TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABREVIATIONS

3GPP	3rd Generation Partnership Project
4G	Forth Generation
5G	Fifth Generation
ADC	Analog-Digital Converter
ADMM	Alternating Direction Method of Multipliers
AEE	Access Energy Efficiency
AO	Alternating Optimization
AtG	Air-to-Ground
AWGN	Additive White Gaussian Noise
BnB	Branch-and-Bound
BS	Base Station
CoMP	Coordinated Multi-Point
CoTDD	Co-channel Time Division Duplexing
C-RAN	Cloud Radio Access Network
CSI	Channel State Information
DAC	Digital-Analog Converter
DBP	Dynamic Back-Pressure
DC	Difference of Convex
DL	Downlink

DSL	Digital Subscriber Loop
EE	Energy Efficiency
FDD	Frequency Division Duplexing
FOTCA	First Order Taylor Convex Approximation
FSO	Free Space Optic
GNCP	General Nonlinear Convex Program
GP	Geometric Programming
GtA	Ground-to-Air
HCN	Heterogeneous Cellular Network
HetNets	Heterogeneous Networks
IoT	Internet of Things
JBPAO	Joint Beamforming and Power Allocation Optimization
JBRAO	Joint Beamforming and Resource Allocation Optimization
KKT	Karush-Kuhn-Tucker
LoS	Line-of-Sight
MAC	Medium Access Control
MBS	Macrocell Base Station
MC	Multi-Cell
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output

MISOCP	Mixed Integer Second Order Cone Programming
MMM	Majorization Minimization Method
MMSE	Minimum Mean Square Error
mmWave	Millimeter Wave
MRC	Maximum Ratio Combining
MRT	Maximum Ratio Transmission
MU	Multi-User
MUE	Macrocell User
NLoS	Non-Line-of-Sight
NOMA	Non-Orthogonal Multiple Access
NP	Non-deterministic Polynomial-time
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational Expenditure
OTN	Optical Transport Network
PC	Power Control
PON	Passive Optical Network
PtMP	Point-to-Multi-Point
PSD	Power Spectral Density
PtP	Point-to-Point
QoE	Quality-of-Experience

QoS	Quality-of-Service
QSI	Queuing State Information
RB	Resource Block
RF	Radio Frequency
RTDD	Reverse Time Division Duplexing
SAP	Small Cell Access Point
SBS	Small Cell Base Station
SC	Small Cell
SCA	Successive Convex Approximation
SCN	Small Cell Network
SDP	Semi-Definite Programming
SE	Spectral Efficiency
SIC	Successive Interference Cancellation
SINR	Signal-to-Interference-plus-Noise Ratio
SISO	Single Input Single Output
SLNR	Signal-to-Leakage-plus-Noise Ratio
SOC	Second Order Cone
SOCP	Second Order Cone Programming
SPCA	Sequential Parametric Convex Approximation
SUE	Small Cell User

SSE	Sum Spectral Efficiency
TDD	Time Division Duplexing
UAV	Unmanned Aerial Vehicle
UDN	Ultra-Dense Network
UL	Uplink
WA	Wireless Access
WAC	Wireless Access Communications
WB	Wireless Backhaul
WBC	Wireless Backhaul Communication
WDM	Wavelength-Division Multiplexing
WMMSE	Weighted Minimum Mean Square Error
ZF	Zero Forcing

LISTE OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathbb{R}$	Set of real variables
$\mathbb{C}$	Set of complex variables
$\mathbb{E}(f(\mathsf{X}))$	Expectation operation of function of random variable X
$\|x\|$	Absolute operation of $x \in \mathbb{C}$
x^*	Complex conjugate of $x \in \mathbb{C}$
$\mathscr{R}(x)$	Real part of $x \in \mathbb{C}$
$\mathscr{I}(x)$	Image part of $x \in \mathbb{C}$
$\mathbf{x}$	Column vector or set of vectors
$\mathscr{X}$	Set of vectors $\mathbf{x}_j$, variables x_j, or index j
$\mathscr{X}_{\sim i}$	Set excluding vector $\mathbf{x}_i$,variable x_i, or index i
$(\mathbf{x})^H$	Hermitian operation of vector $\mathbf{x}$
$(\mathbf{x})^T$	Transpose operation of vector $\mathbf{x}$
$\|\mathbf{x}\|$ or $\|\mathbf{x}\|_2$	ℓ_2 norm of vector $\mathbf{x}$
$\|\mathbf{x}\|_\infty$	ℓ_∞ norm of vector $\mathbf{x}$
$\mathbf{X}$	Matrix
$\|\mathbf{X}\|_F$	Frobonius norm of matrix $\mathbf{X}$
Bdiag$(\mathbf{X}_1,\ldots,\mathbf{X}_N)$	Block diagonalization of matrices $\mathbf{X}_1,\ldots,\mathbf{X}_N)$
$\mathscr{O}$	Big O notation
$\nabla f(x,y)$	Gradient (vector) of scalar function $f(x,y)$

$f(x)'$	Derivative of scalar function $f(x)$ with respect to x
$\langle \mathbf{x}^T, \mathbf{y} \rangle$	Scalar product of vector $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{x} \circ \mathbf{y}$	Hadamard product of vector $\mathbf{x}$ and $\mathbf{y}$
bits	Unit of amount of information
bits/s	Unit of achievable rate
bits/s/Hz	Unit of spectral efficiency
bits/Joule/Hz	Unit of energy efficiency

INTRODUCTION

The evolution of mobile wireless networks towards their Fifth Generation (5G) establishment and beyond (Andrews *et al.*, 2014; Nikolich *et al.*, 2017) has seen an increasingly massive number of connected devices ensuing from the *Internet of Things* (IoT) and *Industry 4.0* phenomena. According to the report in (Lucero, 2016), the amount of wireless equipments will outreach 75 billions by 2025. Among many key enabling technologies (Boccardi *et al.*, 2014) such as massive multi-input multi-output (MIMO), millimeter wave (mmWave), full-duplex (FD), 5G promotes Ultra Dense Network (UDN) technology (Ge *et al.*, 2016) as one significant candidate to fulfill its promises on enhancing user's throughput, catering seamless coverage, and reducing overall system power. In general, UDNs spatially deploys a massive number of small cell base stations (BSs) in the proximity of local users to reduce communication range, thus can ubiquitously guarantee undisrupted services to cell-edge users, significantly improve area spectral efficiency (SE) and efficiently conserve transmit power. Despite many inherited benefits, dense deployment of classical small cell architecture quickly raises the operational cost (OPEX) due to a huge amount of fiber backhaul establishment to the core networks, which causes new economical and deployment challenges.

Recently, wireless backhaul (WB) technology has emerged as a cost-effective and pragmatic architectural solution, which enables the transportation of backhaul data over-the-air to the remote small cell without the needs of fiber connections. Therefore, it can potentially sustain the operation of dense networks. In the context of 5G, WB in dense heterogeneous networks (HetNets) is found most applicable and advantageous when being accommodated under the mmWave, microwave, and sub–6 Gigahertz spectrum bands to exploit both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) communications. Especially in the last case, NLoS WB can fully exploit the availability of existing hardwares and developed wireless access (WA) technologies for immediate WB's experiment towards its future deployment. Despite the introduced advantages, this new WB scheme must ensure the quality of backhaul rate as high as

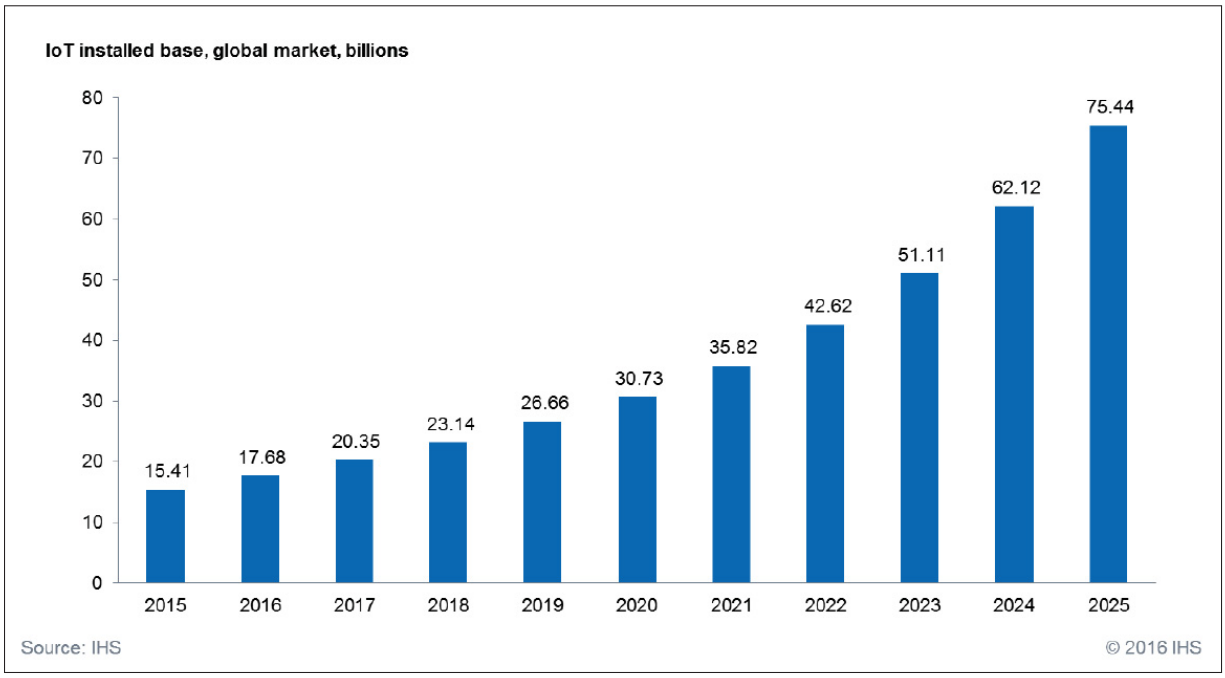

Figure 0.1 The IoT market will be massive (Lucero, 2016).

the conventional wired ones so as to assist the WA communications between the small cells and local users. Note that in the sub–6 Gigahertz case, simultaneous WB communications are concurrently operated over the same spectral channel, also known as Point-to-Multi-Point (PtMP) in-band WB. Under such circumstance, WB signals are severely interfered by other concurrent WB and WA transmissions. This becomes a fundamental bottleneck of PtMP WB HetNets and makes it more difficult to manage a proper system resource allocation to respect the stringent WB-WA rate constraint while optimizing the network performance. Motivated by these observations, this dissertation aims at constructing several novel designs which jointly optimize the resource allocation in NLoS PtMP WB HetNets to improve WB's operation and consolidate its viability in 5G dense networks.

The detailed organization of this dissertation, which contains 7 chapters, is as follows. Chapter 1 introduces the overview of WB in the scope of 5G HetNets, which is followed by its challenges and existing solutions. Then, this chapter presents the dissertation's motivations

which lead directly to its objectives. In addition, it also highlights the significant novel contributions of each technical chapter throughout the dissertation. Towards its end, the main methodology employed to solve all the formulated problems is described.

Chapter 2 presents the first article which studies the operation of NLoS PtMP WB small cell HetNets under the application of RTDD system. Observing that the uplink and downlink WB and WA transmissions are coupled under the proposed scheme, a joint design of transmit and receive beamforming at the macrocell BS (MBS), the transmit power at the small cell access points (SAPs) and users, along with the spectrum partitioning which maximizes the small cell user (SUE) sum rate on both uplink and downlink is presented. Then, a high-complexity branch-and-bound (BnB) algorithm based on monotonic optimization and an efficient low-complexity algorithm based on successive convex approximation (SCA) second order cone programming (SOCP) are developed to solve for a global optimal and a sub-optimal solution, respectively. Achieved results show that the proposed joint consideration of uplink and downlink transmissions significantly outperforms other existing designs. Besides, the developed SCA- and SOCP-based algorithm is shown more efficient in terms of convergent and time-consuming performance compared to the conventional solution approaches.

Chapter 3 presents the second article which investigates a similar WB small cell HetNet, where an equal spectrum partitioning is fixed for the RTDD system. Under this consideration, this work concentrates on a downlink design of transmit beamforming and power allocation which maximizes the access energy-efficiency, defined by the ratio between the user sum rate and the overall network power consumption. In this work, a novel non-linear power consumption model which reflects the mathematical dependence of the consumed power with respect to the small cell backhaul rate is proposed. Followed by the claim of the formulated problem's NP-hardness, this work develops two centralized high- and low-complexity algorithms to achieve the global optimal and sub-optimal solutions, respectively. Then, by leveraging the advantage

of alternating direction method of multipliers (ADMM) method and the capability of message exchanging between the considered BSs, a decentralized algorithm is developed to offload the resource computing tasks among the BSs in order to achieve a distributed solution. Numerical results validate the superiority of the novel power consumption model and the efficiency of the low-complexity centralized and decentralized algorithms.

Chapter 4 presents the third article regarding the downlink resource allocation of a WB small cell HetNet over multiple time slots, where each small cell is equipped with a buffer of finite size. In this work, wireless channels are assumed time-varying after each time slot and a variant of RTDD, namely co-channel TDD (CoTDD), combined with fixed spectrum partitioning is employed. Under this assumption, an optimization problem which maximizes the SUE sum rate of all the considered time slots is formulated. This problem is first theoretically proved to improve the network performance when the small cell buffers are exploited properly. Following these findings, two practical low-complexity online algorithms are developed to solve for a joint solution of transmit power, beamforming, and buffer usage to obtain the expected performance. Through numerical results, these algorithms are shown to perform close to the benchmarking off-line algorithm, which again validates the theoretical statement about the enhancement from the proposed small cell buffer scheme.

