was obtained by utilizing the uplink-downlink duality. Exploiting proposed beamformers and taking the effects of channel estimation error into consideration, we analyzed the per-

formance of underlying cell-free massive MIMO network and derived the capacity lower bound. Through the system level evaluations, the results showed that the proposed algorithm significantly outperforms the conventional resource allocation schemes in practical environments.

Chapter 6

Conclusions

These days, wireless communications are becoming so tightly integrated in our daily lives, especially with the global spread of laptops, tablets and smartphones. This has paved the way to dramatically increasing wireless network dimensions in terms of subscribers and amount of flowing data. Due to ever-increasing usage of wireless mobile devices, it has become necessary to improve the spectral efficiency of wireless networks. The next generation of wireless communication networks, also known as 5*G*, is supposed to do that. The two important fundamental requirements for the future 5*G* wireless networks are abilities to support high data traffic and exceedingly low latency. A likely candidate to fulfill these requirements is multi-cell multi-user multi-input multiple-output (MU-MIMO); also termed as coordinated multi-point (CoMP) transmission and reception in 3*GPP* LTE-Advanced systems. In order to achieve the highest possible performance of this aforementioned candidate technology, a properly designed resource allocation algorithm is needed. By designing a resource allocation algorithm which maximizes the network throughput, this technology is able to manage the exponential growth of wireless network dimensions. Moreover, with the rapidly growing data traffic, interference has become a major limitation in wireless networks. To deal with this issue and in order to manage the interference in the wireless network systems, various interference mitigation techniques have been introduced in literature among which interference alignment (IA) has been shown to significantly improve the network

performance. However, how to practically use IA to mitigate inter-cell interference in downlink multi-cell multiuser MIMO networks still remains an open problem.

In this dissertation, we improved the performance of wireless networks in terms of spectral efficiency, by developing new algorithms and protocols that can efficiently mitigate both inter-cell and intra-cell interference, and allocate the resources. In particular, we focused on designing new precoding algorithms. Furthermore, we mathematically analyzed the performance improvement of multi-user MIMO networks employing proposed techniques and revealed the fundamental relationships between network parameters and the network performance, which provided guidance on the wireless networks design.

In Chapter 2, we reviewed various interference mitigation techniques in multi-cell multi-user MIMO networks in order to develop an efficient interference mitigation method. In this chapter, limited feedback-based co-channel interference mitigation of a multi-cell multi-user MIMO was investigated. To be specific, limited feedback-based interference alignment was introduced to eliminate the inter-cell interference while an improved low complexity iterative leakage-based coordinated beamforming strategy was introduced to mitigate the intra-cell interference. By jointly considering the transmit beamforming, receive beamforming, and the quantization error of the codebook based feedback, our algorithm greatly mitigates the co-channel interference and achieves a better performance compared to the traditional ones. Through system level evaluations, it was shown that the introduced scheme significantly outperforms the conventional interference management schemes in practical environments.

In Chapter 3, we analyzed the state-of-the-art to identify the sub-optimal resource allocation technique in Coordinated Multi-Point transmission and reception networks. In this chapter, a resource allocation problem has been studied for downlink CoMP coordinated beamforming systems where each base station serves its own mobile stations. Due to the coupled interference among mobile stations the resulting optimization problem became non-convex. To solve for optimal resource allocation strategies including downlink precoding and power allocation for CoMP-CB transmissions, we have introduced a stochastic parallel successive convex approximation-based algorithmic

framework for a general non-convex network proportional-fair metric optimization problem. The introduced novel decomposition enabled all base stations to update their optimization variables in parallel by solving a sequence of strongly convex subproblems. Moreover, closed-form expressions of the locally optimal solution were characterized in some special cases as well as in both high and low SNR regimes. Numerical results shown that the introduced scheme significantly improved and optimized the system performance by mitigating inter-cell interference.

In Chapter 4, we reviewed various aspects of wireless caching technology to develop a backhaul-aware network utility objective function. To this end, available studies on this topic was summarized and organized in such a way that helps the reader to develop a critical perspective. In this chapter, we introduced a novel iterative algorithm to increase the network throughput and backhaul saving of a multi-cluster multi-user Cloud-RAN, by jointly optimizing the content delivery and content placement. The proposed algorithm utilized the alternating direction method of multipliers along with the successive convex approximation-based method, and enabled all base stations to update their optimization variables in parallel. Simulation results demonstrated that efficiently designing of the caching placement along with the user association and beamforming design will greatly influence the backhaul saving and network throughput.

Finally, in Chapter 5, We have studied the downlink resource allocation in a cell-free massive MIMO network and proposed an algorithm for maximizing (over the resource allocation: precoding vectors and power allocations) the minimum achievable rate among the users in the network. By utilizing the uplink-downlink duality, the optimal solution of this problem was obtained. Exploiting proposed beamformers and taking the effects of channel estimation error into consideration, we furthermore analyzed the performance of underlying cell-free massive MIMO network and derived the capacity lower bound. Through the system level evaluations, the results showed that the proposed algorithm significantly outperforms the conventional resource allocation schemes in practical environments.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2016âĂŞ2021," Cisco Systems, March 2017.

[2] S. Mosleh, L. Liu, H. Hou, and Y. Yi, "Coordinated data assignment: A novel scheme for big data over cached cloud-RAN," *IEEE Global Communications Conference (GLOBECOM)*, Feb. 2017.

[3] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[4] L. Liu, J. C. Zhang, Y. Yi, H. Li, and J. Zhang, "Combating interference: MU-MIMO, CoMP, and HetNet," *J. of Commun.*, vol. 7, no. 9, pp. 646–655, Sept. 2012.

[5] S. Mosleh, L. Liu, and J. Zhang, "Proportional-fair resource allocation for coordinated multi-point transmission in LTE-advanced," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5355–5367, Aug. 2016.

[6] B. Farhang-Boroujeny, Q. Spencer, and A. L. Swindlehurst, "Layering techniques for space-time communications in multi-user networks," *58th IEEE Vehicular Technology Conf. (VTC) 2003*, pp. 1339–1343, Oct. 2003.

[7] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[8] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, March 2006.

[9] M. Sadek, A. Tarighat, and A. H. Sayed, "Active antenna selection in multiuser MIMO communications," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1498–1510, April 2007.

[10] C. Suh and D. Tse, "Interference alignment for cellular networks," *46th Annu. Allerton Conf. on Commun., Control, and Computing*, pp. 1037–1044, Sept. 2008.

[11] C. Suh, M. Ho, and D. Tse, "Downlink interference alignment," *IEEE Trans. Commun.*, vol. 59, no. 9, pp. 2616–2626, Sept. 2011.

[12] L. Liu and J. Zhang, "New leakage-based iterative coordinated beamforming for multi-user mimo in LTE-advanced," *IEEE Int. Conf. on Commun. (ICC) 2012*, pp. 2308–2312, June 2012.

[13] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3457–3470, Aug. 2008.

[14] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom for the K user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[15] S. Mosleh, J. Abouei, and M. R. Aghabozorgi, "Distributed opportunistic interference alignment using threshold-based beamforming in MIMO overlay cognitive radio," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3783–3793, Oct. 2014.

[16] S. Mosleh, L. Liu, Y. Li, and J. Zhang, "Interference alignment and leakage-based iterative coordinated beam-forming for multi-user MIMO in LTE-Advanced," *IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, Dec. 2015.

[17] R. T. Krishnamachari and M. K. Varanasi, "Interference alignment underlimited feedback for MIMO interference channels," *in Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1339–1343, Jun. 2010.

[18] R. Tresch and M. Guillaud, "Cellular interference alignment with imperfect channel knowledge," *in Proc. IEEE Int. Conf. on Commun. (ICC)*, pp. 1339–1343, Jun. 2009.

[19] O. E. Ayach and R. W. Heath, "Interference alignment with analog channel state feedback," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 626–636, Feb. 2012.

[20] J. Kim, S. Moon, S. Lee, and I. Lee, "A new channel quantization strategy for MIMO interference alignment with limited feedback,," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 358–366, Jan. 2012.

[21] S. Cho, K. Huang, D. Kim, and H. Seo, "Interference alignment for uplink cellular systems with limited feedback," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 960–963, July 2012.

[22] N. Lee, W. Shin, R. W. H. Jr., and B. Clerckx, "Interference alignment with limited feedback for two-cell interfering MIMO-MAC," *Int. Symp. on Wireless Commun. Systems (ISWCS)*, pp. 566–570, Aug. 2012.

[23] H. Gao, T. Lv, D. Fang, S. Yang, and C. Yuen, "Limited feedback-based interference alignment for interfering multi-access channels," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 540–543, April 2014.

[24] P. Cao, A. Zappone, and E. A. Jorswieck, "Grouping-based interference alignment with ia-cell assignment in multi-cell mimo mac under limited feedback," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1336–1351, March 2016.

[25] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[26] C. Yetis, T. Gou, S. Jafar, and A. Kayran, "On feasibility of interference alignment in mimo interference networks," *IEEE Trans. Signal Process.*, vol. 58, no. 9, p. 4771âĂŞ4782, Sep. 2010.

[27] S. Sesia, I. Toufik, and M. Baker, "LTE: The UMTS long term evolution: From theory to practice," *2nd Edition, Wiley*, Aug. 2011.

[28] R. Shafin, L. Liu, J. Zhang, and Y. C. Wu, "DoA estimation and capacity analysis for 3D millimeter wave massive MIMO/FD-MIMO OFDM systems," *accepted to be published in IEEE Trans. Wireless Commun.*, July 2016.

[29] L. Liu, R. Chen, S. Geirhofer, K. Sayana, Z. Shi, and Y. Zhou, "Downlink MIMO in LTE-Advanced: SU-MIMO vs. MU-MIMO," *IEEE Commun. Mag.*, pp. 140–147, Feb. 2012.

[30] C. Zhang, W. Xu, and M. Chen, "Robust MMSE beamforming for multiuser MISO systems with limited feedback," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 588–591, July 2009.

[31] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "A distributed numerical approach to interference alignment and applications to wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3309–3322, June 2011.

[32] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification, 2nd edition," *Wiley*, Nov. 2000.

[33] R. E. Prieto, "A general solution to the maximization of the multidimensional generalized rayleigh quotient used in linear discriminant analysis for signal classification," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 6, pp. VI 157–160, April 2003.

[34] P. Marsch and G. Fettweis, *Coordinated multi-point in mobile communications*. London: Cambridge University Press, 2011.

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

[36] L. Liu, J. C. Zhang, J. C. Yu, and J. Lee, "Inter-cell interference coordination through limited feedback," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, pp. 1–7, 2010.