Chapter 5 presents the forth article which enhances the performance of WB small cell HetNets under the proposed cooperative multi-input single-output (MISO) non-orthogonal multiple access (NOMA). In particular, the MISO NOMA scheme similar to (Hanif *et al.*, 2016) is employed for the downlink WB communications from the MBS to the SAPs, and a novel cooperation scheme between the small cells, stemmed from the successive interference cancellation (SIC) protocol, is proposed to improve the downlink WA communications. Under this scheme, a resource allocation of transmit power, beamforming, together with the decoding order of the SIC NOMA and the cooperative policy between the SAPs are jointly designed to optimize two

separate problems. The first problem is to maximize the user sum rate. The second problem is to maximize the number of admitted users which satisfy the minimum rate requirement. In this work, a method based on the principle of difference of convex (DC) programming combined with Lipschitz continuity is derived to efficiently solve the formulated combinatorial non-linear problem for a high-quality sub-optimal solution. Numerical results are extensively conducted to show that the proposed cooperative NOMA scheme remarkably improves the performance of the WB HetNet and outperforms all the existing approaches in the literature.

Finally, Chapter 6 presents the fifth article which investigates the improvement of the unmanned aerial vehicle (UAV)-assisted WB HetNet compared to the traditional small cell WB HetNets. In this new network, each WB SAP resembles a UAV, which can flexibly fly to any position in addition to the capabilities of receiving data via WB communications and transmit data via WA communications. Then, this work revisits the impact of cooperative NOMA scheme proposed in Chapter 6 on the new UAV-assisted WB networks. The challenge is that the flexible changes of UAV's position modify the channel's discrepancy from the MBS to UAVs, which in turn affect the decision on the SIC's decoding order and UAV's cooperative policy. Thus, we formulate a problem which jointly optimizes the radio resource allocation together with the location of UAVs, the SIC's decoding order, and the UAV cooperation policy which maximizes the user sum rate. Solving the formulated problem is more challenging than all previous works due to a huge amount of mixed integer variables coupled through several non-convex functions. Consequently, employing the developed methods in Chapter 5 is not efficient. Toward this end, a more elegant with lower complexity algorithm based on the combination of DC programming and Lipschitz continuity is developed to overcome the raised challenges of Chapter 6's algorithm, which can more efficiently solve for a sub-optimal solution of the considered problem.

CHAPTER 1

WIRELESS BACKHAUL IN HETEROGENEOUS NETWORKS: OVERVIEW, BENEFITS, CHALLENGES, THESIS OBJECTIVE AND CONTRIBUTIONS, METHODOLOGY

1.1 Overview

Network Densification

The term network densification refers to the dense deployment of small cells over the cellular coverage (Bhushan *et al.*, 2014). Small cell architecture was first proposed to deploy small size, low power, short range base stations, capable of serving indoor and local users, to coexist with the underlying macrocell BSs which are mainly dedicated to serve outdoor mobile users (Chandrasekhar *et al.*, 2008). Towards the establishment of 5G and its subsequent generations, small cell and macrocell BSs are likely to be equipped with a large scale of antenna array, also known as massive MIMO, embraced by mmWave technologies to cope with the expected increase of $1000\times$ user's data traffic, $10-20\times$ SE and system energy-efficiency (EE) (Andrews *et al.*, 2014). In fact, due to the characteristics of wireless signal propagation, improving wireless transmissions by massive MIMO operated at mmWave band is compatible with the short-range communications, which subsequently leads to small cell coverage shrinkage (Ge *et al.*, 2014). Since this shrinkage creates several out-of-service spots (Ge *et al.*, 2014), as depicted in Fig. 1.1 this encourages a dense deployment of small cell BSs to form the UDN in order to ubiquitously serve users at a large scale. Upon the newest 3GPP Release 15 of 5G standard along with the proliferation of IoT and Industry 4.0, UDNs embraced by massive MIMO and mmWave will remain as a dominant candidate to support the expected user's demand of diverse QoS requirements.

In small cell HetNets, the essential component which maintains the backbone of the entire network is the backhaul architecture. Backhaul generally represents the family of connections which spans from the BSs to the core networks. It is important to note that the backhaul traffic

of small cell BSs are often aggregated at the macrocell BS (donor node (Siddique *et al.*, 2015b)) via wired or wireless backhaul connectivities before being routed to the core networks.

Wired Backhaul

A classical wired backhaul connection prefers fiber optic material to bridge the necessary links between the BSs and core networks since fiber backhaul can support high capacity and low delay (Ranaweera *et al.*, 2013). At the macrocell tier, wired backhaul connections can be distinguished based on the physical connection distances between the BSs and the core networks as follows (Wang *et al.*, 2015b):

- More than ten kilometers: optical transport network (OTN) with wavelength-division multiplexing (WDM).
- Less than ten kilometers: point-to-point (PtP) or PtMP unified passive optical network (Uni-PON) which uses an optical splitter to aggregate WDM signals from the backhaul links of multiple cells.

At the small cell tier, small cell BSs can take advantage of the existing indoor wired infrastructure such as digital subscriber loop (DSL) over copper wire (Dahrouj *et al.*, 2015), which is capable of providing up to 1 Gb/s of data rate when more sophisticated technology such as G.fast is employed (Wang *et al.*, 2015b). If there exists a fiber optic connectivity nearby, the small cell traffic via DSL link can be multiplexed into the main fiber optic link and transported towards the aggregated MBS node before forwarding to the core network. In most situations, ubiquitous wired backhaul connections are often deployed less ubiquitously due to their expensive cost. For a dense small cell network, this becomes excessive and impractical, and hence a more viable backhaul solution is always anticipated.

Wireless Backhaul

WB emerges as a replaceable and cost-effective technology which more flexibly and efficiently adapts with the backhauling demands compared to the wired backhaul. In fact, WB concept

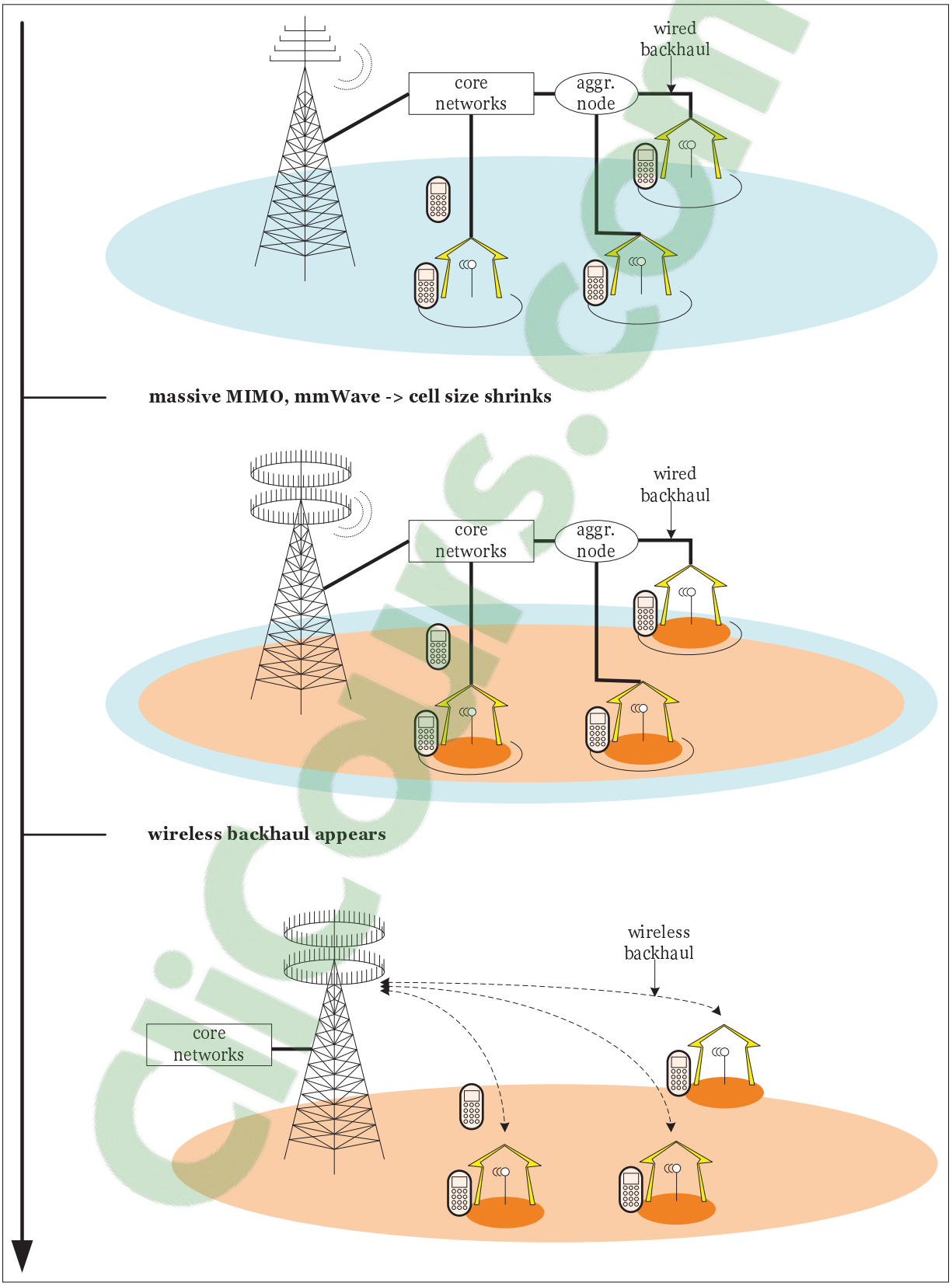

Figure 1.1 Evolution from wired backhaul to wireless backhaul.

is not entirely new and was used to relay information between the MBS and core networks in the traditional homogeneous cellular networks. In particular, WB LoS transmissions based on PtP microwave or free space optic (FSO) (Dahrouj *et al.*, 2015) have been shown to achieve comparatively high capacity as the fiber optic ones; and in some cases, it can be used for small cell backhaul if LoS condition is feasible. In general, when small cells are densely deployed in urban area with several surrounding giant buildings and blockages, NLoS transmissions are more favorable and practical for WB small cell BSs. However, NLoS transmissions often occur in the sub–6 GHz band where conventional wireless access transmissions preoccupy. Thus, accommodating WB technology here could create many concurrent transmissions and degrade each transmission's quality without a proper network resource planning.

According to the report of spectrum feature for WB in (Siddique *et al.*, 2015b), accommodation of WB on each spectrum band has some relations with the environmental conditions, hardware availability and desired quality requirement. The pros and cons characteristics of the WB solutions, which are summarized in Table 1.1, are explained as follows:

- **900 Mhz to sub–6 Ghz bands**: WB operated in these spectrum bands must be licensed but then most suitable for the NLoS communications. Therefore, sub–6 GHz WB is often used for low mobile applications within urban area. Beside maintaining wider coverage with low attenuation transmissions, it also economically benefits from the available hardwares and WA technologies without suffering from the burden of higher frequency migration and antenna alignment. However, since the frequency resource of sub–6 GHz band is scarce, newly operated WB must concurrently coexist with the other WA communications and causes (or suffers) severe interference to each other. Thus, an appropriate resource management is necessary to control the co-channel interference and attain an effective WB deployment on these spectrum bands.
- **600-800 MHz bands (TV white space)**: similar to the 900 Mhz to sub–6 Ghz bands, WB operated in the TV white space unlicensed bands also benefits from the NLoS communications. Although it can support wider coverage at low attenuation transmissions, operation of WB under these bands must deal with the expensive hardware cost and interference issue

in addition to the event of opportunistic availability of spectrum for WB accommodation. Therefore, TV white space WB is more likely to be applied in the sparsely populated area.

- **6–60 Ghz bands (microwave)**: Since these spectrum bands reserve a much wider licensed bandwidth, the new WB communications can be accommodated orthogonally here to avoid the impact of interference. Consequently, the cost of licensed frequency registration, in addition to the hardware cost, is high. Nonetheless, due to shorter wavelength, microwave signals are only suitable with LoS short range transmissions, which require no obstacle in-between since transmitted signals are easily attenuated and impaired during their propagation in the air. To overcome the LoS-related challenges, microwave-assisted antennas are traditionally placed at the high-ground positions on some tall buildings in the urban area, so that they can accurately align their directions to the intended destination and preserve a good quality communications.
- **60–80 Ghz bands (mmWave)**: these spectrum bands are recently considered a spectral gold-mine which attracts many wireless transmissions to migrate to. Most mmWave bands are categorized into unlicensed and light-licensed ones. WB operated under these bands includes all the advantages and limitations as in microwave bands, such as orthogonal transmission availability, interference-free communications, low coverage, high hardware cost, high attenuation transmission, antenna alignment, etc. In mmWave bands, multi-hop transmission strategies are frequently needed to wirelessly transmit backhaul data to the remote terminal. WB operated in the mmWave spectrum bands is often used in dense urban networks.
- **Satelite bands**: these bands span in several sub-bands, e.g., 4–6 (C band), 10–12 (Ku band), and 20–30 GHz (Ka band), which are especially used to support users with high mobility and can guarantee wide and ubiquitous coverage. However, WB transmissions operated under these bands are obviously expensive due to spectrum and hardware costs, and must highly suffer from the delay issues. Thus, it is only used at some rural or remote areas and for some mobile situations.