[37] L. Liu, Y. H. Nam, and J. C. Zhang, "Proportional fair scheduling for multi-cell multi-user MIMO systems," *44th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, March 2010.

[38] D. Gesbert, S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: a new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[39] E. Bjornson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, p. 113âĂŞ381, 2012.

[40] "3GPP TR 36.819 v11.1.0 coordinated multi-point operation for LTE physical layer aspects," Dec. 2011.

[41] "3GPP TR 36.814 v9.0.0 further advancements for E-UTRA physical layer aspects," Mar. 2010.

[42] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[43] L. Liu, J. C. Zhang, Y. Yi, H. Li, and J. Zhang, "Combating interference: MU-MIMO, CoMP, and HetNet," *Journal of Comm. Academy Publisher*, vol. 7, no. 9, pp. 646–655, Sept. 2012.

[44] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-Advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[45] Y. Nam, L. Liu, and J. Zhang, "Cooperative communications for LTE-Advanced âĂŞ relay and CoMP," *Int. J. on Commun. Syst. (Wiley)*, vol. 27, no. 10, pp. 1616–1625, Oct. 2014.

[46] "3GPP TS 36.420 v8.1.0 E-UTRAN X2 general aspects and principles," Dec. 2008.

[47] P. Baracca, F. Boccardi, and N. Benvenuto, "A dynamic clustering algorithm for downlink CoMP systems with multiple antenna UEs," *EURASIP Journal on Wireless Communications and Networking*, vol. 1, pp. 1–14, 2014.

[48] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.

[49] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, June 2013.

[50] S. Shi, M. Schubert, and H. Boche, "Rate optimization for multiuser MIMO systems with linear precoding," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 4020–4030, Aug. 2008.

[51] ——, "Downlink MMSE transceiver optimization for multiuser MIMO systems: Duality and sum-MSE minimization," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5436–5446, Aug. 2007.

[52] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[53] A. Gjendemsj, D. Gesbert, G. E. Oien, and S. G. Kiani, "Binary power control for sum rate maximization over multiple interfering links," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3164–3173, Aug. 2008.

[54] S. Ye and R. S. Blum, "Optimized signaling for MIMO interference systems with feedback," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2839–2848, Nov. 2003.

[55] Q. Shi, M. Razaviyayn, Z. Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[56] K. Kwak, H. Lee, H. W. Je, J. Hong, and S. Choi, "Adaptive and distributed CoMP scheduling in LTE-Advanced systems," *IEEE Vehicular Technology Conference*, pp. 1–5, Sept. 2013.

[57] H. W. Je, H. Lee, K. Kwak, S. Choi, J. Hong, and B. Clerckx, "Long-term channel information-based CoMP beamforming in LTE-Advanced systems," *IEEE Global Telecommunications Conference*, pp. 1–6, Dec. 2011.

[58] B. O. Lee, H. Je, O. S. Shin, and K. B. Lee, "A novel uplink MIMO transmission scheme in a multicell environment," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 4981–4987, Oct. 2009.

[59] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, July 2004.

[60] L. Liu, R. Chen, S. Geirhofer, K. Sayana, Z. Shi, and Y. Zhou, "Downlink MIMO in LTE-advanced: SU-MIMO vs. MU-MIMO," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 140–147, Feb. 2012.

[61] Z. Q. Luo and S. Zhang, "Spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–72, Feb. 2008.

[62] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, , and J. S. Pang, "Decomposition by partial linearization: parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb. 2014.

[63] A. Hjorungnes, *Complex-Valued Matrix Derivatives With Applications in Signal Processing and Communications*. London: Cambridge University Press, 2011.

[64] C. Shi, R. A. Berry, and M. L. Honig, "Monotonic convergence of distributed interference pricing in wireless networks," *In proceedings of the 2009 IEEE International conference on Symposium on Information Theory (ISIT)*, vol. 3, pp. 1619–1623, Jun. 2009.

[65] A. Molisch, *Wireless Communications*, 2nd ed. Wiley, 2010.

[66] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks: A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, March 2015.

[67] "C-RAN: the road towards green RAN," *China Mobile,White paper, ver. 2.5*, pp. 405–426, Oct. 2011.

[68] T. X. Tran and D. Pompili, "Dynamic radio cooperation for downlink cloud-rans with computing resource sharing," *in Proc. IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, 2017.

[69] ——, "Dynamic radio cooperation for user-centric cloud-ran with computing resource sharing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2379–2393, Apr. 2017.

[70] "Cloud radio access network (c-ran): a primer."

[71] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 1077–1081, July 2013.

[72] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," *11th International Symposium on Wireless Communications Systems (ISWCS)*, pp. 1–5, 2014.

[73] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *http://arxiv.org/pdf/1602.00173*, Jan. 2016.

[74] A. Liu and V. K. N. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[75] T. X. Tran and D. Pompili, "Octopus: A cooperative hierarchical caching strategy for cloud radio access networks," *in Proc. IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, 2016.

[76] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili, "Collaborative multi-bitrate video caching and processing in mobile-edge computing networks," *in Proc. IEEE Conference on Wireless On-demand Network Systems and Services (WONS)*, 2017.

[77] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, July 2014.

[78] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[79] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

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

[81] T. X. Tran, D. V. Le, G. Yue, and D. Pompili, "Cooperative Hierarchical Caching and Request Scheduling in a Cloud Radio Access Network," *IEEE Transactions on Mobile Computing*, vol. PP, pp. 1–15, 2018.

[82] T. X. Tran, F. Kazemi, E. Karimi, and D. Pompili, "Mobee: Mobility-Aware Energy-Efficient Coded Caching in Cloud Radio Access Networks," in *IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2017, pp. 461–465.

[83] T. X. Tran and D. Pompili, "Adaptive Bitrate Video Caching and Processing in Mobile-Edge Computing Networks," *IEEE Transactions on Mobile Computing, under revision*, Mar. 2018.

[84] T. X. Tran, A. Younis, and D. Pompili, "Understanding the Computational Requirements of Virtualized Baseband Units using a Programmable Cloud Radio Access Network Testbed," in *IEEE International Conference on Autonomic Computing (ICAC)*, 2017, pp. 221–226.

[85] T. X. Tran and D. Pompili, "Joint task offloading and resource allocation for multi-server mobile-edge computing networks," *IEEE Transactions on Vehicular Technology, under revision*, 2017.

[86] "(mobile edge computing: Recent efforts and five key research directions."

[87] J. Zhao, T. Q. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, p. 2762âĂŞ2775, Jun. 2013.

[88] B. Dai and W. Yu, "Sparse beamforming for limited-backhaul network MIMO system via reweighted power minimization," *in Proc. IEEE Global Communications Conference (GLOBECOM)*, Jun. 2014.

[89] F. Zhuang and V. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, p. 684âĂŞ693, Feb. 2014.

[90] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," *IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communication (PIMRC)*, pp. 1370–1374, Sept. 2014.

[91] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *Special Section in IEEE Access:Recent Advances in Cloud Radio Access Networks*, vol. 2, pp. 1326–1339, Oct. 2014.

[92] H. Zhou, M. Tao, E. Chen, and W. Yu, "Content-centric multicast beamforming in cache-enabled cloud radio access networks," *in Proc. IEEE Global Communications Conference (GLOBECOM)*, Feb. 2016.

[93] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[94] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *IEEE INFOCOM*, p. 1107âĂŞ1115, Mar. 2012.

[95] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, p. 8402âĂŞ8413, Dec. 2013.

[96] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, p. 3553âĂŞ3568, Oct. 2015.

[97] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[98] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100–1113, Jun. 2014.

[99] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 37–42, May 2014.

[100] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, p. 3665âĂŞ3677, Oct. 2014.

[101] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," *in Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, p. 1078âĂŞ1086, Apr. 2014.

[102] K. Naveen, L. Massoulie, E. Baccelli, A. C. Viana, and D. Towsley, "On the interaction between content caching and request assignment in cellular cache networks," *in Proc. 5th Workshop Things Cellular, Oper., Appl. Challenges (AllThingsCellular), New York, NY, USA*, p. 37âĂŞ42, 2015.

[103] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," *in Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, p. 936âĂŞ944, Apr. 2015.

[104] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, p. 2275âĂŞ2284, Aug. 2016.

[105] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *Special Section in IEEE Access: Wireless Cachiing Technique for 5G*, vol. 4, pp. 8625–8633, Dec. 2016.

[106] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2016.

[107] S. Mosleh, L. Liu, J. D. Ashdown, E. Perrins, and K. Turck, "Content-Based User Association and MIMO Operation over Cached Cloud-RAN Networks," *IEEE Transactions on Communications (under revision)*.

[108] E. Bastug, J. L. Guenego, and M. Debbah, "Proactive small cell networks," *20th International Conference on Telecommunications (ICT)*, pp. 1–5, May 2013.

[109] C. T. K. Ng and H. Huang, "Linear precoding in cooperative mimo cellular networks with limited coordination clusters," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Dec. 2010.

[110] S. Mosleh, J. D. Ashdown, J. D. Matyjas, M. J. Medley, J. Zhang, and L. Liu, "Interference alignment for downlink Multi-Cell LTE-Advanced systems with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8107–8121, Dec. 2016.

[111] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," *IEEE Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOMâĂŹ99)*, vol. 1, p. 126âĂŞ134, 1999.

[112] M. Hefeeda and O. Saleh, "Traffic modeling and proportional partial caching for peer-to-peer systems," *IEEE/ACM Transactions on Networking*, vol. 16, no. 6, p. 1447âĂŞ1460, Dec. 2008.

[113] J. Kleinberg and E. Tardos, "Algorithm design," *Boston, MA: Addison-Wesley*, 2012.

[114] Q. Shi, M. Razaviyayn, Z. Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[115] T. Ma, Q. Shi, and E. Song, "Qos-constrained weighted sum-rate maximization in multi-cell multi-user MIMO systems: An ADMM approach," *35th Chinese Control Conference (CCC)*, July 2016.

[116] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[117] M. R. Garey and D. S. Johnson, *Computers and intractability: A Guide to the Theory of NP-completeness*, 1990.

[118] G. L. Nemhauser, M. W. P. Savelsbergh, and G. S. Sigismondi, "MINTO, a Mixed INTeger Optimizer," *Research Letters, Elsevier*, vol. 15, pp. 47–58, 1994.

[119] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, p. 641âĂŞ656, Feb. 2014.

[120] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[121] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, Sep. 2012.

[122] B. K. L. E. G. L. T. L. M. O. E. F. Rusek, D. Persson and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[123] Greentouch, "accessed on apr. 5, 2016. [online] available: http://www.greentouch.org/."

[124] 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, no. 3, pp. 1834–1350, Mar. 2017.