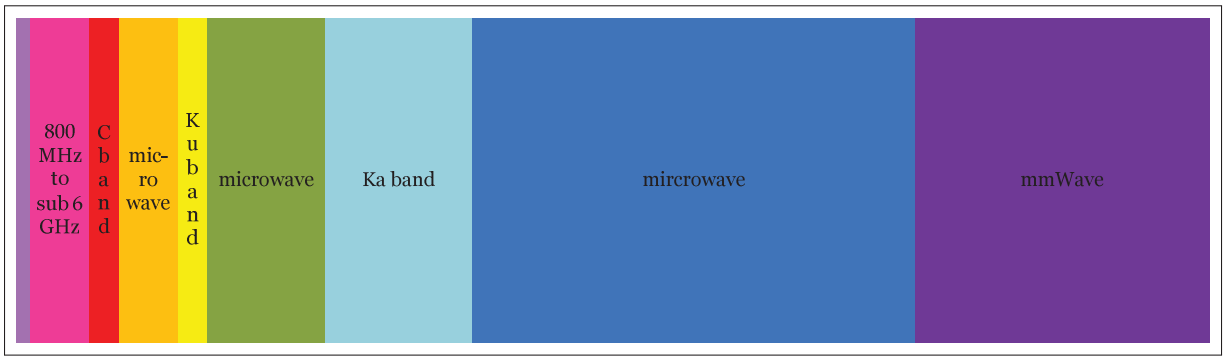

Figure 1.2 Available spectrum for WB accommodation.

Table 1.1 Simulation parameters

Backhaul spectrum features	Benefits	Limitations	Applications
Sub-6 GHz 800 MHz–6 GHz Licensed NLOS 70 Mb/s @ 20 MHz Urban: 1.5–2.5 km Rural: 10 km @ 3.5 GHz	No additional spectrum No new hardware required Easy O and M Wider coverage Antenna alignment not required Low attenuation	Limited spectrum High cost spectrum Interference issues	Low mobility scenarios Rural and urban areas Conversational voice and video (live streaming), real-time gaming
Microwave 6–60 GHz Licensed LoS 1 Gb/s+ 2 ∼ 4 km	High capacity Medium coverage High directivity	Additional spectrum cost Hardware cost Require antenna alignment High attenuation	Urban and rural areas Real-time as well as non-real-time services
Millimeter-wave 60 GHz Unlicensed LoS 1 Gb/s+ ∼ 1 km	Bulk of unused spectrum High capacity Low coverage High directivity Small form factor Zero spectrum cost Noise limited	Hardware cost Multi-hopping required High attenuation Multiple antennas required Require antenna alignment	Dense urban areas Real-time as well as non-real-time services
Millimeter-wave 70–80 GHz Light licensed LoS 1 Gb/s ∼ 3 km	Bulk of unused spectrum High capacity Low coverage High directivity Small form factor Noise limited	Hardware cost Multi-hopping required High attenuation Multiple antennas required Require antenna alignment	Dense urban areas Real-time as well as non-real-time services
TV White Space 600–800 MHz Unlicensed NLoS 18 Mb/s 1 ∼ 5 km urban	Antenna alignment not required Wider coverage Low attenuation	Primary user constraints Hardware cost Opportunistic availability Interference issues	Sparse areas Conversational voice and video (live streaming), real-time gaming
Satellite 4–6, 10–12, 20–30 GHz Licensed LoS 2–10 Mb/s downlink 1–2 Mb/s uplink	Wider coverage Supports high mobility Ubiquitous coverage	Additional spectrum cost Hardware cost Antenna alignment issues Jitter, time delay	Rural and remote areas Mobile situations Buffered streaming

1.2 Challenges and Existing Solutions

1.2.1 Fundamental Challenges of PtMP WB Small Cell HetNets

The Bottleneck of Backhaul and Access Transmission Relationship

In WB small cell HetNets, it is important to concentrate on two activities of small cell BSs, which are: the transmission/reception of backhaul data to/from the core networks via the WB links, and the transmission/reception of these data to/from the intended users via WA links. It is apparent that the small cell BSs cannot transmit/receive more WA data rate than what they receive/transmit, denoted as WB data rate, from/to the core networks. The fundamental design of WB small cell HetNets must allocate the radio resource such that the relationship where WB rate must be greater than or equal to WA rate are always maintained (Wang *et al.*, 2016, 2015a). However, such constraints imply a limitation on the overall system performance due to the following reason. When the condition of WB channels are bad, but the condition of WA channels are good, the resource allocation on the WA link must be accordingly reduced to let the WA rate be no greater than the WB rate. This means that WA channel is underexploited. In general, the fluctuation of WB rate, in responding to time-varying channel, enforces the achievable WA rate to accordingly vary. In other word, it imposes a finite upper bound of the access rate in the entire WB HetNets. This is a bottleneck compared to the conventional wired backhaul networks, where channel condition remains temporally constant. In this case, wired backhaul rate is often high so that the upper bound on the access link is neglected and WA channel can always be fully exploited.

Co-channel Interference

WB operated in the sub–6 GHz spectrum bands envisages the co-channel interference issue, especially when the network is dense and the spectrum resource is scarce. The small cells in WB HetNets which are functional with WB transmission/reception capability in this band are mainly divided into in-band and out-of-band half-duplex categories (Siddique *et al.*, 2015b).

In the in-band half-duplex case, the small cells accommodate their WB communications on the spectrum resource dedicated for the WA communications. This causes co-channel interference at each considered access and backhaul receiver on the uplink and downlink sides, which degrades both the WB and WA rates. On the other hand, in the out-of-band half-duplex small cell case, the small cells accommodate their WB communications on the non-overlapped spectrum with the WA ones. This approach has one drawback since spectrum is inefficiently utilized. Although it can avoid interference causing on the WA communications, WB transmissions on the same spectral channel must compete with each other for its best performance. Under the dense network scenario, these two in-band and out-of-band WB small cells will eventually suffer from severe interference from neighboring nodes. Therefore, an optimal strategy for interference management should be taken into account to capture the best network performance.

Signaling Overhead

Another practical challenge of WB is the excessive signaling overhead (Siddique *et al.*, 2015b). This issue occurs due to the huge amount of information required to be exchanged from the small cell to the core networks; or between the neighboring small cells in multi-hop communications via wireless channel. Signaling overhead usually comes from the demand of channel state information (CSI) transportation, frequent handover, interference management, load balancing or any cooperative communication strategy. Consequently, small cells must manage their overall system parameters in order to minimize as much signaling overhead as possible to preserve some sufficiently good wireless channels for the WB communications.

Distributed and Self-organized WB Networks

The inherent drawback of conventional small cell HetNets is its limitation on centralized computation of resource allocation. Due to the requirement of excessive signaling overhead and the emergence of dense small cell deployment, a distributed resource allocation is preferred and anticipated (Goonewardena, 2017). This requirement also applies to the WB small cell HetNets. However, due to many newly introduced design challenges such as WB-WA relationship,

co-channel interference, limited signaling overhead, developing a distributed algorithm in WB is often difficult. This is because these WB-related challenges modify the characteristics of the conventional small cell networks, so that existing solution approaches must be redesigned to cope with these new system changes.

Delay

A transmission failure in the medium access control (MAC) layer leads to a request of retransmission or announcement of drop call. In either case, a delay occurs. This also applies to the scenario of WB, where there exists the factors of channel fading and severe interference from concurrent WB and WA transmissions. These impacts might degrade the backhaul system reliability and overall performance so that characterizing the backhaul delay behavior is highly important in the WB networks (Chen *et al.*, 2015).

1.2.2 Existing Solutions

Massive MIMO combined with mmWave for WB

In recent years, the advanced extensions of spectral and spatial (antenna equipment) domains in wireless technology opens many alternatives to improve the system performance. By migrating wireless transmissions to a much higher frequency of GHz, mmWave technology was able to provide up to several Gb/s of achievable rate (Pi *et al.*, 2016; Heath *et al.*, 2016) and is obviously selected as an enabling candidate for WB. Despite the fact that it highly requires directional antennas and mostly compatible with LoS communications, mmWave WB can leverage a large amount of underutilized bandwith to achieve higher rate. For example, 1 Gb/s of backhaul capacity can be achieved over a 250 MHz in E-band (71-76 GHz and 81-86 GHz) (Gao *et al.*, 2015). Another characteristic is that mmWave only supports short range communications since mmWave signals quickly attenuate when propagating in space further than 510-700 meters in V-band or several kilometers in E-band. Therefore, inter-cell interference can be remarkably reduced, while using orthogonal spectrum allocation on these bands, intra-

cell interference is also suppressed. However, this characteristic is also a system drawback since mmWave signal cannot traverse to the remote receiver while retaining a good signal's quality. To deal with this, there are two solutions: multi-hop transmission and hybrid BS. In the multi-hop scheme, neighboring small cell BSs can communicate with each other in order to forward the backhaul data from the core networks through multi-hop to the edge small cell BSs. In the hybrid BS scheme, the macrocell and small cell BSs are assumed to flexibly switch their operation between mmWave and sub-6 GHz bands. The small cells locating near the macrocell base stations are preferred to communicate at the mmWave band due to their short distances, while the small cells at the edge may switch to the sub-6 GHz band to receive NLoS backhaul signal.

In addition, mmWave technology facilitates the installation of much more number of antennas within a small size antenna circuit without causing antenna correlation. This allows the BSs to be equipped with a large number of antenna (massive MIMO) (Larsson *et al.*). It is worth to mention that massive MIMO is also chosen as the key enabler for 5G standard due to its capability to significantly increase the area SE and EE. Largely distinguished from the classical MIMO technology, massive MIMO relies on simple beamforming techniques such as maximum ratio transmission (MRT), Zero forcing (ZF), etc. (Ngo *et al.*, 2013; Nguyen *et al.*, 2015b; Nguyen & Le, 2015), to serve multiple users within the same time-frequency resource while satisfying the 5G's QoS and quality-of-experience (QoE). According to the principle of large number theory applying on massive MIMO, BSs equipped with a large number of antenna can basically focus their transmit power towards to intended users while substantially canceling co-channel inter-user interference, thus effectively improve the user rate. In light of this, massive MIMO is very promising to embrace the WB small cell HetNets performance.

In some scenario, mmWave can be combined with massive MIMO to inherit the fruitful advantages from these two technologies to achieve the substantial improvement for a viable WB solution. Despite these benefits, mmWave based on massive MIMO also faces some challenges, which are shown as below:

- First, the cost and complexity of designing a mmWave based on massive MIMO transceiver, including the high-speed analog-digital converters (ADCs) and digital-analog converters (DACs), synthesizers, mixers, etc. are high. Therefore, using a large scale of inexpensive antennas with a finite number of expensive baseband RF chains is more appealing. However, this subsequently leads to other challenges related to precoding/combining schemes.
- Channel estimation in mmWave based on massive MIMO can be more difficult even when TDD leveraging the channel reciprocity is considered. Even for TDD-based mmWave communications, the synchronization and calibration error of RF chains to guarantee the channel reciprocity are not trivial (Chan. et al., 2006).

Half Duplex Time-Frequency Scheduling

A half-duplex transceiver refers to the one which either transmits or receives signal, but not simultaneously, at the same time-frequency resource. Most classical macrocell and small cell BSs operate in half-duplex mode. In the half-duplex system, wireless communications are divided into the uplink and downlink ones, and are accommodated separately either in time domain, named as time division duplexing (TDD), or in frequency domain, named as frequency division duplexing (FDD) (Tse & Viswanath, 2005). For the conventional wired backhaul small cell HetNets, the downlink (and uplink) communications of the small cell are allocated in the same resource with the downlink (and uplink) communications of the macrocell, since backhaul data transportation is isolated in the wired channel (Chandrasekhar *et al.*, 2008; Nguyen *et al.*, 2013, 2012a,b; Nguyen & Le, 2014b,a). However, the rise of WB communications between small cell and macrocell BSs complicates these uplink and downlink accommodations. For instance, assuming that the macrocell and small cells are operating on the downlink of the TDD system. Macrocell BS must transmit backhaul and access data to the small cell BSs and macrocell users (MUEs), respectively. However, since the small cell BSs operating on the downlink must also transmit at this stage, it is impossible to simultaneously trigger both the receiving and transmitting capabilities in half-duplex small cell BS.

To overcome this obstacle, some work in (Hoydis *et al.*, 2013; Sanguinetti *et al.*, 2015) proposes a novel variant schemes of TDD, namely RTDD and CoTDD combined with spectrum partitioning to effectively allocate the uplink and downlink of the access and backhaul transmissions. As depicted in Fig. 1.3, RTDD divides the time domain into time slots. While the macrocell BS transmits access and backhaul data to the MUEs and small cell BSs, the small cell BSs operate on the uplink side to receive signals from both the macrocell BS and small cell users (SUEs). Similarly, while the macrocell BS receives access and backhaul data to the MUEs and small cell BSs, the small cell BSs operate on the downlink side to transmit signals to both the macrocell BS and SUEs. It is worth to mention that the spectrum partitioning in RTDD is proposed to facilitate the cross-tier interference received at each small cell BS. The same arguments are also applied for the CoTDD system. In fact, these schemes have been widely investigated in the literature and embraced as a viable solution for future WB small cell HetNets.

Full Duplex

Full-duplex communications refer to the simultaneous transmissions and receptions of information in the same frequency band at the same time so as to improve spectral efficiency (Sabharwal*et al.*, 2014). BSs equipped with full-duplex transceiver are often categorized by two particular antenna configurations, namely shared and separated antenna configurations. The shared antenna configuration uses a single antenna for simultaneous in-band transmissions and receptions through a three-port circulator (c.f Fig. 1.4). On the other hand, the separated antenna configuration requires separate antennas for transmissions and receptions (Liu *et al.*, 2015). Based on these characteristics, it is obvious that WB small cells should be coupled with the full-duplex module to improve WB rate while facilitating the time-frequency division in RTDD from the half-duplex case (Pitaval *et al.*, 2015). As depicted in Fig. 1.4, small cell BSs can simultaneously transmit and receive access backhaul data to and from the intended users and macrocell BSs within the same time slot and spectrum. According to (Pitaval *et al.*, 2015), using full-duplex can also facilitate the problem related to control signal communication re-

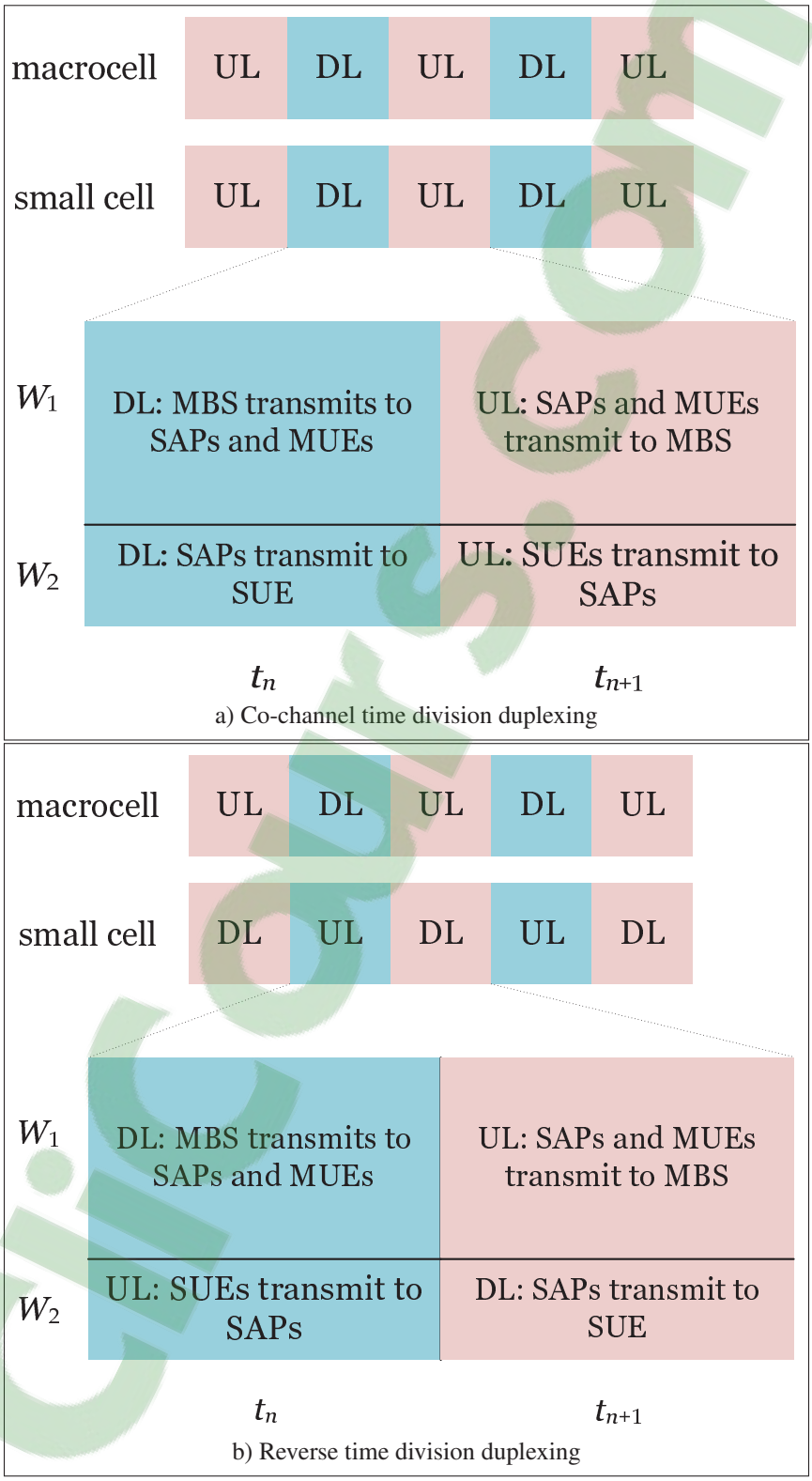

Figure 1.3 Two variant of time division duplexing user for half-duplex small cell in WB small cell HetNets.

striction, thus shortens the length of a subframe and potentially reduces transmission latency and energy consumption.

Despite the above potentials, the gains of full-duplex for WB small cell HetNets are fundamentally limited by the overwhelming self-interference, which is caused by the transmitter to its own collocated receivers. However, thanks to the three complementary techniques to mitigate self-interference in full-duplex, namely spatial isolation, RF cancellation, and digital cancellation, self-interference can be pushed down to the level of noise-floor in low-power devices (Bharadia *et al.*, 2013). Thus, by enabling small cell with the capability of full-duplex to forward backhaul data to the users, spectral efficiency can be effectively increased when self-interference is radically handled.

Cache and Buffer at Small Cell

Cache is a concept which stores (or prefetches) the potentially interested databefore it is requested by the user (Wang *et al.*; Bastug *et al.*). This data can be predicted by certain methods, such as machine learning, based on the data's likelihood from a given content distribution. In recent years, cache technology from computer network field has been adopted to the small cell wireless networks in order to reduce the overload of backhaul transmissions (Shanmugam *et al.*, 2013). In particular, a small cell can be equipped with a local cache of small size to prefetch data and then wirelessly transmit it to the small cell users, instead of requesting backhaul data to be directly transmitted from the core networks. If local cache has sufficiently large size, more data can be prefetched and cache hits ratio, defined as the number of request found in the cache over the total request, increases. In practice, small cell cache's size is rather small so that whenever there is a cache miss, data must be requested from the macrocell BS to be wirelessly transmitted to the small cell BS and then SUEs. Therefore, an effective scheme which jointly controls the cache content management together with the wireless backhaul transmissions is always interested.

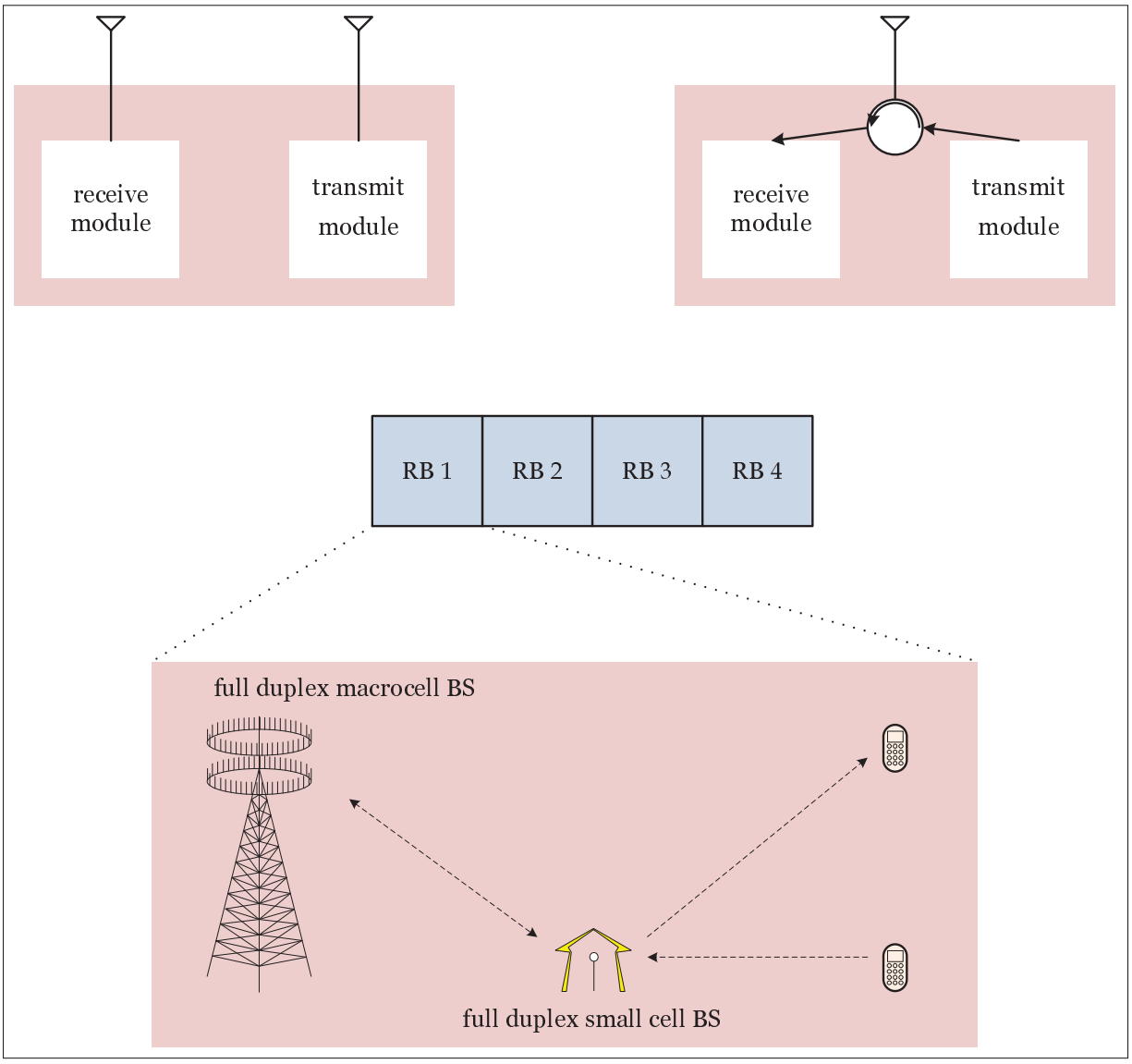

Figure 1.4 Full-duplex in WB small cell HetNets.

Different from cache, buffer is a protocol which receives causal incoming data, store it shortly within the buffer queue, and this data will be released to be transmitted to the receiver (Xia *et al.*; Zlatanov *et al.*). When buffer is equipped at a small cell, the buffered data is often received directly from the macro BS's wireless transmissions. This is different from the cache since prefetched data is typically routed from different other sources, through multiple gateways, to reach the final destination at the local cache. Although buffer and cache aim at reducing the overload for backhaul links, buffer in a much small time period, targets to store data

and release it to transmit as soon as possible so that it can save more queue space for more incoming data. This is to avoid buffer overflow and prevent buffer delay (Guo *et al.*; Georgiadis *et al.*, 2006). However, the prefetching process of cache is updated after a longer time period, only when the distribution of request content changes.

In practical system, cache and buffer can be collocated at the small cell to effectively leverage the overall WB small cell performance (Xiang *et al.*). On one hand, small cell cache serves as a long-term alternative to proactively predict and prefetch popular content in case the SUEs request it. On the other hand, there might be some requested content which is not cached, and must be wirelessly transmitted from the macro cell BS. Here, buffer is used to improve the WB data reception under the time-varying wireless channel environment in order to satisfy the QoS and QoE requirement. The joint management of cache and buffer at the small cell has recently become the ever-demanding topic which attracts many research in the field.

Small Cell Coordination

Coordination among small cells in the conventional wired backhaul small cell HetNets can significantly mitigate interference and improve link quality, and consequently boost the overall system capacity. Such a strategy was widely investigated and standardized in the 4G LTE standard as Coordinated Multi Point (CoMP) (Lee *et al.*, 2012). The small cell BSs when participating in CoMP transmissions are discriminated by two coordination levels. In the first approach, which is known as multi-point joint coordination, the cooperative small cell BSs is only capable of sharing CSI but not the transmit signals. In this case, this CoMP system resembles the multi-cell network configuration. In particular, each small cell BS serve its local users by transmitting the planned signals, while it causes inter-cell interference to the other users belonging to other small cells. Under the assumption that CSI is known at the coordinator, which is located at a central server, joint radio resources allocation (such as transmit beamforming or power) can be computed across all the small cell BSs to effectively control interference and achieve the optimal network performance.

In the second approach, also known as multi-point joint transmission, coordinated small cell BSs share both the CSI and transmit data. We can imagine these small cell BSs now forms a centralized multi-antenna BS system similar to the MISO multi-user networks to simultaneously beamform data to all users. These small cell BSs are now in full coordination mode. The differences between the joint coordination and joint transmission are illustrated in Fig. 1.5. A subtle point should be highlighted that users in this scenario still suffer co-channel inter-user interference, but with beamforming techniques, interference can be better managed to achieve a much higher gain in system performance compared to the first approach.

It is important to note that cell coordination poses extra signaling overhead burdens to the scarce small cell backhaul channel resource. The extra overhead is due to the exchanging of information between the neighboring small cell, via the wired X2/S1 interface (Yang *et al.*, 2013; Hou & Yang, 2011). In wireless backhaul scenario, there is no available wired connections between small cells or small cell to the core networks. Therefore, it is more challenging to enable the wireless message exchanging between these BSs via the wireless X2/S1 interface (Ge *et al.*, 2014). The availability of cell coordination and performance gain due to coordination depend on the quality of WB channel conditions and how many dedicated channel resources are given for the signaling overhead.