[125] ——, "Cell-free massive mimo: Uniformly great service for everyone," *in Proc. 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pp. 201–205, Jun. 2015.

[126] 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, no. 7, pp. 4445–4459, Jul. 2017.

[127] S. Mosleh, H. Almosa, E. Perrins, and L. Liu, "Downlink Resource Allocation in Cell-Free Massive MIMO Systems," in *International Conference on Computing, Networking and Communications (ICNC 2019)*, 2019.

[128] S. Buzzi and C. DAndrea, "Cell-free massive mimo: User-centric approach," *IEEE Wirless Commun. Letters*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[129] A. Tang, J. Sun, and K. Gong, "Mobile propagation loss with a low base station antenna for NLOS street microcells in urban area," *IEEE Veh. Technol. Conf.*, pp. 333–336, May. 2001.

[130] C. T. Secretariat, "Digital mobile radio toward future generation systems: Cost 231 final report," *Brussels: European COST Action 231*, 1999.

[131] E. K. T. Z. Wang and A. R. Nix, "Joint shadowing process in urban peer-to-peer radio channels," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 52–64, Jan. 2008.

[132] H. Almosa, R. Shafin, S. Mosleh, Z. Zhou, Y. Li, J. C. Zhang, and L. Liu, "Downlink channel estimation and precoding for FDD 3D Massive MIMO/FD-MIMO systems," in *26th Wireless and Optical Communication Conference (WOCC)*, 2017.

[133] H. Almosa, S. Mosleh, E. Perrins, and L. Liu, "Downlink Channel Estimation with Limited Feedback for FDD Multi-User Massive MIMO with Spatial Channel Correlation," in *IEEE International Conference on Communications (ICC)*, 2018.

[134] S. M. Kay, "Fundamentals of statistical signal processing, volume 1: Estimation theory," *Englewood Cliffs, NJ, USA: Prentice-Hall*, 1993.

[135] E. G. L. E. Bjornson and M. Debbah, "Massive mimo for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

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

[137] A. G. S. Cui and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.

[138] E. L. D. Ng and R. Schober, "Energy-efficient resource allocation in ofdma system with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.

[139] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[140] D. Hoang and R. A. Iltis, "Non-cooperative eigencoding for MIMO Ad-hoc networks," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 865 – 869, Feb. 2008.

[141] B. Kulis, S. Sra, and I. Dhillon, "Convex perturbations for scalable semidefinite programming," *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 296–303, 2009.

[142] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento, "A parallel stochastic approximation method for nonconvex multi-agent optimization problems," *IEEE Trans. Signal Process.*, 2014, submitted, online available: http://arxiv.org/abs/1410.5076.

[143] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. London: Cambridge University Press, 2004.

Appendix A

Appendix

A.0.1 Proof of Theorem 1

As mentioned earlier, [31] shows that the condition in (9) is automatically satisfied almost surely if $\mathbf{Q}_i$ and $\mathbf{T}_j$ can be found to satisfy condition (2.8). Thus, solving the linear IA feasibility problem amounts to solve the polynomial equations in (2.8). Our approach is to view the alignment problem as a system of multivariate polynomial equations, and determine its solvability by comparing the number of equations and variables. It is clear that the system is solvable if and only if the number of independent equation does not exceed the number of variables. To do so, we rewrite the conditions in (2.8) as the following:

$$\mathbf{q}_{i_n}^H \sum_{\substack{j=1 \\ j\neq i}}^{G} \mathbf{H}_{i_k,j}\mathbf{t}_{j_m} = 0,\ \forall i_k \in \mathscr{I} \tag{A.1}$$

where $\mathbf{t}_{j_m}$ and $\mathbf{q}_{i_n}$, $\forall m \in \{1,2,\cdots,S\}$ and $\forall n \in \{1,2,\cdots,N_r^{(i_k)} - S\}$, are the cell-specific transmit and receive beamforming vectors. Hence, the number of equations is directly obtained from (A.1) and it is equal to $N_{\text{eq}} = \sum_{i=1}^{G}\sum_{k=1}^{I_i}(N_r^{(i_k)} - S)S$. However, calculating the number of variables is less straightforward, since we have to exclude inessential variables that are not involved with IA. Inspired by the analysis in [26], the number of variables to be designed for the transmit beamforming $\mathbf{T}_i$ is $(N_t^{(i)} - S)S$. Likewise, the actual number of variables to be designed for the receive beam-

forming matrix $\mathbf{Q}_i$ is $(N_r^{(i_k)} - S)S$. As a result, the total number of variables of interests become

$$\sum_{i=1}^{G} \left((N_t^{(i)} - S)S + \sum_{k=1}^{I_i} (N_r^{(i_k)} - S)S \right).$$

Therefore, when $\sum_{i=1}^{G} N_t^{(i)} > GS$, IA becomes feasible.

A.0.2 Proof of Theorem 2

To solve a closed-form expression for $\hat{\mathbf{T}}_i$, we will first calculate $\mathbb{E}_H\{\text{SLNR}_{i_k}\}$ using the Jensen's inequality:

$$\begin{aligned} &\mathbb{E}_H\{\text{SLNR}_{i_k}\} = \\ &\mathbb{E}_H\left\{ \frac{\text{Tr}\left(\hat{\mathbf{T}}_i^H \mathbf{H}_{i_k,i}^H \mathbf{H}_{i_k,i} \hat{\mathbf{T}}_i \right)}{\text{Tr}\left(\hat{\mathbf{T}}_i^H (\frac{\sigma_{i_k}^2}{\rho_{i_k} S} \mathbf{I}_S + \sum_{j=1, j\neq i}^{G} \sum_{m=1}^{I_j} \mathbf{H}_{j_m,i}^H \mathbf{H}_{j_m,i}) \hat{\mathbf{T}}_i \right)} \right\} \\ &\geq \frac{\text{Tr}\left(\hat{\mathbf{T}}_i^H \mathbb{E}_H\left\{ \mathbf{H}_{i_k,i}^H \mathbf{H}_{i_k,i} \right\} \hat{\mathbf{T}}_i \right)}{\text{Tr}\left(\hat{\mathbf{T}}_i^H (\frac{\sigma_{i_k}^2}{\rho_{i_k} S} \mathbf{I}_S + \sum_{j=1, j\neq i}^{G} \sum_{m=1}^{I_j} \mathbb{E}_H\left\{ \mathbf{H}_{j_m,i}^H \mathbf{H}_{j_m,i} \right\}) \hat{\mathbf{T}}_i \right)}. \end{aligned} \tag{A.2}$$

By invoking the SVD of $\mathbf{H}_{i_k,i}$, we have

$$\mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{i_k,i}^H\mathbf{H}_{i_k,i}\right\} = \mathbb{E}_{\mathbf{H}}\left\{\mathbf{V}_{\mathbf{H}_{i_k,i}}\blacksquare_{\mathbf{H}_{i_k,i}}^H\blacksquare_{\mathbf{H}_{i_k,i}}\mathbf{V}_{\mathbf{H}_{i_k,i}}^H\right\} \tag{A.3}$$

$$\overset{(a)}{=} \mathbb{E}_{\mathbf{A},\mathbf{B},\mathbf{R}}\left\{(\mathbf{A}_{\mathbf{H}_{i_k,i}}\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,i}} + \mathbf{B}_{\mathbf{H}_{i_k,i}}\mathbf{R}_{\mathbf{H}_{i_k,i}})^H \right.$$

$$\left.(\mathbf{A}_{\mathbf{H}_{i_k,i}}\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,i}} + \mathbf{B}_{\mathbf{H}_{i_k,i}}\mathbf{R}_{\mathbf{H}_{i_k,i}})\right\}$$

$$\overset{(b)}{=} \widehat{\mathbf{V}}_{\mathbf{H}_{i_k,i}}\mathbb{E}_{\mathbf{A}}\left\{\mathbf{A}_{\mathbf{H}_{i_k,i}}^H\mathbf{A}_{\mathbf{H}_{i_k,i}}\right\}\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,i}}^H$$

$$+ \mathbb{E}_{\mathbf{B},\mathbf{R}}\left\{\mathbf{R}_{\mathbf{H}_{i_k,i}}^H\mathbf{B}_{\mathbf{H}_{i_k,i}}^H\mathbf{B}_{\mathbf{H}_{i_k,i}}\mathbf{R}_{\mathbf{H}_{i_k,i}}\right\}$$

$$\overset{(c)}{=} N_t(1-\delta)\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,i}}^H\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,i}} + \delta N_t\mathbf{I}_{N_t}.$$

Since $\sum_{k=1}^{S}\mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{i_k,i}^H\mathbf{H}_{i_k,i}\right\}$ is Hermitian and positive semidefinite and $(\frac{\sigma_{i_k}^2}{\rho_{i_k}S}\mathbf{I}_S + \sum_{j=1,j\neq i}^{G}\sum_{m=1}^{I_j}\mathbb{E}_{\mathbf{H}}\{\mathbf{H}_{j_m,i}^H\mathbf{H}_{j_m,i}$ is Hermitian and positive definite, by generalized eigenvalue decomposition, there exists an invertible matrix $\mathbf{J}_i \in \mathbb{C}^{N_t\times N_t}$ such that

$$\mathbf{J}_i^H\mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{i_k,i}^H\mathbf{H}_{i_k,i}\right\}\mathbf{J}_i = \blacksquare_i = \mathrm{diag}(\delta_{i_1},\ldots,\delta_{i_{N_t}}) \tag{A.4}$$

$$\mathbf{J}_i^H(\frac{\sigma_{i_k}^2}{\rho_{i_k}S}\mathbf{I}_S + \sum_{j=1,j\neq i}^{G}\sum_{m=1}^{K}\mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{j_m,i}^H\mathbf{H}_{j_m,i}\right\})\mathbf{J}_i = \mathbf{I}_{N_t}$$

with $\delta_{i_1} \geq \delta_{i_2} \geq \ldots \geq \delta_{i_{N_t}} \geq 0$. Here, the columns of $\mathbf{J}_i$ and the diagonal entries of $\blacksquare_i$ are the generalized eigenvectors and eigenvalues of the pair $(\sum_{k=1}^{S}\mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{i_k,i}^H\mathbf{H}_{i_k,i}\right\}, (\frac{\sigma_{i_k}^2}{\rho_{i_k}S}\mathbf{I}_S + \sum_{j=1,j\neq i}^{G}\sum_{m=1}^{I_j}\mathbb{E}_{\mathbf{H}}\left\{\mathbf{H}_{j_m,i}^H\mathbf{H}_{j_m,i}\right\}))$, respectively. Then, the optimal precoder maximizing the lower bound of the objective function in Eq. (2.21) can be obtained by extracting the leading S columns of $\mathbf{J}_i$ as given by Theorem 2.

A.0.3 Proof of Theorem 3

For given $\hat{\mathbf{T}}_j$, the expectation of $\mathscr{L}_{i_k}(\mathbf{W}_i,\hat{\mathbf{T}}_j)$ can be expressed as (A.5), where (a) follows readily from applying the definition of residual interference at user i_k in Eq. (2.17); (b) comes from substituting (2.18); (c) holds since $\mathbf{R}_{\mathbf{H}_{i_k,j}}$ is independent of $\mathbf{A}_{\mathbf{H}_{i_k,j}}$ and $\mathbf{B}_{\mathbf{H}_{i_k,j}}$, and $\mathbb{E}\{\mathbf{R}_{\mathbf{H}_{i_k,j}}\} = \mathbf{0}$.

$$\mathbb{E}_{\mathbf{H}}\{\mathscr{L}_{i_k}(\mathbf{W}_i,\hat{\mathbf{T}}_j)\} \overset{(a)}{=} \mathbb{E}_{\mathbf{H}}\left\{\mathrm{Tr}\left\{\mathbf{W}_i^H\left(\sum_{j=1,j\neq i}^{G}\Lambda_{\mathbf{H}_{i_k,j}}\mathbf{V}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{V}_{\mathbf{H}_{i_k,j}}\Lambda^H_{\mathbf{H}_{i_k,j}}\right)\mathbf{W}_i\right\}\right\} \tag{A.5}$$

$$\overset{(b)}{=} \mathbb{E}_{\mathbf{A},\mathbf{B},\mathbf{R}}\left\{\mathrm{Tr}\left\{\mathbf{W}_i^H\left(\sum_{j=1,j\neq i}^{G}(\mathbf{A}_{\mathbf{H}_{i_k,j}}\widehat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}+\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}})\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H(\mathbf{A}_{\mathbf{H}_{i_k,j}}\widehat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}+\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}})^H\right)\mathbf{W}_i\right\}\right\}$$

$$\overset{(c)}{=} \mathrm{Tr}\left\{\mathbf{W}_i^H\left(\sum_{j=1,j\neq i}^{G}\mathbb{E}_{\mathbf{A},\mathbf{B},\mathbf{R}}\left\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\widehat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}+\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\mathbf{B}^H_{\mathbf{H}_{i_k,j}}\right\}\right)\mathbf{W}_i\right\}$$

$$\overset{(d)}{=} \mathrm{Tr}\left\{\mathbf{W}_i^H\sum_{j=1,j\neq i}^{G}\left(\mathbb{E}_{\mathbf{A}}\left\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\widehat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\widehat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\right\}+\mathbb{E}_{\mathbf{B},\mathbf{R}}\left\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\mathbf{B}^H_{\mathbf{H}_{i_k,j}}\right\}\right)\mathbf{W}_i\right\}$$

Moreover, (d) holds due to the fact that the expectation of the first and second terms in (c) have nothing to do with $\mathbf{B}_{\mathbf{H}_{i_k,j}},\mathbf{R}_{\mathbf{H}_{i_k,j}}$ matrices and $\mathbf{A}_{\mathbf{H}_{i_k,j}}$ matrix, respectively.

We start with the calculation of the first term in Eq. (A.5). Note that the expected value of a random matrix is a matrix whose elements are the expected values of the individual random variables that are the elements of the random matrix. Let $\mathbf{X}_{i_k,j}\triangleq\widehat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j$ and $\mathbf{e}_s$ denote the vectors having "1" at the s-th position and zero elsewhere which the dimension of $\mathbf{e}_s$ will be clear from the context. Moreover, we define a $N_r\times N_t$ matrix $\mathbf{L}_s$, such that $\mathbf{L}_s\triangleq[\mathbf{I}_{N_r\times N_r}\ \mathbf{0}]$. We note that for $1\leq r,t\leq N_r$:

$$
\begin{aligned}
&\left(\mathbb{E}_{\mathbf{A}}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\}\right)_{rt} \\
&=\left(\mathbb{E}_{\mathbf{A}}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\mathbf{X}_{i_k,j}\mathbf{X}^H_{i_k,j}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\}\right)_{rt} \\
&=\mathbb{E}_{\mathbf{A}}\{\mathbf{e}_r^T\mathbf{A}_{\mathbf{H}_{i_k,j}}\mathbf{X}_{i_k,j}\mathbf{X}^H_{i_k,j}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\mathbf{e}_t\} \\
&=\mathbb{E}_{\mathbf{A}}\{\sum_{k=1}^{S}\mathbf{e}_r^T\mathbf{A}_{\mathbf{H}_{i_k,j}}\mathbf{X}_{i_k,j}\mathbf{e}_k\mathbf{e}_k^T\mathbf{X}^H_{i_k,j}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\mathbf{e}_t\} \\
&=\sum_{k=1}^{S}\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbb{E}_{\mathbf{A}}\{\mathbf{e}_r^T\mathbf{A}_{\mathbf{H}_{i_k,j}}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{X}_{i_k,j}\mathbf{e}_k\mathbf{e}_k^T\mathbf{X}^H_{i_k,j}\mathbf{e}_m\mathbf{e}_m^T\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\mathbf{e}_t\} \\
&=\mathbf{e}_r^T\left(\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbf{e}_\ell^T\mathbf{X}_{i_k,j}\mathbf{X}^H_{i_k,j}\mathbf{e}_m\mathbb{E}_{\mathbf{A}}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\mathbf{e}_\ell\mathbf{e}_m^T\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\}\right)\mathbf{e}_t \\
&=\mathbf{e}_r^T\Bigg(\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbf{e}_\ell^T\mathbf{X}_{i_k,j}\mathbf{X}^H_{i_k,j}\mathbf{e}_m \\
&\qquad\qquad\mathbb{E}_{\mathbf{A}}\{\hat{\lambda}^{(\ell)}_{i_k,j}\sqrt{1-z^{(\ell)}_{i_k,j}}\hat{\lambda}^{(m)}_{i_k,j}\sqrt{1-z^{(m)}_{i_k,j}}\mathbf{D}_{\ell m}\}\Bigg)\mathbf{e}_t,
\end{aligned}
\tag{A.6}
$$

where $\mathbf{D}_{\ell m}$ denotes an $N_r \times N_r$ matrix with one in the ℓm-th position and zeros elsewhere. For $N_r > N_t$, the above expectation can be expressed as the following expression

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\} \\
&=\begin{pmatrix}\mathbf{X}_{i_k,j}\odot\mathbf{E} & \mathbf{0}_{N_t\times(N_r-N_t)}\\ \mathbf{0}_{(N_r-N_t)\times N_t} & \mathbf{0}_{(N_r-N_t)\times(N_r-N_t)}\end{pmatrix},
\end{aligned}
\tag{A.7}
$$

where $\mathbf{E}$ is an $\tilde{d}\times\tilde{d}$ matrix with $(\mathbf{E})_{\ell m}=\hat{\lambda}^{(\ell)}_{i_k,j}\hat{\lambda}^{(m)}_{i_k,j}\left(1-\frac{1}{2}\delta\frac{N_t-1}{N_t}-\frac{1}{8}\delta^2\frac{N_t-1}{N_t+1}-\frac{1}{16}\delta^3\frac{N_t-1}{N_t+2}\right)^2$ and $\odot$ is the Hadamard product. For $N_r < N_t$, we partition $\mathbf{X}_{i_k,j}$ as follows $\mathbf{X}_{i_k,j}=\begin{pmatrix}(\mathbf{X}_{i_k,j})_{11} & (\mathbf{X}_{i_k,j})_{12}\\ (\mathbf{X}_{i_k,j})_{21} & (\mathbf{X}_{i_k,j})_{22}\end{pmatrix}$ where $\left(\mathbf{X}_{i_k,j}\right)_{11}$ is a $\tilde{d}\times\tilde{d}$ matrix and can be expressed as $\left(\mathbf{X}_{i_k,j}\right)_{11}=\mathbf{L}_s\mathbf{X}_{i_k,j}\mathbf{L}_s^T$. Therefore, we have

$$\mathbb{E}_{\mathbf{A}}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\} = \left(\mathbf{L}_s\mathbf{X}_{i_k,j}\mathbf{L}_s^T\right)\odot\mathbf{E}, \tag{A.8}$$

and for $N_r = N_t$, the above expression simplifies to

$$\mathbb{E}_{\mathbf{A}}\{\mathbf{A}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{V}}^H_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\hat{\mathbf{V}}_{\mathbf{H}_{i_k,j}}\mathbf{A}^H_{\mathbf{H}_{i_k,j}}\} = \mathbf{X}_{i_k,j}\odot\mathbf{E}. \tag{A.9}$$

The second term in Eq. (A.5) can be calculated as

$$\begin{aligned}&\mathbb{E}_{\mathbf{B},\mathbf{R}}\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\mathbf{B}^H_{\mathbf{H}_{i_k,j}}\} = \\ &\mathbb{E}_{\mathbf{B}}\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbb{E}_{\mathbf{R}}\{\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\}\mathbf{B}^H_{\mathbf{H}_{i_k,j}}\},\end{aligned} \tag{A.10}$$

where equality holds due to the law of total expectation. At first, we calculate the inner expectation as follows

$$\begin{aligned}&\left(\mathbb{E}_{\mathbf{R}}\{\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\}\right)_{rt} \\ &= \mathbb{E}_{\mathbf{R}}\{\mathbf{e}_r^T\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\mathbf{e}_t\} \\ &= \mathbb{E}_{\mathbf{R}}\{\sum_{k=1}^{S}\mathbf{e}_r^T\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\mathbf{e}_k\mathbf{e}_k^T\hat{\mathbf{T}}_j^H\mathbf{R}^H_{\mathbf{H}_{i_k,j}}\mathbf{e}_t\}\end{aligned} \tag{A.11}$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{R}}\{\sum_{k=1}^{S}\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbf{e}_r^T\mathbf{R}_{\mathbf{H}_{i_k,j}}\mathbf{e}_\ell\mathbf{e}_\ell^T\hat{\mathbf{T}}_j\mathbf{e}_k\mathbf{e}_k^T\hat{\mathbf{T}}_j^H\mathbf{e}_m\mathbf{e}_m^T\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\mathbf{e}_t\} \\
&= \mathbf{e}_r^T\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbf{e}_\ell^T\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{e}_m\mathbb{E}_{\mathbf{R}}\{\mathbf{R}_{\mathbf{H}_{i_k,j}}\mathbf{e}_\ell\mathbf{e}_m^T\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\}\mathbf{e}_t \\
&= \mathbf{e}_r^T\frac{1}{N_t-1}\sum_{m=1}^{N_t}\mathbf{e}_m^T\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{e}_m\left(\mathbf{I}_{N_t}-\hat{\mathbf{v}}_{i_k,j}^{(m)}\hat{\mathbf{v}}_{i_k,j}^{(m)H}\right)\mathbf{e}_t,
\end{aligned}
\tag{A.12}
$$

where the last equality comes from the fact that for every $\ell, m \in \{1,\ldots,N_t\}$ and $\ell \neq m$, $\mathbf{r}_{i_k,j}^{(\ell)} \perp\!\!\!\perp \mathbf{r}_{i_k,j}^{(m)}$, $\mathbb{E}\{\mathbf{r}_{i_k,j}^{(\ell)}\} = \mathbf{0}$, $\mathbb{E}\{\mathbf{r}_{i_k,j}^{(\ell)}\mathbf{r}_{i_k,j}^{(\ell)H}\} = 1/(N_t-1)(\mathbf{I}_{N_t}-\hat{\mathbf{v}}_{i_k,j}^{(\ell)}\hat{\mathbf{v}}_{i_k,j}^{(\ell)H})$. Hence,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{R}}\{\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\} \\
&= \frac{1}{N_t-1}\sum_{m=1}^{N_t}\mathbf{e}_m^T\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{e}_m\left(\mathbf{I}_{N_t}-\hat{\mathbf{v}}_{i_k,j}^{(m)}\hat{\mathbf{v}}_{i_k,j}^{(m)H}\right) \triangleq \mathbf{Y}_{i_k,j}.
\end{aligned}
\tag{A.13}
$$