Resource Allocation

Resource allocation in WB small cell HetNets is the task to allocate radio resource such as transmit power at the BSs or users for the WB and WA links. In general, to attain a proper allocation, the first step is to formulate an optimization problem with respect to the radio resource variables in order to maximize the objective function subject to some certain constraints. The objective function and the constraints can be involved in different form of network utilities, which are the functions of the radio resources, e.g., the rate function computed according to Shannon capacity formula, the power function, the energy efficiency function, etc. In this dissertation, since WB communications are operated in the sub–6 GHz band, most WB and WA communications are in the interference-limited regime so that radio resource are tightly cou-

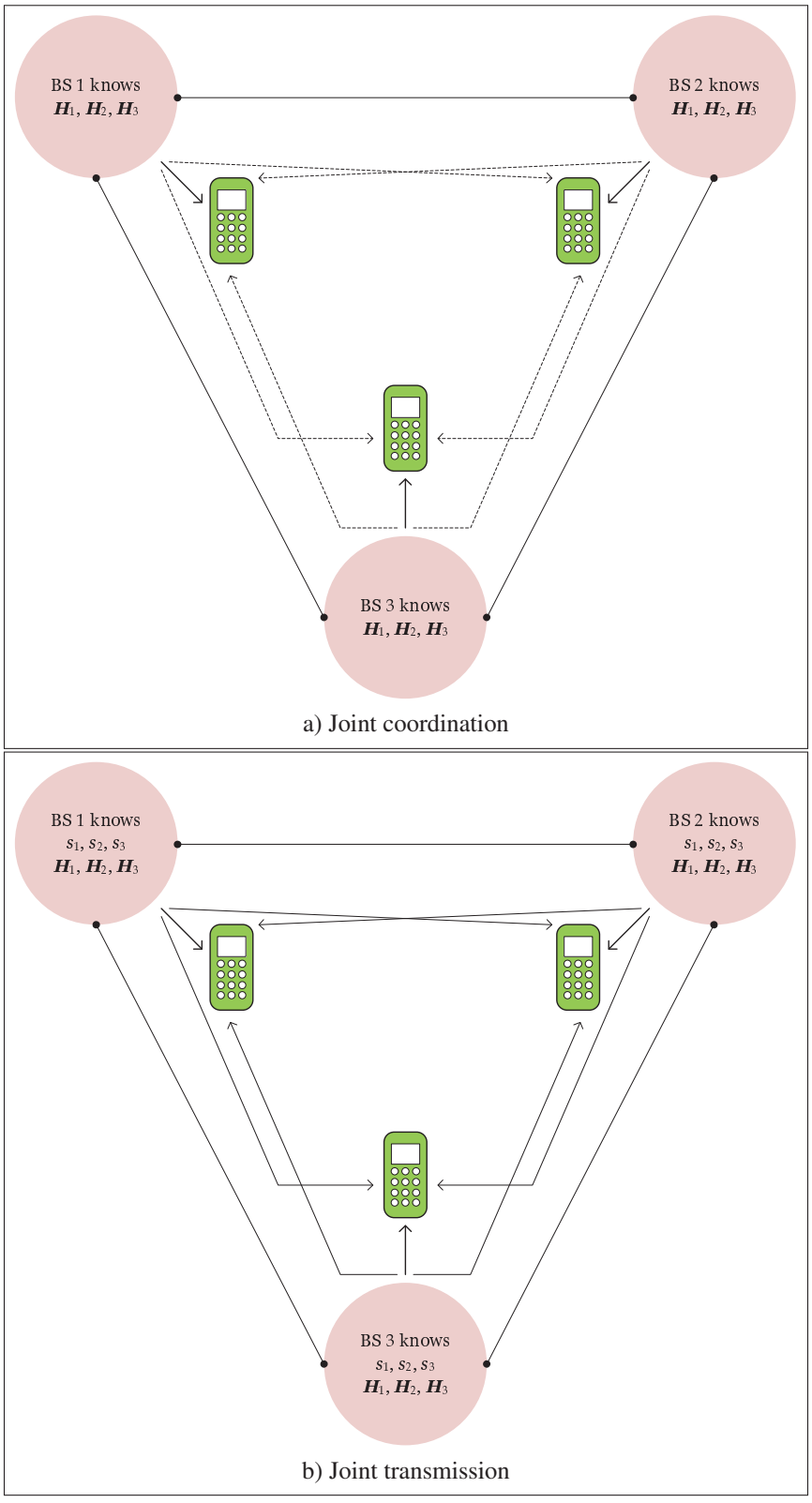

Figure 1.5 Small cell coordinated multi-point.

pled together in the formulated problems. This makes the second step, which aims at solving the optimization problems more challenging. It is important to note that since these problems are formulated on the WB small cell HetNet environment, the constraints related to WB-WA relationship are always considered, which further complicates the solution approach (Wang *et al.*, 2016; Zhao *et al.*, 2015; Hur *et al.*, 2013b).

1.3 Related Work

This section aims at elaborating the state-of-the-art related to resource allocation work which spans from a general conventional wired backhaul to WB small cell HetNets. It is worth to recall that a typical HetNet is a superimposition of one macrocell tier and multiple small cell tiers, e.g., picocell, microcell, femtocell. Each tier consists of several serving BSs and many users within its coverage. BSs that belong to each particular tier are distinguished by some key parameters such as maximum power budget and operating transmission range. In addition, users belonging to different tiers might possibly be regulated by different minimum QoS requirements. In other words, we can treat a traditional wired backhaul HetNet similar a multi-cell multi-user network. Thus solutions proposed for this network type are also applicable for the wired backhaul multi-tier HetNet scenario as well. A literature related to resource allocation optimization problems in multi-cell multi-user networks were investigated by considering various design objectives such as power minimization, sum rate maximization, and utility optimization.

1.3.1 Power Minimization

In the early stage of designing the effective solutions for power minimization problems in cellular network, power control (PC) technique proposed by (Foschini & Miljanic, 1993) was seen as the most popular algorithm not only by its superior, fast-converging behavior and compatibility for distributed implementation but also by its applicability to more sophisticated network scenarios. In PC technique, each user can independently update its power at one time instance based on the acquired signal-to-interference-plus-noise (SINR) information fed back from the

BS in the previous time instance and the predefined target SINR. As long as the target SINR is feasible, this PC algorithm is shown to converge fast to a fixed set of power values which is equivalent to the minimum aggregate transmit power of the users. In case of infeasible SINR target, user scheduling should be incorporated (Andersin *et al.*, 1995).

Inspired by the result in (Foschini & Miljanic, 1993), Rashid-Farrokhi *et al.* (1998b,a) employed the existing PC to investigate extended network schemes with multiple antennas equipped at the BS. Rashid-Farrokhi *et al.* (1998b) devised a joint computation of receive beamforming and uplink power allocation algorithm to iteratively compute the receive beamforming vector with a fixed allocated power, and then update the power based on the SINR value calculated from the recent obtained receive beamforming value. Further, Dahrouj & Yu (2010) visited a similar problem of transmit beamforming design on the downlink of a multi-antenna BS and provided a theoretical framework that shows the dual relationship between downlink transmit beamforming vector and the existing joint receive beamforming and uplink power control in (Rashid-Farrokhi *et al.*, 1998b,a). This finding is useful and important in the consideration of a very large scale of antenna number at the BS, where power can be updated using its asymptotically formula to overcome the burden of matrix inversion computation that was involved in receive beamforming updating (Huang *et al.*, 2013). On the contrary, instead of exploiting the uplink and downlink duality, designing transmit beamforming can be directly tackled by equivalently reformulating the initial optimization problem into the standard form of second order cone programming (SOCP) in which modern dedicated solvers are available for numerical solution (Liu *et al.*, 2011a; Cheng *et al.*, 2013; Gershman *et al.*, 2010). In some cases, resorting the problem into semidefinite programming (SDP) (Gershman *et al.*, 2010) and relaxing the rank one constraints shows more effectiveness than other approaches when some additional non-convex constraints are introduced into the problem (Nasir *et al.*, 2015). Once obtaining the relaxed solution from SDP, a post processing method is proposed to convert the covariance-matrix typed variables into the feasible transmit beamforming vector. In OFDMA HetNets, these problems are further considered by jointly optimizing the power allocation (or transmit beamforming) together with sub-carrier assignment which minimizes the overall power con-

sumption. Note that in a OFDMA system, one communication can be allocated on multiple sub-carriers, and one sub-carrier can be accommodated with different communications. Therefore, the mathematical formulations of these joint power and sub-carrier allocation problems often fall into the mixed integer non-linear programming category, which is generally NP-hard (as proved in (Liu *et al.*, 2011a)) and more challenging to solve. Despite these difficulties, Lopez-Perez *et al.* (2013) considered an OFDMA system where multiple users share multiple discrete sub-carriers, but they restricted that at most one user is allowed to transmit on each sub-carrier. Then, they developed a centralized and a distributed algorithm to jointly solve for the power and sub-carrier allocation. In (Cheng *et al.*, 2013), a joint optimization of beamforming, BS selection, user association which minimizes the power consumption for CoMP transmission was considered, where the authors developed an algorithm based on MISOCP and another algorithm based on SOCP combined with Inflation technique to solve for efficient solutions. This MISOCP based technique is extended to the limited fronthaul capacity cloud radio access networks (C-RANs) in (Luong *et al.*, 2017c,b,a, 2016a; Luo *et al.*, 2015). In (Li *et al.*, 2016), a problem which jointly optimizes BS operation, user association, sub-carrier assignment, and power allocation to minimize the average energy consumption was formulated. The authors developed the Steerable Energy ExpenDiture algorithm (SEED) which can efficiently solve for a local optimal solution within a polynomial time and significantly reduce power consumption against other existing schemes. In (Yadav *et al.*, 2016a,b), the authors studied the energy minimization for the hybrid small cell which can either exploit energy from the grid or the harvested renewable energy through battery.

1.3.2 Rate Maximization

Unlike the power problem case, rate-related resource allocation problems aims at maximizing the achievable rate given limited radio resource. These resources can simply be the maximum power budget at each transmitter, or/and the number of available sub-carriers to accommodate the required communications. The objective function in rate-related problems can be the maximization of sum rate, or minimum rate, of sum of log rate. It is worth to mention that

the two latter cases are known as the rate maximization problems with fairness. In (Papandriopoulos & Evans, 2009; Kha *et al.*, 2012; Chiang *et al.*, 2007), the authors proposed different approaches to develop low-complexity algorithms to solve the sum achievable rate problem. In particular, (Papandriopoulos & Evans, 2009) rewrites the non-concave rate as a function of respective power variables into a logarithm concave approximation with additionally introduced parameters and reformulate the original problem into a lower-bounded concave one. In this way, an algorithm with polynomial time complexity is proposed to recursively solve the approximated concave problem and update the corresponding parameters until it converges to a local optimal point. Akin to (Papandriopoulos & Evans, 2009), Kha *et al.* (2012); Chiang *et al.* (2007) approached the problem by rewriting it in the difference of convex (DC) and geometric programming (GP), respectively. Then, Kha *et al.* (2012) proposed to employ the first Taylor order linear approximation, while Chiang *et al.* (2007) employed the geometric-mean approximation to approximate the non-concave problem following the similar process used in (Papandriopoulos & Evans, 2009) and develop different low complexity algorithms based on their proposed methods. A good summary of these three techniques is reported in (Ngo *et al.*, 2014b), in which the authors jointly solve the sub-carriers assignment and power allocation to maximize the sum achievable rate in the OFDMA environment. In (Luong *et al.*, 2016b, 2014a), the authors considered the throughput analysis in the coexistence of IEEE 802.15.4 and 802.11 networks under unsaturated traffic. Different versions of these approaches can be found in (Wang & Vandendorpe, 2010; Tan *et al.*, 2011; Wang & Vandendorpe, 2012) and the references therein.

Moving to the multi-cell multiple antenna scheme, the design of sum rate maximization problem needs to be reconsidered more carefully since the optimization domain now deals with the functions of complex-valued beamforming variables, which is more complicated than real-valued power variables in the single antenna case. To convey this, Tran *et al.* (2012) invokes the framework of sequential parametric convex approximation (SPCA) in (Beck *et al.*, 2010) to propose a low complexity, fast converging iterative SOCP-based algorithm that attains a local optimal set of complex-valued beamforming vector and maximizes the objective function. Dif-

ferent from that, the authors in (Ha *et al.*, 2015) resort the proposed problem into a weighted minimum mean square error (WMMSE) minimization problem, in which separate block of variables, each has different physical meaning, is iteratively solved using Gauss-Seidel method when other blocks are fixed by some values (Grippo & Sciandrone, 2000). For a general setting of MIMO network, the above methods applied for MISO can also be reused with simple manipulation, which have been extensively studied in (Shi *et al.*, 2011; Nguyen & Le-Ngoc, 2014).

Similar to the above power problems, design beamforming in multi-cell can be coupled with other parameters, such as sub-carrier, or user association in different network scenarios, which raise the computational complexity. Sun *et al.* (to appear) considered a problem of jointly optimizing the beamforming together with sub-carrier allocation which maximizes the user sum rate. The authors proposed a method based on SDP combined with DC programing to relax the considered problem into a more tractable convex form, so that an efficient low-complexity algorithm can be developed to solve for a solution. In (Sun *et al.*, 2015b; Sanjabi *et al.*, 2014; Hong *et al.*, 2013), two joint optimization problems of beamforming and user association were considered to maximize a system wide utility function related to the rate formula on the downlink and uplink of a multi-cell multi-user network. The authors first proved the NP-hardness of the formulated problems. Then, in (Sanjabi *et al.*, 2014), they employed the Block Coordinated Descent method to iteratively solve their problem until convergence. In (Sun *et al.*, 2015b), they developed a low-complexity algorithm based on game theory approach to compute the local optimal solution in a distributed manner.

1.3.3 Energy Efficiency Maximization

In the recent years, research on green communication has become an active trend in wireless communication systems (Hu & Qian, 2014; Li *et al.*, 2011; I *et al.*, 2014; Soh *et al.*, 2013). Unlike conventional wireless communication networks where the metric of interest is the achievable rate computed in bits per second, achieving greener communications can be interpreted as obtaining the maximum number of information bits over the energy unit Joule,

which is referred to as the maximization of EE. In fact, according to the report in (Auer *et al.*, 2011), almost 80 % the total network energy is spent at BS sites; thus, saving more energy while attempting to maintain an acceptable achievable rate simply translates to greener and more economical communications. Motivated by the need of EE amendment, which also helps in lowering operational cost for mobile network operators and contributes to the decrease of CO_2 emission, optimally managing the radio resource is equally vital to attain the best EE.