Therefore,

$$
\begin{aligned}
&\left(\mathbb{E}_{\mathbf{B}}\left\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{Y}_{i_k,j}\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\right\}\right)_{rt} = \mathbf{e}_r^T\mathbb{E}_{\mathbf{B}}\left\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{Y}_{i_k,j}\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\right\}\mathbf{e}_t \\
&= \mathbb{E}_{\mathbf{B}}\left\{\mathbf{e}_r^T\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{Y}_{i_k,j}\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\mathbf{e}_t\right\} \\
&= \mathbb{E}_{\mathbf{B}}\left\{\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbf{e}_r^T\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{Y}_{i_k,j}\mathbf{e}_m\mathbf{e}_m^T\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\mathbf{e}_t\right\} \\
&= \mathbf{e}_r^T\left(\sum_{\ell=1}^{N_t}\sum_{m=1}^{N_t}\mathbf{e}_\ell^T\mathbf{Y}_{i_k,j}\mathbf{e}_m\mathbb{E}_{\mathbf{B}}\left\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{e}_\ell\mathbf{e}_m^T\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\right\}\right)\mathbf{e}_t.
\end{aligned}
\tag{A.14}
$$

Therefore, for $N_r > N_t$, we have

$$
\mathbb{E}_{\mathbf{B},\mathbf{R}}\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\} = \begin{bmatrix}\mathbf{Y}_{i_k,j}\odot\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}.
$$

By partitioning $\mathbf{Y}_{i_k,j}$ as $\begin{bmatrix}(\mathbf{Y}_{i_k,j})_{11} & (\mathbf{Y}_{i_k,j})_{12}\\ (\mathbf{Y}_{i_k,j})_{21} & (\mathbf{Y}_{i_k,j})_{22}\end{bmatrix}$ where $\left(\mathbf{Y}_{i_k,j}\right)_{11}$ is a $\tilde{d}\times\tilde{d}$ matrix and can be expressed as $\left(\mathbf{Y}_{i_k,j}\right)_{11} = \mathbf{L}_s\mathbf{Y}_{i_k,j}\mathbf{L}_s^T$, for $N_r < N_t$ we get $\mathbb{E}_{\mathbf{B},\mathbf{R}}\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\} = \left(\mathbf{L}_s\mathbf{Y}_{i_k,j}\mathbf{L}_s^T\right)\odot\mathbf{G}$, and for $N_r = N_t$, we obtain $\mathbb{E}_{\mathbf{B},\mathbf{R}}\{\mathbf{B}_{\mathbf{H}_{i_k,j}}\mathbf{R}_{\mathbf{H}_{i_k,j}}\hat{\mathbf{T}}_j\hat{\mathbf{T}}_j^H\mathbf{R}_{\mathbf{H}_{i_k,j}}^H\mathbf{B}_{\mathbf{H}_{i_k,j}}^H\} = \mathbf{Y}_{i_k,j}\odot\mathbf{G}$, where $\mathbf{G}$ is a $\tilde{d}\times\tilde{d}$

matrix which its ℓm-th component is equal to $(\mathbf{G})_{\ell m} = \hat{\lambda}_{i_k,j}^{\ell}\hat{\lambda}_{i_k,j}^{m}\left(1-\frac{N_t-1}{8N_t^2(1+N_t)}\right)^2$. Hence, $\mathbb{E}\{\tilde{\mathbf{R}}_{i_k}\}$ can be expressed as given by Theorem 3.

A.0.4 Proof of Theorem 4

To solve the optimization problem (2.32), at first we need to calculate the statistical expectations $\mathbb{E}\left\{\mathbf{h}_{i_k}^H\mathbf{h}_{i_k}\right\}$ for $i_k \in \mathscr{I}$. By substituting (2.30) into $\mathbf{h}_{i_k}^H\mathbf{h}_{i_k}$, we obtain

$$\begin{aligned}\mathbb{E}\left\{\mathbf{h}_{i_k}^H\mathbf{h}_{i_k}\right\} &= \mathbb{E}\left\{\sum_{\ell=1}^{d}\sum_{m=1}^{d} g_{i_k}^{(\ell)}\lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)*}\mathbf{e}_{i_k,i}^{(\ell)}g_{i_k}^{(m)*}\lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(m)}\mathbf{e}_{i_k,i}^{(m)H}\right\}\\ &= \sum_{\ell=1}^{d}\sum_{m=1}^{d} g_{i_k}^{(\ell)}g_{i_k}^{(m)*}\lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)*}\lambda_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(m)}\mathbb{E}\left\{\mathbf{e}_{i_k,i}^{(\ell)}\mathbf{e}_{i_k,i}^{(m)H}\right\},\end{aligned}$$

where $d \triangleq \min\{\mathrm{S},\mathrm{N_r}-\mathrm{S}\}$. Then, we calculate the expectation of $\mathbf{e}_{i_k,i}^{(\ell)}\mathbf{e}_{i_k,i}^{(m)H}$ which is given as follows:

$$\begin{aligned}\mathbb{E}\left\{\mathbf{e}_{i_k,i}^{(\ell)}\mathbf{e}_{i_k,i}^{(m)H}\right\} &\overset{\text{(a)}}{=} \mathbb{E}\Bigg\{\sqrt{z_{i_k,i}^{(\ell)}}\sqrt{z_{i_k,i}^{(m)}}\mathbf{r}_{i_k,i}^{(\ell)}\mathbf{r}_{i_k,i}^{(m)H} +\\ &\sqrt{z_{i_k,i}^{(m)}}\sqrt{1-z_{i_k,i}^{(\ell)}}\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\mathbf{r}_{i_k,i}^{(m)H} + \sqrt{z_{i_k,i}^{(\ell)}}\sqrt{1-z_{i_k,i}^{(m)}}\mathbf{r}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(m)H}\\ &+\sqrt{1-z_{i_k,i}^{(\ell)}}\sqrt{1-z_{i_k,i}^{(m)}}\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(m)H}\Bigg\}\\ &\overset{\text{(b)}}{=}\begin{cases}(1-\delta)\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(\ell)H}+\frac{\delta}{S}\mathbf{I}_S & ,\text{ if } \ell = m;\\ (1-\frac{\mathrm{S}-1}{2\mathrm{S}}\delta-\frac{\mathrm{S}-1}{8(\mathrm{S}+1)}\delta^2)^2\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(m)H} & ,\text{ if } \ell \neq m.\end{cases}\end{aligned}$$

where (a) comes from Eq. (2.7) and (b) holds due to the fact that $z_{i_k,i}^{(\ell)} \perp\!\!\!\perp \mathbf{r}_{i_k,i}^{(\ell)}$, $z_{i_k,i}^{(m)} \perp\!\!\!\perp \mathbf{r}_{i_k,i}^{(m)}$, $z_{i_k,i}^{(m)} \perp\!\!\!\perp \mathbf{r}_{i_k,i}^{(\ell)}$, $z_{i_k,i}^{(\ell)} \perp\!\!\!\perp \mathbf{r}_{i_k,i}^{(m)}$, $z_{i_k,i}^{(\ell)} \perp\!\!\!\perp z_{i_k,i}^{(m)}$, $\mathbf{r}_{i_k,i}^{(\ell)} \perp\!\!\!\perp \mathbf{r}_{i_k,i}^{(m)}$, $\mathbb{E}\{\mathbf{r}_{i_k,i}^{(\ell)}\} = \mathbb{E}\{\mathbf{r}_{i_k,i}^{(m)}\} = \mathbf{0}$, $\mathbb{E}\{\mathbf{r}_{i_k,i}^{(m)}\mathbf{r}_{i_k,i}^{(m)H}\} = 1/(S-1)(\mathbf{I}_S - \hat{\mathbf{e}}_{i_k,i}^{(m)}\hat{\mathbf{e}}_{i_k,i}^{(m)H})$, $\mathbb{E}\{\mathbf{r}_{i_k,i}^{(\ell)}\mathbf{r}_{i_k,i}^{(\ell)H}\} = 1/(S-1)(\mathbf{I}_S - \hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(\ell)H})$ and $\mathbb{E}\{z_{i_k,i}^{(\ell)}\} = \mathbb{E}\{z_{i_k,i}^{(m)}\} = \delta(\mathrm{S}-1)/\mathrm{S}$. Hence,

we have

$$\mathbb{E}\left\{\mathbf{h}_{i_k}^H\mathbf{h}_{i_k}\right\} = \sum_{\ell=1}^{d}|g_{i_k}^{(\ell)}|^2|\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)}|^2\left((1-\delta)\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(\ell)H}+\frac{\delta}{\mathrm{S}}\mathbf{I}_{\mathrm{S}}\right)+$$
$$\sum_{\ell=1}^{d}\sum_{\substack{m=1\\m\neq\ell}}^{d}g_{i_k}^{(\ell)}g_{i_k}^{(m)*}\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)*}\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(m)}(1-\frac{\mathrm{S}-1}{2\mathrm{S}}\delta-\frac{\mathrm{S}-1}{8(\mathrm{S}+1)}\delta^2)^2\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(m)H}. \tag{A.15}$$

By substituting the equation (A.15) into (2.32), the optimization problem in (2.32) is equivalent to

$$\underset{\|\mathbf{v}_{i_k}\|^2=1}{\arg\max}\ \frac{\mathbf{v}_{i_k}^H\left(\Psi_{i_k}+\sum_{\ell=1}^{d}\alpha_{i_k,\ell}\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(\ell)H}\right)\mathbf{v}_{i_k}}{\frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}+\mathbf{v}_{i_k}^H\left(\sum_{\substack{m=1\\m\neq k}}^{I_i}\left(\Psi_{i_m}+\sum_{\ell=1}^{d}\alpha_{i_m,\ell}\hat{\mathbf{e}}_{i_m,i}^{(\ell)}\hat{\mathbf{e}}_{i_m,i}^{(\ell)H}\right)\right)\mathbf{v}_{i_k}}, \tag{A.16}$$

where $\alpha_{i_k,\ell}\triangleq|g_{i_k}^{(\ell)}|^2|\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)}|^2(1-\delta)$, $i_k\in\mathscr{I}$, and

$$\Psi_{i_k}\triangleq\sum_{\ell=1}^{d}\sum_{\substack{m=1\\m\neq\ell}}^{d}g_{i_k}^{(\ell)}g_{i_k}^{(m)*}\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)*}\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(m)}(1-\frac{\mathrm{S}-1}{2\mathrm{S}}\delta-\frac{\mathrm{S}-1}{8(\mathrm{S}+1)}\delta^2)^2$$
$$\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(m)H}+\sum_{\ell=1}^{d}\frac{\delta}{\mathrm{S}}|g_{i_k}^{(\ell)}|^2|\hat{\lambda}_{\mathbf{H}_{i_k,i}^{\text{eff}}}^{(\ell)}|^2\mathbf{I}_{\mathrm{S}}. \tag{A.17}$$

Since the above optimization problem is a generalized Rayleigh quotient problem, for a given receive filters, the optimal closed-form solution for transmit beamforming vectors are available and can be expressed as follow

$$\mathbf{v}_{i_k}^{\text{opt}}=\nu_{\max}\left\{\left(\frac{\tilde{\sigma}_{i_k}^2}{\rho_{i_k}}\mathbf{I}_{\mathrm{S}}+\sum_{\substack{m=1\\m\neq k}}^{I_i}\left(\Psi_{i_m}+\sum_{\ell=1}^{d}\alpha_{i_m,\ell}\hat{\mathbf{e}}_{i_m,i}^{(\ell)}\hat{\mathbf{e}}_{i_m,i}^{(\ell)H}\right)\right)^{-1}\right.$$
$$\left.\left(\Psi_{i_k}+\sum_{\ell=1}^{d}\alpha_{i_k,\ell}\hat{\mathbf{e}}_{i_k,i}^{(\ell)}\hat{\mathbf{e}}_{i_k,i}^{(\ell)H}\right)\right\}. \tag{A.18}$$

A.0.5 Proof of Theorem 5

The partial Lagrangian of (3.14) is given by

$$\begin{aligned}\mathscr{L}_i(\mathbf{P}_i;\mu_i) = w_i \log_2 \det(\mathbf{I}_{N_r} + \mathbf{C}_i\mathbf{P}_i) \\ - \operatorname{Tr}\left((\mathbf{E}_i + \mu_i\mathbf{I})\mathbf{P}_i\right) + \mu_i p_{i,\max}\end{aligned} \tag{A.19}$$

where μ_i is the Lagrange multiplier corresponding to the power constraint. The dual function is

$$\mathscr{D}_i(\mu_i) = \underset{\mathbf{P}_i \succeq \mathbf{0}}{\text{maximize}} \;\; \mathscr{L}_i(\mathbf{P}_i;\mu_i) \tag{A.20}$$

Then the optimal solution can be found by solving the dual problem: $\underset{\mu_i \geq 0}{\text{minimize}} \;\; \mathscr{D}_i(\mu_i)$. To start with, the Lagrangian can be written as

$$\begin{aligned}\mathscr{L}_i(\mathbf{P}_i;\mu_i) =& w_i \log_2 \det\left(\mathbf{I}_{N_r} + \tilde{\mathbf{V}}_i^H \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \tilde{\mathbf{V}}_i\right) \\ & - \operatorname{Tr}\left(\tilde{\mathbf{V}}_i^H (\tilde{\mathbf{E}}_i + \mu_i\mathbf{I})\tilde{\mathbf{V}}_i\right) + \mu_i p_{i,\max}\end{aligned}$$

where $\tilde{\mathbf{V}}_i = \mathbf{V}_i \mathbf{P}_i^{1/2}$ is unnormalized transmit precoder of i-th BS. Using Theorem 2, we use the fact that, the generalized eigenmatrix $\mathbf{V}_i$ has the following diagonalization properties [139]

$$\mathbf{V}_i^H \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{V}_i = \mathbf{C}_i \;\; , \;\; \mathbf{V}_i^H (\tilde{\mathbf{E}}_i + \mu_i\mathbf{I})\mathbf{V}_i = \mathbf{G}_i \tag{A.21}$$

where $\mathbf{C}_i = \operatorname{diag}(c_i)$ and $\mathbf{G}_i = \operatorname{diag}(g_i)$ are diagonal matrices. Hence, the Lagrangian can be rewritten as

$$\mathscr{L}_i(\mathbf{P}_i;\mu_i) = w_i \log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{C}_i\mathbf{P}_i\right) - \operatorname{Tr}\left(\mathbf{G}_i\mathbf{P}_i\right) \tag{A.22}$$

and the optimization problem (A.20) can be expressed as

$$\underset{\mathbf{P}_i \succeq \mathbf{0}}{\text{maximize}} \;\; w_i \log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i\right) - \mathrm{Tr}\left(\mathbf{G}_i \mathbf{P}_i\right). \tag{A.23}$$

Taking the lagrangian of the above problem into account we have

$$\begin{aligned} \mathscr{L}(\mathbf{P}_i; \Psi) = & -w_i \log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i\right) \\ & + \mathrm{Tr}\left(\mathbf{G}_i \mathbf{P}_i\right) - \mathrm{Tr}\left(\Psi \mathbf{P}_i\right) \end{aligned} \tag{A.24}$$

which by taking derivative with respect to $\mathbf{P}_i$ of the above equation we get

$$\nabla_{\mathbf{P}_i} \mathscr{L} = -w_i (\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i)^{-1} \mathbf{C}_i + \mathbf{G}_i - \Psi \tag{A.25}$$

Hence, the first order condition (Karush-Kuhn-Tucker (KKT) conditions) can be written as

$$\begin{aligned} \mathbf{P}_i \succeq \mathbf{0}, \;\; \Psi \succeq \mathbf{0}, \;\; \Psi \mathbf{P}_i = \mathbf{0}, \\ w_i (\mathbf{I}_{N_r} + \mathbf{C}_i \mathbf{P}_i)^{-1} \mathbf{C}_i + \Psi = \mathbf{G}_i \end{aligned} \tag{A.26}$$

then, it can be seen that the optimal $\mathbf{P}_i$ and Ψ must be diagonal. By defining $\mathbf{P}_i = \mathrm{diag}(p_i)$ and $\Psi = \mathrm{diag}(\psi_i)$, the KKT conditions are

$$p_k \geq 0, \;\; \psi_k \geq 0, \;\; p_k \psi_k = 0, \;\; \frac{w_i c_k}{1 + c_k p_k} + \psi_k = g_k \tag{A.27}$$

more compactly, we can write

$$\mathbf{P}_i(k,k) = \left[\frac{w_i \mathbf{C}_i(k,k) - \mathbf{G}_i(k,k)}{\mathbf{C}_i(k,k) \mathbf{G}_i(k,k)} \right]^+ \tag{A.28}$$

where the remaining elements of $\mathbf{P}_i$ are zero. Note that, to solve (3.14), one needs to find the optimal Lagrange multiplier μ_i by solving $\underset{\mu_i \geq 0}{\text{minimize}} \;\; \mathscr{D}_i(\mu_i)$. This can be accomplished using a

simple bisection search.

A.0.6 Proof of Theorem 6

Our proof is inspired by the work of [140]. Using Cholesky decomposition of the matrix $\tilde{\mathbf{E}}_i + \mu_i \mathbf{I} = \mathbf{L}_i \mathbf{L}_i^H$ and defining $\bar{\mathbf{V}}_i = \mathbf{L}_i^H \tilde{\mathbf{V}}_i$, $\mathscr{L}_i(\mathbf{P}_i; \mu_i)$ is rewritten as

$$\begin{aligned}\mathscr{L}_i(\mathbf{P}_i; \mu_i) =& w_i \log_2 \det\left(\mathbf{I}_{N_r} + \bar{\mathbf{V}}_i^H \mathbf{L}_i^{-1} \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{L}_i^{-H} \bar{\mathbf{V}}_i\right) \\ & - \mathrm{Tr}\left(\bar{\mathbf{V}}_i^H \bar{\mathbf{V}}_i\right)\end{aligned}$$

Let $\bar{\mathbf{R}}_i = \bar{\mathbf{V}}_i \bar{\mathbf{V}}_i^H$ and $\mathbf{L}_i^{-1} \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{L}_i^{-H} = \mathbf{U}_i \mathbf{D}_i \mathbf{U}_i^H$ be the eigenvalue/eigenvector decomposition of $\mathbf{L}_i^{-1} \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{L}_i^{-H}$ where $\mathbf{U}_i$ is unitary and $\mathbf{D}_i$ is diagonal with the diagonal entries arranged in decreasing order. It is not difficult to verify that $\mathscr{L}_i(\mathbf{P}_i; \mu_i)$ is equivalent to the following equation

$$\mathscr{L}_i(\mathbf{P}_i; \mu_i) = w_i \log_2 \det\left(\mathbf{I}_{N_r} + \tilde{\mathbf{R}}_i \mathbf{D}_i\right) - \mathrm{Tr}\left(\tilde{\mathbf{R}}_i\right) \tag{A.29}$$

where $\tilde{\mathbf{R}}_i = \mathbf{U}_i^H \bar{\mathbf{R}}_i \mathbf{U}_i$. By Hadamard's inequality, it can be seen that the optimal $\tilde{\mathbf{R}}_i$ must be diagonal. Hence,

$$\tilde{\mathbf{R}}_i^{-1/2} \mathbf{U}_i^H \bar{\mathbf{R}}_i \mathbf{U}_i \tilde{\mathbf{R}}_i^{-1/2} = \tilde{\mathbf{R}}_i^{-1/2} \mathbf{U}_i^H \bar{\mathbf{V}}_i \bar{\mathbf{V}}_i^H \mathbf{U}_i \tilde{\mathbf{R}}_i^{-1/2} \tag{A.30}$$

which that means $\bar{\mathbf{V}}_i^H \mathbf{U}_i \tilde{\mathbf{R}}_i^{-1/2}$ is a unitary matrix and the optimal solution $\bar{\mathbf{V}}_i$ only depends on $\bar{\mathbf{R}}_i$. It is well known that if $\bar{\mathbf{V}}_i$ is an optimal solution, the product of $\bar{\mathbf{V}}_i$ with an arbitrary unitary matrix is also optimal. Therefore, $\bar{\mathbf{V}}_i = \mathbf{U}_i \tilde{\mathbf{R}}_i^{1/2}$. It is worth mentioning that, because $\tilde{\mathbf{R}}_i$ is a diagonal matrix and $\mathbf{U}_i$ is an eigenmatrix of $\mathbf{L}_i^{-1} \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{L}_i^{-H}$, then $\bar{\mathbf{V}}_i$ is an eigenmatrix of $\mathbf{L}_i^{-1} \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{L}_i^{-H}$ with unnormalized columns, i.e.,

$$\mathbf{L}_i^{-1} \mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \mathbf{L}_i^{-H} \bar{\mathbf{V}}_i = \bar{\mathbf{V}}_i \Lambda_i \tag{A.31}$$