Solving a general EE maximization problem is complicated since it carries the combination of difficulties in the two previous problem categories in the form of a nonlinear fraction. A common approach to overcome this obstacle is to transform the fractional form into linear subtractive form using Dinkelbach method (Dinkelbach, 1967) and directly apply the developed method used for the rate and power optimization to solve this equivalent problem. This solution direction has been frequently exploited in many work in the literature. Specifically, optimal beamforming and zero-forcing based beamforming designs for maximizing the EE in MC MU HetNets were studied in (He *et al.*, 2014) and (Xu *et al.*, 2014), respectively. In (Ng *et al.*, 2012), the authors addressed the joint power allocation and antenna design that maximizes the system EE in OFDMA massive MIMO networks. However, these works only aimed to develop centralized algorithms with full coordination between cells. This approach often requires high computational cost and signaling overhead since each node needs to have the full knowledge of global channel state information (CSI) to compute the joint optimal solution. To overcome these difficulties, the authors in (Huang *et al.*, 2014) proposed a decentralized algorithm based on alternative optimization to design the beamforming that maximizes the system EE. Here, limited information exchange between uncoordinated BSs are assumed. The work in (Ngo *et al.*, 2014b) employs the framework of primal decomposition in (Palomar & Chiang, 2006b) to decompose the coupled problem into smaller problems to iteratively solve and update the involved parameters until convergence. To enhance the convergence behavior of the primal decomposition method, (Tervo *et al.*, 2015a) develops a decentralized algorithm based on ADMM.

In some cases, when there are more antennas at each small cell, the ratio of power for maintaining radio frequency chains or switching the BS ON/OFF to the total power consumption becomes more important since it is now comparable to the transmit power. This motivated the authors in (Zhou *et al.*) to consider BS sleep/active or in (Tervo *et al.*, 2015b) for antenna selection to refine the achieved EE in comparison with fixed circuit power. In (Luong *et al.*, 2017c), a joint optimization of beamforming, BS sleep/active selection, together with user association is considered to maximize the system EE under the limited fronthaul capacity C-RANs. Moreover, at the WB small cell BS, the energy consumed in decoding the collected backhaul data from the MBS cannot be ignored. This is due to the fact that small cell BS have small range of operation and the energy consumed in the decoding process is eligible in comparison to the one used in transmitting the data (Luong *et al.*, 2018b,a, 2017b, 2011; Rubio & Pascual-Iserte, 2014). For the downlink of WB small cell HetNets, each SAP relates its decoding rate proportionally to the amount of power to decode the backhaul messages before forwarding them to its own users. Hence, this decoding power is fundamentally important (Rubio & Pascual-Iserte, 2014) and should be considered in the problem of EE optimization. More studies on energy efficiency in WiFi-cellular HetNets can be found in (Luong *et al.*, 2016c, 2014b)

1.3.4 Wireless Backhaul Communications

The problem of radio resource allocation optimization in WB networks has been started over ten years. However, the intensity of this research topic just recently grows faster and attracts more researchers in the 5G networks. One of the first works on WB was proposed in (Cao *et al.*, 2007) in 2007. In this work, the authors consider a joint design of routing, medium access control scheduling and beamforming in order to achieve the minimum power consumption performance while still guarantee some minimum achievable rate requirement. In this work, WB communication are planted in an orthogonal resource dimension (e.g., spectrum) to the wireless access channel so that there is no cross interference between these two communication types. To overcome the difficulty of coupling factor in the non-convex problem, the authors propose to decouple the first problem into smaller subproblems corresponding to each

layer and iteratively solve each until convergence. Similarly, (Lee *et al.*) proposed different methodologies and strategies to activate the multi-hop WB in medium access control layer to meet the proposed criteria such as call admission or delay.

Moving to the perspective of 5G, the work in (Hur *et al.*, 2013b) is found as one of the first work to propose the solution for WB in small cell networks. Hur *et al.* (2013b) proposed the use of outdoor mmWave communications for WB between small cells. To overcome the impairment of mmWave propagation in outdoor environment, this work proposes to use massive MIMO and beam alignment to mitigate the sensitivity of the narrow beam to the movement and wind-induced factor, which are considered as the main causes to performance degradation of mmWave communication. Motivated from this work, Singh *et al.* considered the scenario of self-backhauled mmWave small cell, where a part of the small cell BSs, namely A-BSs, have wired backhaul and the rest backhaul wirelessly to A-BSs. The authors of (Niu *et al.*, 2017) presented an optimization problem of scheduling and power control which maximizes the mmWave WB energy efficiency. The analysis on the coverage and rate distribution are characterized where it is concluded that in dense network, increasing the fraction of A-BSs improves the performance of the network while the user rate coverage saturates if the density of the A-BSs is kept constant.

In the scenario of reuse the sub–6 GHz spectrum bands for WB, (Wang *et al.*, 2016) proposes a model of large scale antenna at the MBS to efficiently mitigate interference causing to their users and SBSs on the wireless access and backhaul links, respectively. In addition, the WB communications between MBS and SBSs are accommodated on an orthogonal time/frequency domain by employing the reverse time division duplex and spectrum splitting as a strategy of interference management. A joint optimization problem of spectrum allocation and user association is formulated such that the sum log-rate of the network is maximized. Interestingly, another work from (Sanguinetti *et al.*, 2015) also follows the same spirit as in (Wang *et al.*, 2016). In (Sanguinetti *et al.*, 2015), the authors setup an almost similar system model of massive MIMO and interference management with reverse time division duplex and spectrum splitting to let the BSs communicate with two types of user: low and fast mobility. Unlike

(Wang *et al.*, 2016), an analytical framework to evaluate the consumed power at the large regime of antenna are conducted by some simple formulas to give some insights to the system behavior. On the other hand, to further investigate the effect of interference management, the work in (Thomsen *et al.*, 2014; Phan *et al.*, 2015) introduce the idea of applying network coding to develop a new communication strategy between SBSs and MBS. With a scenario of two SBSs and two users, an optimization problem is formulated and solved to give rise to the optimal setup of the proposed strategy. To better exploit the available spectrum, (Li *et al.*, 2018; Siddique *et al.*, 2017) introduced a hybrid strategy of out-of-band and in-band full-duplex and formulated a problem which optimizes the spectrum allocation under the proposed scheme.

The problem of optimally deploying BS with WB capability in space is also of great interest. In general, this deployment issue is related to the problem of minimizing the cost function with some guarantee on the QoS requirement. In (Zhao *et al.*, 2015), the meaning of cost minimization is translated into the definition of how many maximum WB SBSs are admitted in the considered network to achieve some user minimum rate requirement on the DL. As such, the number of SBSs underlaid with macrocell that can be admitted to the network is optimized. In (Islam *et al.*, 2014), two-tier HetNets are considered where SBSs that cannot access the MBS are assisted by single-hop wireless link aggregator nodes to ensure backhaul connectivity. A joint cost function of node locations, power control, channel scheduling and routing is optimized to achieve the best deployment scenario of the aggregator nodes in spatial domain.

Recently, a framework of stochastic geometry is applied to analyze the energy efficiency behavior of WB two-tier HetNets in (Yang *et al.*, 2016a). By assuming that all the BSs and users are distributed according to different independent Poisson Point Processes, this paper computes the average achievable rate and energy efficiency on the uplink and downlink and shows that WB can be significantly more energy efficient in two tiers than in one tier.

1.4 Motivations and Objectives

The evolution of wireless technology towards its future generations will always seek for a more economical, efficient but viable and practical solution. Denser network deployment to cope with the exponential growth of mobile equipment is inevitable. This raises an increasing importance of designing a more effective WB solution for the future networks. Motivated to overcome the raised challenges together with bridge the gap from the benefits and drawbacks of the existing solution discussed above, the main objective of this thesis is to address the fundamental challenges and achieve the optimal performance of the overall WB systems in terms of sum achievable rate and network energy efficiency. To achieve these goals, this dissertation jointly optimizes the allocation of multi-dimensional resource, such as transmit power, beamforming, buffered content, small cell ON/OFF selection, etc., which maximizes the considered network utility function while always explicitly respecting the WB and WA relationship constraints.

In general, the formulated optimization problems when considering the WB and WA relationship are generally difficult and provably NP-hard. According to the mathematical terminology, there exists no algorithm which can solve a NP-hard problem within a polynomial time. To solve for global optimal solution of such problem requires a exhaustive search based algorithm. However, the computational complexity of this method exponentially scales with the problem size, which is not applicable for real-time wireless application. Motivated by this, this dissertation aims at developing a pragmatic low-complexity algorithm to solve for a high-quality sub-optimal solution.

1.5 Highlighted Novel Contributions

The novel contributions of this dissertation are elaborated in each chapter as follows:

- The main contribution of Chapter 2 lies in the realization that the RTDD combined with spectrum partitioning which couples the uplink and downlink WB and WA communications. Consequently, it proposes to simultaneously optimize the resource allocation on the

uplink and downlink sides. In fact, this important characteristics must not be overlooked as in (Wang *et al.*, 2016), since if the resource allocation is only for the downlink (or the uplink) side without respecting the uplink (or downlink) WB–WA relationship, this could jeopardize the operation of WB small cell HetNets under the proposed RTDD. Another contribution of this chapter lies in the solutio approach. In particular, a novel result which extend from (Nguyen *et al.*, 2015a) is derived to approximate the generalized exponential cone constraint into a system of conic constraints. This derivation supports the developed algorithm in reducing more computational complexity compared to the existing approaches, which is helpful in achieving an practical and wireless-application-oriented solution.

- The main contribution of Chapter 3 lies in the novel power consumption model, which reflects the dependence between the consumed power and the decoding backhaul data from the received signal. Compared to the existing linear power consumption model in the literature, this newly proposed model offer an appropriate method to quantify the overall power consumption. This provides a more precise solution of resource allocation which optimizes the system energy efficiency. Another important contribution is that this work is the first one which claims and rigorously proves the NP-hardness of the formulated EE problem. It should be noted that the derived proof does not only apply for this particular problem of EE in WB HetNet, but is also useful to prove the NP-hardness of an EE problem in a general interference-limited regime of wireless networks.
- The main contribution of Chapter 4 is the analysis and design which study the improved impact of small cell buffer on WB HetNet over multiple time slots. Due to the wireless channel variation, there is some time instances where in the first time slot, the channel of WB is better than WA, and in the second time slot, the channel of WB is worse than WA. If we allocate the resource to achieve a strictly higher WB data rate than the achievable WA one in the first time slot, the backhaul data which is not used by the WA transmission will be stored in the buffer queue. Without the buffer, this data will be dropped. In the second time slot, we recall that the channel condition in the second time slot leads to a poor WB data rate, and therefore WA communication cannot achieve a high rate. Small cell with buffer can overcome this rate-related obstacles by exploiting the buffered data together with

the received data from the WB transmission. Without the buffer, the small cell can only the receive data from the WB transmission without getting any extra data from the buffer, which obviously results in a less performance than the buffer case. The contribution of this work is to effectively integrate this intuition by jointly designing the radio resource allocation together with the buffer usage according to the changes of channel condition in order to maximize the user sum rate over time.

- The main contribution of Chapter 5 lies in the novel cooperative NOMA scheme employed on the WB and WA communications. The novelty of this scheme is the proposed cooperation between the small cell BSs as the consequence of the successive interference cancellation (SIC) NOMA protocol from the WB communications. This work is the first one which proposes such scheme, as well as the approach to optimize the SIC decoding order in the MISO NOMA scenario. By analyzing the potential benefits of a proper decoding order decision and the cooperation policy between the small cell BSs, a joint design of resource allocation together with SIC decoding order and small cell cooperation policy is considered to maximize the two objectives of interest.
- The main contribution of Chapter 6 lies in the novel proposal of employing multiple UAVs as the flying WB small cells to consider the UAV-assisted PtMP WB networks. Moreover, this work revisits the impact of cooperative NOMA in Chapter 5 on this new UAV-assisted PtMP WB networks. A subtle observation is that UAV can flexibly change their spatial position, which in turn affects the WB channel's discrepancies and consequently change the performance of the proposed cooperative NOMA scheme. Based on these analyses, a joint design of UAV placement together with radio resource allocation, SIC decoding order, and cooperation policy is formulated to maximize the user sum rate.

1.6 Methodology

Most of the optimization problems formulated throughout Chapter 2 to 6 retain in a non-linear non-convex form, where this form's NP-hardness was proved in (Luo & Zhang, 2008; Liu *et al.*, 2011a). This implies a general difficulty in solving these problems for an optimal solution. The

difficulty is that one must customize a high-complexity based on exhaustive search method to exhaustively seek for a possible optimal solution over excessive data set within the solution domain. This is not prominent for a typical wireless application. In light of this, the main methodology of this dissertation is to invoke the recent results on SCA techniques to develop a more appealing low-complexity algorithms, which can efficiently solve for a high-quality sub-optimal solution. The highlighted benefit of this employed SCA approach is that it can always achieve a solution better than any existing method in the literature. In general, the framework to solve a formulated problem involves in the following basic steps:

- **Step 1:** Transforming the original problem into an equivalent but more tractable form, which will be more amenable to the application of SCA method.
- **Step 2:** Appropriately employing the SCA technique on the non-convex parts to approximate the equivalently transformed problem into a sequence of convex approximated ones, where each of them is a convex approximate of the original non-convex problem. It is important to note that the employed approximation step introduces some additional input parameter to each approximated problem.
- **Step 3:** Further deriving a new controlled-accuracy based on SOCP approximation method to manipulate each approximated problem into a standard form of conic programming, which is more amenable to the dedicated commercial solvers.
- **Step 4:** Developing an algorithm to iteratively solve the sequence of the convex approximated problems and update the introduced parameters in Step 2 until convergence
- **Step 5:** Deriving the computational complexity analysis for each developed algorithm.

CHAPTER 2

RESOURCE ALLOCATION IN TWO-TIER WIRELESS BACKHAUL HETEROGENEOUS NETWORKS

Tri Minh Nguyen[1], Animesh Yadava[2], Wessam Ajib[2], Chadi Assi[3]

[1] Département de Génie Électrique, École de Technologie Supérieure,
[2] Département d'Informatique, Université de Québec à Montréal
[3] Concordia Institute for Information Systems Engineering, Concordia University

This article was published in *IEEE Transaction on Wireless Communications* in October 2016 (Nguyen *et al.*, 2016b).