Substituting back $\bar{\mathbf{V}}_i = \mathbf{L}_i^H \tilde{\mathbf{V}}_i$, we have

$$\mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii} \tilde{\mathbf{V}}_i = \mathbf{L}_i \bar{\mathbf{V}}_i \Lambda_i = \mathbf{L}_i \mathbf{L}_i^H \tilde{\mathbf{V}}_i \Lambda_i = (\tilde{\mathbf{E}}_i + \mu_i \mathbf{I}) \tilde{\mathbf{V}}_i \Lambda_i$$

hence, $\tilde{\mathbf{V}}_i$ is the generalized eigenmatrix of $\mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii}$ and $\tilde{\mathbf{E}}_i + \mu_i \mathbf{I}$. It is worth mentioning that $\mathbf{V}_i$ is also the generalized eigenmatrix of $\mathbf{H}_{ii}^H \mathbf{R}_i^{-1} \mathbf{H}_{ii}$ and $\tilde{\mathbf{E}}_i + \mu_i \mathbf{I}$.

A.0.7 Proof of Theorem 7

Since (3.18) is convex optimization problem, we can use a general-purpose method such as an interior point method to solve it in polynomial time. However, this problem admits a closed-form solution as can be derived from the KKT conditions. To this purpose, the partial Lagrangian of (3.18) is given by

$$\begin{aligned} \mathscr{L}_i(\mathbf{P}_i; \mu_i, \Psi) = & -w_i \log_2 \det(\mathbf{H}_{ii} \mathbf{V}_i \mathbf{P}_i \mathbf{V}_i^H \mathbf{H}_{ii}^H) \\ & + \mathrm{Tr}\,(\mathbf{A}_i \mathbf{P}_i) + \mu_i (\mathrm{Tr}(\mathbf{P}_i) - p_{i,\max}) - \mathrm{Tr}(\Psi \mathbf{P}_i) \end{aligned} \tag{A.32}$$

Hence, the gradient of the above Lagrangian can be written as

$$\nabla_{\mathbf{P}_i} \mathscr{L} = -w_i \mathbf{P}_i^{-1} + \mathbf{A}_i + \mu_i \mathbf{I} - \Psi \tag{A.33}$$

and the KKT conditions can be expressed as

$$\begin{gathered} \mathrm{Tr}(\mathbf{P}_i) \leqslant p_{i,\max} \ , \ \mathbf{P}_i \succ \mathbf{0} \\ w_i \mathbf{P}_i^{-1} + \Psi = \mathbf{A}_i + \mu_i \mathbf{I} \ , \ \Psi \succeq \mathbf{0} \\ \mu_i \left(\mathrm{Tr}(\mathbf{P}_i) - p_{i,\max}\right) = 0 \ , \ \Psi \mathbf{P}_i = \mathbf{0} \ , \ \mu_i \geqslant 0 \end{gathered} \tag{A.34}$$

To simplify the KKT conditions, let $\mathbf{A}_i = \mathbf{U} \Sigma \mathbf{U}^H$ be the eigenvalue/eigenvector decomposition of $\mathbf{A}_i$, where $\mathbf{U}$ is unitary and Σ is diagonal with the diagonal entries arranged in decreasing order

(denote the eigenvalues σ_i). It is not difficult to verify that the KKT conditions are equivalent to the following conditions

$$\begin{aligned} \mathrm{Tr}(\tilde{\mathbf{P}}_i) \leqslant p_{i,\max}\ ,\ \tilde{\mathbf{P}}_i \succ 0\ , \\ w_i \tilde{\mathbf{P}}_i^{-1} + \tilde{\Psi} = \Sigma + \mu_i \mathbf{I}\ ,\ \tilde{\Psi} \succeq 0 \\ \mu_i \left(\mathrm{Tr}(\tilde{\mathbf{P}}_i) - p_{i,\max}\right) = 0\ ,\ \tilde{\Psi}\tilde{\mathbf{P}}_i = \mathbf{0}\ ,\ \mu_i \geqslant 0 \end{aligned} \tag{A.35}$$

where $\tilde{\mathbf{P}}_i = \mathbf{U}^H \mathbf{P}_i \mathbf{U}$ and $\tilde{\Psi} = \mathbf{U}^H \Psi \mathbf{U}$. Assuming the optimal $\tilde{\mathbf{P}}_i$ and $\tilde{\Psi}$ are diagonal with $\tilde{\mathbf{P}}_i = \mathrm{diag}(\tilde{p}_i)$ and $\tilde{\Psi} = \mathrm{diag}(\psi)$, the KKT conditions become

$$\begin{aligned} \sum_k \tilde{p}_k \leqslant p_{i,\max},\ \tilde{p}_k > 0,\ \mu_i \geqslant 0,\ \psi \geq 0,\ \psi \tilde{p}_k = 0, \\ \frac{w_i}{\tilde{p}_k} + \psi = \sigma_k + \mu_i\ ,\ \mu_i \left(\sum_k \tilde{p}_k - p_{i,\max} \right) = 0\ . \end{aligned} \tag{A.36}$$

by looking into detail at the KKT conditions we have

$$\mathbf{P}_i = \mathbf{U}^H \,\mathrm{diag}\left[\frac{w_i}{\sigma_k + \mu_i} \right] \mathbf{U} \tag{A.37}$$

For example, for a network consisting of two users $\mathbf{A}_i = w_2 \mathbf{V}_1^H \mathbf{H}_{21}^H \mathbf{R}_2^{-1} \mathbf{H}_{21} \mathbf{V}_1 = \mathbf{U}\,\mathrm{diag}(\sigma_i)\mathbf{U}^H$ where $\mathbf{R}_2$ is evaluated at $\mathbf{P}_1 = \bar{\mathbf{P}}_1$.

A.0.8 Proof of Theorem 8

Instead of solving (3.21), we can also solve

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad \mathrm{Tr}\left((w_i \mathbf{C}_i - \mathbf{D}_i)\mathbf{P}_i\right) + \varepsilon \log_2 \det(\mathbf{P}_i) \tag{A.38}$$

where $\varepsilon \in [0, \bar{\varepsilon})$ is a pre-specified constant and $f(\mathbf{P}_i) = \log_2 \det(\mathbf{P}_i)$ is the perturbation function. Even though other perturbation functions could also be considered, the log det function has two

crucial benefits. First, it makes the problem strictly convex enabling us to adapt a successive projections technique to obtain an efficient algorithm. Second, it proves indispensable in enforcing positive definiteness without any eigenvalue computations because $\log_2 \det(\mathbf{P}_i)$ is a natural barrier function for the cone of positive definite matrices. It has shown that for an appropriate value of ε, a solution of (3.21) is also a solution of (A.38) [141, Theorem 2.1.]. Since (A.38) is strictly convex, solving it allows us to obtain a unique solution amongst all possible solutions of (3.21).

The partial Lagrangian of (A.38) is given by

$$\begin{aligned}\mathscr{L}_i(\mathbf{P}_i;\mu_i,\Psi) = &-\operatorname{Tr}\left((w_i\mathbf{C}_i-\mathbf{D}_i)\mathbf{P}_i\right)-\operatorname{Tr}(\Psi\mathbf{P}_i)\\ &-\varepsilon\log_2\det(\mathbf{P}_i)+\mu_i(\operatorname{Tr}(\mathbf{P}_i)-p_{i,\max})\end{aligned}$$

and the gradient of the Lagrangian is

$$\nabla_{\mathbf{P}_i}\mathscr{L} = -w_i\mathbf{C}_i+\mathbf{D}_i-\varepsilon\mathbf{P}_i^{-1}+\mu_i\mathbf{I}-\Psi \tag{A.39}$$

The KKT conditions are

$$\begin{aligned}\operatorname{Tr}(\mathbf{P}_i)\leqslant p_{i,\max}\ ,\ \mathbf{P}_i\succeq\mathbf{0}\ ,\\ \varepsilon\mathbf{P}_i^{-1}+\Psi = -w_i\mathbf{C}_i+\mathbf{D}_i+\mu_i\mathbf{I}\ ,\ \Psi\succeq\mathbf{0}\ ,\\ \mu_i\left(\operatorname{Tr}(\mathbf{P}_i)-p_{i,\max}\right)=0\ ,\ \Psi\mathbf{P}_i=\mathbf{0}\ ,\ \mu_i\geqslant 0.\end{aligned} \tag{A.40}$$