2.1 Introduction

Network densification is becoming one of the architectural shift towards the emerging 5G wireless networks to support the ever-increasing needs of data and services from users (Bhushan *et al.*, 2014; Andrews *et al.*, 2014). By densely deploying more low-power small cells in multi-tier HCNs (Ghosh *et al.*, 2012), the operators are able to deliver seamless coverage to the cell-edge users and enhance the area spectral efficiency. However, since data destined to the local users is often received from the core network through backhaul transmissions, this also means that a denser backhaul connection establishment with the core network is required.

Conventionally, small cell networks (SCNs) rely on fast and reliable fiber links to establish the transmissions with the backhaul center. However, for SCNs densification, a fiber link solution may not be a cost-effective alternative. Thus, a novel architecture of wireless backhaul network which concurrently accounts for cost saving and convenient installation has emerged. Wireless backhaul (WB) technology (Ge *et al.*, 2014; Hur *et al.*, 2013b) is introduced as a viable solution that reuses available spectrum to communicate backhaul data via the wireless X2 interface. By replacing (or coexisting with) the fiber connection to exchange backhaul data, the cellular operators can rely on WB technology to reduce the installation cost and overcome the difficulties of backhaul deployment at some rural areas. Although providing a promising cost-effective

alternative, WB communications (WBC) must guarantee the high speed, long range and reliable backhaul transmission requirements under a low latency constraint. Furthermore, WBC needs to share the available spectrum and power resources with the forward communication. Thus, efficient resource allocation becomes a challenging task (Zhang *et al.*, 2014; Zhang *et al.*, 2015; Zhang *et al.*, 2015).

2.1.1 Related Work

Multiuser multiple-input multiple-output (MU-MIMO) has emerged in wireless cellular networks as it exploits both the advantages of MIMO and MU diversity gains. In MU-MIMO environment, establishing WBC concurrently with existing wireless access communication (WAC) is simply equivalent to introducing more users with simultaneous transmissions into the network. In contrast, these WB nodes act as special users that are required to attain sufficiently high data rates to support their access links. It is worth to note that integrating multiple antennas and activating the WBC and WAC simultaneously on the same radio resource makes the resource management more complicated. In fact, an appropriate radio resource allocation (such as power, frequency, time, and antennas) plays a crucial role in enhancing the wireless network performance. This has been indeed an active area of research over the last decade. In particular, numerous resource allocation problems for the MU single-cell MIMO systems have been investigated considering various design objectives such as power minimization (Rashid-Farrokhi *et al.*, 1998b; Dahrouj & Yu, 2010; Tran *et al.*, 2014), sum-rate maximization (Tran *et al.*, 2012), and utility optimization (Liu *et al.*, 2011b). The early work in (Rashid-Farrokhi *et al.*, 1998b) addresses the problem of uplink (UL) power minimization. The authors employed the Foschini-Miljanic power control protocol to devise an iterative receive beamforming and transmit power update. Dahrouj & Yu (2010) revisited the same problem on the UL and downlink (DL) and derived a different algorithm based on the DL–UL duality relationship. Moving to the problem of maximizing the network sum-rate, the authors in (Tran *et al.*, 2012) developed an efficient algorithm based on the successive convex approximation combined with second order cone programming (SOCP), which is numerically shown to quickly converge to

the optimal solution. A very good survey of various network utility optimization problems is reported in (Liu *et al.*, 2011b), where each problem is conducted with complexity analysis and efficient algorithm.

Communicating backhaul data on wireless link was first introduced in (Viswanathan & Mukherjee, 2006; Cao *et al.*, 2007). In these works, design criteria pertaining to medium access control (MAC) layer are taken into account for the optimal system design. In (Viswanathan & Mukherjee, 2006), a linear optimization framework is proposed for the optimum routing and scheduling with the objective of throughput maximization in a wireless mesh network. Whereas in (Cao *et al.*, 2007), a cross-layer design is considered for multi-hop wireless backhaul networks. Recently, the concept of WBC applying to HCNs is revisited in the deployment of small cells. To leverage the potential of WBC in multi-tier cellular networks, the work in (Sanguinetti *et al.*, 2015) proposed the reverse time division duplexing (RTDD) model to accommodate transmissions of different tiers in each time slot reversely where in each time slot WBC and WAC are split into two orthogonal resource blocks (RBs) to exchange data. Based on RTDD model, Wang *et al.* (2016) considers the joint bandwidth allocation and user association that maximizes the achievable DL sum rate of small cell when the macro base station is equipped with a large antenna number (massive MIMO). In (Zhao *et al.*, 2015), the authors studied the admission control of small cell access points (SAPs) in order to permit WB to serve as many SAPs as possible while guaranteeing predetermined quality of service (QoS) rates. Apart from these works, in (Thomsen *et al.*, 2014; Phan *et al.*, 2015), the authors introduced the idea of wireless backhaul via the framework of relay and network coding.

2.1.2 Contributions

In this paper, we propose a novel interference management based on RTDD and orthogonal spectrum splitting for the WB HCNs that caters to both UL and DL transmissions. In particular, WBC between multi-antenna MBS and single antenna SAPs are enabled among the conventional WAC of MBS-MUEs and SAPs-SUEs. In order to share available spectrum resource and mitigate cross-tier interference, we propose to arrange the WBC and WAC according to the

setting of RTDD and spectrum splitting as in Fig. 2.2. Unlike the related work in (Wang *et al.*, 2016), our work realizes that the achievable rate of UL and DL transmissions are coupled under the proposed RTDD interference management to further consider the joint design of transmit beamforming, power allocation and spectrum splitting factors that takes into account both UL and DL transmissions. The contributions and novelties of this paper are summarized as follow:

- We formulate a novel problem of maximizing the achievable sum rate of small cells on both UL and DL. We consider two distinguished levels of bandwidth partition located at two consecutive time slots to reflect the coupling factor between UL and DL transmissions in designing the system parameter.
- The formulated problem is non-convex. To solve it optimally, we first apply the Branch-and-Bound exhaustive search algorithm to find the global optimal solution. Then, we propose a low complexity efficient iterative joint beamforming and resource allocation optimization (JBRAO) algorithm. Specifically, we transform the original problem into a tractable form and invoke the framework of sequential parametric convex approximation (SPCA) in (Beck *et al.*, 2010) to approximate the original non-convex problem by its lower bound concave problem.
- We study the network performance in the presence of imperfect channel state information (CSI). In particular, we formulate a robust optimization problem in which the proposed low complexity algorithm can be employed to achieve its solution.
- The approximation of the optimization problem naturally leads to the set of second order cone (SOC) and general exponential cone constraints. Solving the convex exponential cone is often more complicated and requires modern convex programming solver to achieve the results. Motivated by the perspective of SOCP approximation in (Ben-Tal & Nemirovski, 2001a), we arrive at a novel approach that can approximate the problems with general exponential cone constraints into a form that only involved with SOC constraints.

The rest of this chapter is organized as follows. Section 2.2 introduces the system model and formulates the optimization problem. Section 2.3 and 2.4 develop a Branch-and-Bound algorithm and a low complexity algorithm based on SPCA to solve the formulated optimization

problem. In Section 2.5, we revisit the convex approximated problem to transform general exponential cone constraint into SOC form. Section 2.6 presents our numerical results and discussions. Finally, the concluding remarks of the chapter are given in Section 2.7.

2.2 System Model

2.2.1 Spatial model

Consider a two-tier HCN consisting of a MBS in the macrocell tier and F SAPs in the small cell tier as shown in Fig. 2.1. The MBS is equipped with N antennas to communicate with M macrocell users (MUEs) and SAPs. Each SAP and user is equipped with only single antenna. For simplicity, we assume that each SAP serves only one SUE. The general case of multiple SUEs in one small cell can be extended by following the same framework. In this model, besides the transmissions between the pair of MBS-and-MUE and SAP-and-SUE, we further consider the operation of WBC of the SAPs on the UL and DL to communicate with the MBS, where the SAPs are allowed to transmit and receive backhaul data concurrently with the macrocell on the same spectrum. It is worth to note that this spectrum is often available for conventional communications, which is referred to as WAC for convenience. Therefore, the communications in our model are categorized into: WBC between the SAP-and-MBS and WAC between the MBS(or SAP)-and-MUE (or SUE), respectively.

2.2.2 Reverse time division duplex (RTDD)

In this model, we assume that the system exploits the channel reciprocity property, where the channel condition on the UL and DL time slot is the same within a coherence time. Thus, to facilitate the CSI estimate acquisition, we consider the network operating in the TDD mode, where the channel gain can be estimated via orthogonal pilot transmissions over the time domain(Tse & Viswanath, 2005). Recently, some modified versions of TDD, namely RTDD (Fig. 2.2), studied in the context of (Hoydis *et al.*, 2013) and (Sanguinetti *et al.*, 2015), are considered as a novel option to accommodate WBC in the two-tier HCNs. In particular, (Hoy-

dis *et al.*, 2013) recalibrates the setting of TDD system to reverse the UL/DL time slot in two tiers so that when the MBS transmits its signals to the MUEs on the DL, each SUE transmits its signal to its serving SAP on the UL. A similar transmission protocol applies to the UL of the macrocell and DL of the small cell. However, when WBC coexists with macrocell and small cell transmissions as in (Sanguinetti *et al.*, 2015), assuming in-band half-duplex SAPs, an additional resource dimension such as frequency should be taken into account to avoid the self-interference arising at the SAPs.

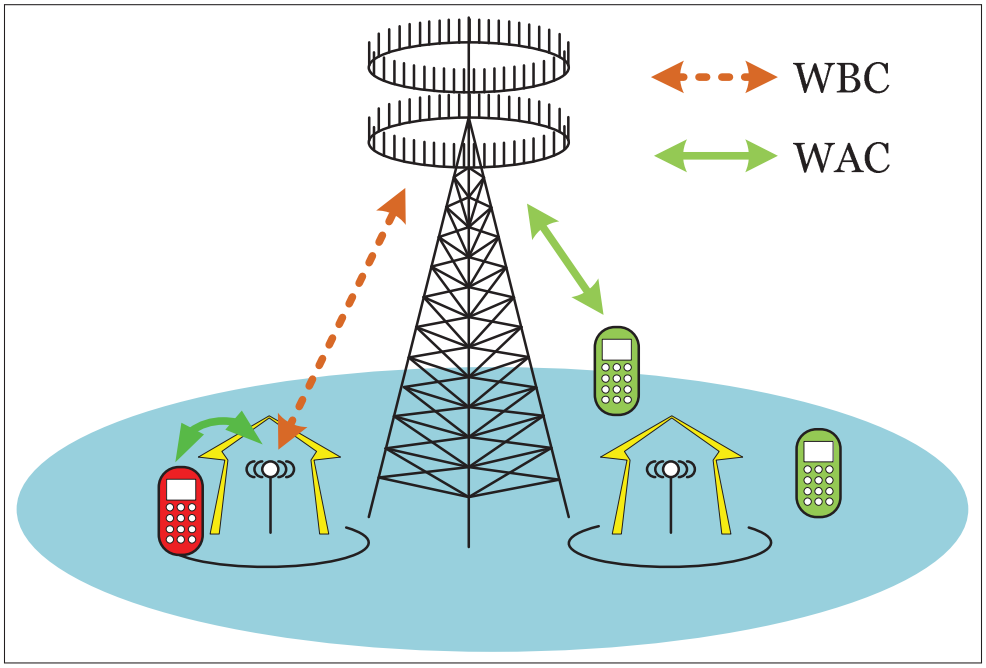

Figure 2.1 Two-tier HCNs with wireless backhaul communications.

2.2.3 Signal Model

We assume the channel is flat over the spectrum and time-invariant within the coherence time T_c where T_c is much larger than the duration of two consecutive time slots. Therefore, we only consider the performance of the network in two consecutive time slots, named as, UL and DL time slots. For simplicity, we denote the duration of the DL time slot T_{α^d} as the period which

accounts for macrocell DL and small cell UL; and the duration of the UL time slot T_{α^u} as the period which accounts for macrocell UL and small cell DL transmissions.

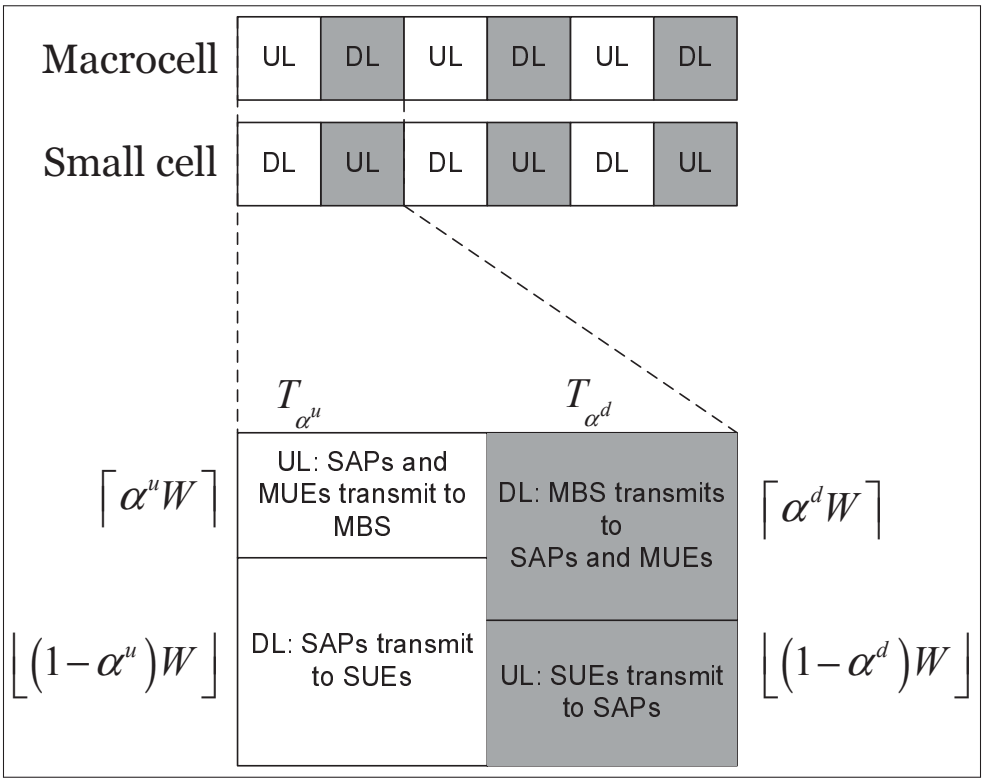

Figure 2.2 Two-tier HCNs with WBC and RTDD setting.