To simplify the KKT conditions, let $-w_i\mathbf{C}_i+\mathbf{D}_i = \mathbf{V}\Gamma\mathbf{V}^H$ be the eigenvalue/eigenvector decomposition of $-w_i\mathbf{C}_i+\mathbf{D}_i$, where $\mathbf{V}$ is unitary and Γ is diagonal with the diagonal entries arranged in decreasing order (denote the eigenvalues γ_i). It is not difficult to verify that the KKT conditions

are equivalent to the following conditions

$$\begin{aligned} \mathrm{Tr}(\tilde{\mathbf{P}}_i) \leqslant p_{i,\max}\ ,\ \tilde{\mathbf{P}}_i \succeq \mathbf{0}\ , \\ \varepsilon\tilde{\mathbf{P}}_i^{-1} + \tilde{\Psi} = \Gamma + \mu_i\mathbf{I}\ ,\ \tilde{\Psi} \succeq \mathbf{0} \\ \mu_i\left(\mathrm{Tr}(\tilde{\mathbf{P}}_i) - p_{i,\max}\right) = 0\ ,\ \tilde{\Psi}\tilde{\mathbf{P}}_i = \mathbf{0}\ ,\ \mu_i \geqslant 0\ , \end{aligned} \tag{A.41}$$

where $\tilde{\mathbf{P}}_i = \mathbf{V}^H\mathbf{P}_i\mathbf{V}$ and $\tilde{\Psi} = \mathbf{V}^H\Psi\mathbf{V}$. Assuming the optimal $\tilde{\mathbf{P}}_i$ and $\tilde{\Psi}$ are diagonal with $\tilde{\mathbf{P}}_i = \mathrm{diag}(\tilde{p}_i)$ and $\tilde{\Psi} = \mathrm{diag}(\psi)$, the KKT conditions become

$$\begin{aligned} \sum_k \tilde{p}_k \leqslant p_{i,\max}\ ,\ \tilde{p}_k \geq 0\ ,\ \mu_i \geqslant 0\ ,\ \psi \geq 0 \\ \frac{\varepsilon}{\tilde{p}_k} + \psi = \gamma_k + \mu_i\ ,\ \mu_i\left(\sum_k \tilde{p}_k - p_{i,\max}\right) = 0\ ,\ \psi\tilde{p}_k = 0 \end{aligned} \tag{A.42}$$

by looking into detail at the KKT conditions we have

$$\mathbf{P}_i = \mathbf{V}^H\,\mathrm{diag}\left[\frac{\varepsilon}{\gamma_k + \mu_i}\right]\mathbf{V} \tag{A.43}$$

which for a network consists of two users, γ_k is the eigenvalue of matrix $-w_1\mathbf{C}_1 + w_2\mathbf{Y}_2^H\chi_2\mathbf{Y}_2$ where $\mathbf{Y}_2$ is evaluated at $\mathbf{P}_1 = \bar{\mathbf{P}_1}$.

A.0.9 Proof of Theorem 9

$\mathbf{C}_i \succcurlyeq \mathbf{0}$ can be partitioned into two blocks, its positive definite and zero parts, and $\mathbf{P}_i$ as:

$$\mathbf{C}_i \triangleq \begin{bmatrix} \mathbf{C}_{i,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{P}_i \triangleq \begin{bmatrix} \mathbf{P}_{i,11} & \mathbf{P}_{i,12} \\ \mathbf{P}_{i,21} & \mathbf{P}_{i,22} \end{bmatrix} \tag{A.44}$$

where $\mathbf{C}_{i,11} \succeq \mathbf{0}$ and $\mathbf{P}_{i,11}$ and $\mathbf{C}_{i,11}$ have the same dimensions. Then, (3.23) can be rewritten as (A.45).

Since $\mathbf{P}_i \in \mathscr{K}$, by definition $\mathbf{P}_{i,11} \in \mathscr{K}$. Let $\breve{\mathbf{P}} = \mathbf{P}_{i,11}$, then,(A.45) is equivalent to (A.46). This

$$\underset{\mathbf{P}_i \in \mathscr{K}}{\text{maximize}} \quad w_i \log_2 \det(\mathbf{I} + \mathbf{P}_{i,11}\mathbf{C}_{i,11}) - \mathrm{Tr}\,(\mathbf{E}_i\mathbf{P}_i) - \tau_i \|\mathbf{P}_i - \bar{\mathbf{P}}_i\|_F^2. \tag{A.45}$$

$$\begin{aligned} \underset{\mathbf{P}_i,\, \mathbf{P}}{\text{maximize}} \quad & w_i \log_2 \det(\mathbf{I} + \check{\mathbf{P}}\mathbf{C}_{i,11}) - \mathrm{Tr}\,(\mathbf{E}_i\mathbf{P}_i) - \tau_i \|\mathbf{P}_i - \bar{\mathbf{P}}_i\|_F^2, \\ \text{subject to} \quad & \mathbf{P}_i \in \mathscr{K},\ \check{\mathbf{P}} = \mathbf{P}_{i,11} \in \mathscr{K} \end{aligned} \tag{A.46}$$

problem can be solved via dual decomposition [142, Lemma 2] and we have

$$\mathbf{P}_i = \left[\bar{\mathbf{P}}_i - \frac{1}{2\tau_i}(\mu^* \mathbf{I} + \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}) \right]^+ \tag{A.47}$$

where $\mathbf{Z}$ is the matrix of lagrangian multipliers associated to the linear constraints $\check{\mathbf{P}} = \mathbf{P}_{i,11}$, $[\mathbf{X}]^+$ denotes the projection of $\mathbf{X}$ onto the cone of positive semidefinite matrices, and μ^* is the multiplier which can be found by bisection.

A.0.10 $f_i(\mathbf{Q}_i, \mathbf{Q}_{-i})$ is a convex function of $\mathbf{Q}_i$ for fixed $\mathbf{Q}_{-i}$ for any $i \in \mathscr{L}$.

The claim will be established by using the following property of convex functions. A function is convex if and only if it is convex when restricted to any line that intersects its domain [143, Chapter 3]. To be specific, given an arbitrary function $f(x)$ and two different feasible points x_1 and x_2, define $g(t) = f(tx_1 + (1-t)x_2), 0 \le t \le 1$. Then, $f(x)$ is a convex function of x if and only if $g(t)$ is a convex function of t for any feasible x_1 and x_2, which is equivalent to $(d^2 g(t)/dt^2) \ge 0$ for $0 \le t \le 1$. Without loss of generality, we consider $i = 1$. Define $R_s = \sum_{i=1}^{L} w_i R_i$ for any fixed $\{\mathbf{Q}_i\}_{i=2}^{L}$. Then we get

$$\begin{aligned} R_s = {} & w_1 \log_2 \det(\mathbf{I} + \mathbf{R}_1^{-1}\mathbf{H}_{11}\mathbf{Q}_1\mathbf{H}_{11}^H) \\ & + \sum_{j=2}^{L} w_j \log_2 \det(\mathbf{I} + \mathbf{R}_j^{-1}\mathbf{H}_{jj}\mathbf{Q}_j\mathbf{H}_{jj}^H) \end{aligned} \tag{A.48}$$

Based on the aforementioned property of convex functions, by investigating the convexity of

R_s restricted to any line, we can prove that R_s is a convex function of $\mathbf{Q}_1 \in \mathscr{K}$. Consider

$$\mathbf{Q}_1 = t\mathbf{Z}_1 + (1-t)\mathbf{X}_1 = \mathbf{X}_1 + t\mathbf{Y}_1,\ 0 \le t \le 1 \tag{A.49}$$

where $\mathbf{X}_1 \in \mathscr{K}$ and $\mathbf{Z}_1 \in \mathscr{K}$. Note that $\mathbf{Y}_1 \notin \mathscr{K}$ but $\mathbf{Y}_1$ is Hermitian. Moreover, $\mathbf{R}_1$ and $\{\chi_j\}_{j=2}^{L}$ do not depend on t. Recalling that $d\ln(\det(\mathbf{Z})) = \mathrm{Tr}\left\{\mathbf{Z}^{-1}d\mathbf{Z}\right\}$ for all $\mathbf{Z}$ such that $\det\mathbf{Z} \neq 0$ [63, Prop. 3.14], the first-order differential of R_s is given by (up to a constant positive factor)

$$\begin{aligned}\frac{dR_s}{dt} &= \mathrm{Tr}\Big\{(\mathbf{R}_1+\chi_1)^{-1}\mathbf{B}_1\Big\} \\ &\quad + \sum_{j=2}^{L}\mathrm{Tr}\Big\{-(\mathbf{R}_j+\chi_j)^{-1}\chi_j\mathbf{R}_j^{-1}\mathbf{B}_j\Big\}\end{aligned} \tag{A.50}$$

where $\mathbf{B}_j = \mathbf{H}_{j,1}\mathbf{Y}_1\mathbf{H}_{j,1}^H$ is Hermitian. Further, we have

$$\begin{aligned}\frac{d^2R_s}{dt^2} &= \mathrm{Tr}\Big\{-(\mathbf{R}_1+\chi_1)^{-1}\mathbf{B}_1(\mathbf{R}_1+\chi_1)^{-1}\mathbf{B}_1\Big\} \\ &\quad + \sum_{j=2}^{L}\mathrm{Tr}\Big\{(\mathbf{R}_j+\chi_j)^{-1}\mathbf{B}_j(\mathbf{R}_j+\chi_j)^{-1}\chi_j\mathbf{R}_j^{-1}\mathbf{B}_j \\ &\quad + (\mathbf{R}_j+\chi_j)^{-1}\chi_j\mathbf{R}_j^{-1}\mathbf{B}_j\mathbf{R}_j^{-1}\mathbf{B}_j\Big\}\end{aligned} \tag{A.51}$$

We define $\mathbf{F}_j \triangleq (\mathbf{R}_j+\chi_j)^{-1}\chi_j\mathbf{R}_j^{-1} = \mathbf{R}_j^{-1} - (\mathbf{R}_j+\chi_j)^{-1}$. Since χ_j is a positive semi-definite (PSD), we have $\mathbf{R}_j+\chi_j \succeq \mathbf{R}_j$. Then, $(\mathbf{R}_j+\chi_j)^{-1} \preceq \mathbf{R}_j^{-1}$ which means that $\mathbf{F}_j$ is PSD and so there exists matrix $\mathbf{J}_j$ such that $\mathbf{F}_j = \mathbf{J}_j\mathbf{J}_j^H$. So, we get

$$\begin{aligned}\mathrm{Tr}\Big\{(\mathbf{R}_j+\chi_j)^{-1}\mathbf{B}_j\mathbf{F}_j\mathbf{B}_j\Big\} &= \mathrm{Tr}\Big\{\mathbf{K}_j\mathbf{K}_j^H\mathbf{B}_j\mathbf{J}_j\mathbf{J}_j^H\mathbf{B}_j\Big\} \\ &= \mathrm{Tr}\Big\{(\mathbf{K}_j^H\mathbf{B}_j\mathbf{J}_j)(\mathbf{K}_j^H\mathbf{B}_j\mathbf{J}_j)^H\Big\} \ge 0\end{aligned} \tag{A.52}$$

where we use the fact that $(\mathbf{R}_j+\chi_j)^{-1}$ is a PSD and can be rewrite as $\mathbf{K}_j\mathbf{K}_j^H$. Moreover, $(\mathbf{K}_j^H\mathbf{B}_j\mathbf{E}_j)(\mathbf{K}_j^H\mathbf{B}_j\mathbf{E}_j)^H$ is PSD. Similarly, the other terms are greater than or equal to zero and so, $d^2R_s/dt^2 \ge 0$ which means R_s is a convex function of $\mathbf{Q}_1$.