2.2.3.1 Macrocell DL - Small Cell UL

This subsection considers the time slot dedicated for macrocell DL and small cell UL transmissions, i.e., T_{α^d}. The spectrum is assumed to be divided into W resource blocks (RBs) of 1 Hz each. The W RBs are split into two parts with the splitting factor of α^d , where α^d takes some real value between 0 and 1 so that both $\alpha^d W$ and $(1-\alpha^d)W$ should take some integer values from the available W RBs. On $\alpha^d W$ RBs, the MBS transmits to its MUEs and the SAPs via WAC and WBC, respectively, while in the remaining $(1-\alpha^d)W$ RBs, each SUE simultaneously transmits data to its serving SAP. By bundling the set of SAP and MUE indices into one group, denoted as $\mathscr{G} = \{\mathscr{F}, \mathscr{M}\} = \{\{1,\ldots,F\},\{F+1,\ldots,F+M\}\}$ where SAP indices

are from $1,\ldots,F$ and MUE indices are from $F+1,\ldots,F+M$, the received signal, within T_{α^d} in $\alpha^d W$ RBs, at the jth receiver is (Tse & Viswanath, 2005):

$$y_j = \mathbf{v}_j^H \mathbf{h}_j x_j + \sum_{k\neq j}^{F+M} \mathbf{v}_k^H \mathbf{h}_j x_k + n_j, \tag{2.1}$$

where $\mathbf{h}_j \in \mathbb{C}^{N\times 1}$ is the channel state vector which includes fading gain and pathloss components and $\mathbf{v}_j \in \mathbb{C}^{N\times 1}$ is the beamforming vector from the MBS to the jth receiver. x_j is the message intended for the jth receiver with unit average power, i.e., $\mathbb{E}\left\{x_j x_j^*\right\} = 1$. n_j is the additive white Gaussian noise (AWGN) vector at the jth receiver. The fading gain and noise are modeled as independent and identically distributed (i.i.d.) circularly-symmetric complex Gaussian random variables from the distribution $\mathscr{CN}(0,1)$ and $\mathscr{CN}(0,N_{\alpha^d})$, respectively, where $N_{\alpha^d} = \alpha^d W \sigma_0$ is the noise power over the allocated RBs and σ_0 is the noise power spectral density (PSD). Treating interference as noise, the achievable rate at the jth receiver is given by $\mathsf{R}_j^d = \alpha^d W \log(1+\Gamma_j^d)$ where

$$\Gamma_j^d = \frac{\left|\mathbf{v}_j^H \mathbf{h}_j\right|^2}{\sum\limits_{k\neq j}^{F+M} \left|\mathbf{v}_k^H \mathbf{h}_j\right|^2 + \alpha^d W \sigma_0}, \tag{2.2}$$

where we denote the set of transmit beamforming at the MBS as $\mathbf{v} \triangleq \left[\mathbf{v}_1^T,\ldots,\mathbf{v}_{F+M}^T\right]^T$ for later usages. Note that in this paper, we compute the formula of rate in nats/s/Hz, so that the base of all the $\log(.)$ function is e. On the other hand, in the other $(1-\alpha^d)W$ RBs, each SUE in the ith small cell transmits data to its intended SAP with power p_i^u. By denoting h_{ij}^u as the scalar channel coefficient from the ith SAP to the SUE in the jth small cell which includes fading gain and pathloss components and $\mathbf{p}^u = \left[p_1^u,\ldots,p_F^u\right]^T$ as the set of transmit powers of SUEs, we can write the achievable rate at ith SAP on the UL as $\mathsf{r}_i^u = (1-\alpha^d)W\log(1+\gamma_i^u)$, where

$$\gamma_i^u = \frac{p_i^u \left|h_{ii}^u\right|^2}{\sum\limits_{j\neq i} p_j^u \left|h_{ij}^u\right|^2 + (1-\alpha^d)W\sigma_0}. \tag{2.3}$$

2.2.3.2 Macrocell UL - Small Cell DL

Similarly, this subsection considers the time slot dedicated for macrocell UL and small cell DL. In T_{α^u}, W RBs are divided into two parts by a factor α^u, where in $\alpha^u W$ RBs, the MUEs and SAPs transmit signals to the MBS through WAC and WBC, respectively, while in $(1-\alpha^u)W$ RBs, each SAP simultaneously transmits its signal to its SUE. Similar to the previous scenario, it is also important to constrain α^u to take some real value between 0 and 1 so that $\alpha^u W$ and $(1-\alpha^u)W$ can take some integer values from the W RBs. By applying the above notation for the set of indices at the SAPs and MUEs to denote the set of transmit power of the SAPs and MUEs as $\rho = [\rho_1, \ldots, \rho_F, \rho_{F+1}, \ldots, \rho_{F+M}]^T$, we can write the received signal at the MBS within T_{α^u} period in $\alpha^u W$ RBs as

$$\mathbf{y} = \sum_{j=1}^{F+M} \mathbf{h}_j \sqrt{\rho_j} s_j + \mathbf{n}, \tag{2.4}$$

where $\mathbf{n} \in \mathbb{C}^{N\times 1}$ is the AWGN vector at the MBS with distribution $\mathscr{CN}(0, \alpha^u W \sigma_0 \mathbf{I})$, where $\mathbf{I}$ is the $N \times N$ identity matrix. s_j is the message from the MUE or SAP with unit average power, i.e., $\mathbb{E}\left\{s_j s_j^*\right\} = 1$. The minimum mean square error (MMSE) receive beamforming $\mathbf{w}_j = d_j \left(\sum_{k\neq j} \rho_k \mathbf{h}_k \mathbf{h}_k^H + \alpha^u W \sigma_0 \mathbf{I}\right)^{-1} \mathbf{h}_j$ is applied at the MBS to detect the transmitted signal from the jth SAPs and MUEs, where d_j is the normalized coefficient. Again, by treating interference as noise, the achievable rate at the jth user can be presented as in (Tse & Viswanath, 2005)

$$\mathsf{R}_j^u = \alpha^u W \log(1+\Gamma_j^u) = \alpha^u W \log(1+\rho_j \mathbf{h}_j^H \Sigma_j^{-1} \mathbf{h}_j), \tag{2.5}$$

where $\Sigma_j = \sum_{k\neq j}^{F+M} \rho_k \mathbf{h}_k \mathbf{h}_k^H + \alpha^u W \sigma_0 \mathbf{I}$ is the covariance matrix, with $j \in \mathscr{G}$. On the other hand, in the other $(1-\alpha^u)W$ RBs, each ith SAP transmits data to its intended SUE in the ith small cell with power p_i^d. By denoting h_{ij}^d as the channel from the jth SAP to the SUE in the ith small cell and $\mathbf{p}^d = \left[p_1^d, \ldots, p_F^d\right]^T$ as the set of transmit powers of the SAPs, the achievable rate at

each small cell on the DL can be written as $\mathsf{r}_i^d = (1-\alpha^u)W\log\left(1+\gamma_i^d\right)$ where

$$\gamma_i^d = \frac{p_i^d \left|h_{ii}^d\right|^2}{\sum\limits_{j\neq i} p_j^d \left|h_{ij}^d\right|^2 + (1-\alpha^u)W\sigma_0}. \tag{2.6}$$

2.2.4 Resource allocation optimization problem

In conventional HCNs, it is always of great interest to obtain the optimal achievable rate of the small cell with some protection on the macrocell performance, e.g., MUE minimum rate QoS (Ngo *et al.*, 2014a). However, with WB, the required access rate to guarantee a reliable and high quality transmission at each SAP should be upper bounded by the backhaul rate. This applies to both the UL and DL transmissions. Thus, we are interested in formulating a problem to jointly optimize the beamforming, power allocation together with fraction of bandwidth usage α^d and α^u to maximize the sum achievable rate at the small cell on the UL and DL. Denoting $\mathbf{p} = \left[\left(\mathbf{p}^u\right)^T, \left(\mathbf{p}^d\right)^T\right]^T$ as the set of transmit power at the SUEs and SAPs in the WAC and $\alpha = \left[\alpha^u, \alpha^d\right]^T$, the optimization problem can be casted as:

$$\max_{\alpha,\mathbf{v},\rho,\mathbf{p}} \sum_{i=1}^{F}\left(\mathsf{r}_i^d + \mathsf{r}_i^u\right) \tag{2.7a}$$

$$\text{s.t.} \quad \mathsf{R}_i^o \geq \mathsf{r}_i^o, \forall i \in \mathscr{F}, \tag{2.7b}$$

$$\mathsf{R}_j^o \geq \mathsf{r}_{\min}^o, \forall j \in \mathscr{M}, \tag{2.7c}$$

$$0 \leq \rho_n \leq \bar{\rho}_f, \forall n \in \mathscr{G}; 0 \leq p_i^o \leq \bar{p}^o, \forall i \in \mathscr{F}, \tag{2.7d}$$

$$\sum_{k=1}^{F+M} \|\mathbf{v}_k\|^2 \leq P_m, \tag{2.7e}$$

$$\alpha^o W, (1-\alpha^o)W \in \mathbb{Z}^+, \tag{2.7f}$$

$$1 \leq \alpha^o W \leq W-1; \forall o \in \mathscr{A} = \{d,u\}, \tag{2.7g}$$

where the objective function (2.7a) is the sum achievable rate at the small cell on the DL and UL. Constraint (2.7b) means that the WB transmission rate from the MBS to the ith SAP on

the DL should be larger than the WA data transmission rate from the ith SAP to its intended SUE. The same rule applies to the WB and WA rate on the UL. With (2.7a) and (2.7b), the rate at each ith small cell should be regulated by the smaller value between the rate of the WBC and WAC at that small cell. Note that the relationship between the WB and WA rates in (2.7b) is important since the WBC must rely on an appropriate resource allocation solution to attain sufficient rate to support the corresponding WAC. Intuitively, if we consider the maximization of WA small cell sum rate on the UL and DL without (2.7b), the achieved WA rate might exceed the WB rate, which practically violates their relationship. On the other hand, the achievable rates at the MUEs via WAC are regulated by (2.7c) to be larger than some fixed minimum rate requirements $\mathsf{r}^o_{\min}$. Constraints (2.7d)–(2.7e) account for the maximum power at the SAPs, MUEs, SUEs, and MBS. (2.7f) requires that the RB assignment is a positive integer number. In (2.7g), due to constraint (2.7c), there must be at least 1 RB allocated for the MBS-to-MUEs transmissions to ensure the QoS requirement, which subsequently implies that the WBC and WAC of the small cell on the UL and DL must have at least 1 RB. Since $\alpha^o W, (1-\alpha^o) W \geq 1$, we have $\alpha^o W, (1-\alpha^o) W \leq W - 1$. The difference between (2.7) and the conventional MU-MISO problem lies in (2.7b) where the lower bound of WB rate of each SAP is not fixed as in (2.7c). A summary of the optimization variables is presented in Table 2.1.

It is easy to check that (2.7a) is a function of both α and $\mathbf{p}$ which can be decoupled. However, these variables are then coupled in (2.7b), which make it challenging to be decomposed into two independent problems, and thus admits a joint approach to solve. Solving (2.7) is difficult due to the non-concave objective function in (2.7a) and the non-convex constraints (2.7b) and (2.7c). (2.7) is also difficult owing to the integer constraint (2.7f). To solve this problem, we first employ multiple equivalent transformations on (2.7) to turn it into a more tractable form. Then, we apply a Branch-and-Bound algorithm and propose a low complexity iterative algorithm to arrive at the global and local optimal solutions, respectively.

2.3 Global Optimal Solution

In this subsection, we solve (2.7) optimally using Branch-and-Bound (BnB) exhaustive search algorithm. First, we proceed to equivalently transform (2.7) into the following form

$$\max_{\alpha,\mathbf{v},\rho,\mathbf{p},\chi\geq 0} \sum_{o\in\mathscr{A}}\sum_{i=1}^{F}\left(1-\alpha^{\mathscr{A}\setminus o}\right)\chi_i^o \tag{2.8a}$$

$$\text{s.t.} \quad \Gamma_i^o \geq \exp\left(\frac{\left(1-\alpha^{\mathscr{A}\setminus o}\right)\chi_i^o}{\alpha^o}\right)-1, \forall i\in\mathscr{F}, \tag{2.8b}$$

$$\Gamma_i^o \geq \exp\left(\frac{\mathrm{r}_{\min}^o}{\alpha^o}\right)-1, \forall i\in\mathscr{M}, \tag{2.8c}$$

$$\gamma_i^o = e^{\chi_i^o}-1, \forall i\in\mathscr{F}; (2.7\text{d})-(2.7\text{g}), \tag{2.8d}$$

where $\chi=\left[\chi_1^u,\ldots,\chi_F^u,\chi_1^d,\ldots,\chi_F^d\right]^T$ is the set of newly introduced variables and $\chi_i^o\geq 0,\forall i\in\mathscr{F},o\in\mathscr{A}$. The equivalence between (2.8) and (2.7) can be seen by replacing χ_i^o with $\gamma_i^o,\forall i\in\mathscr{F}$ from (2.8d). By observing (2.8), we note two important properties. First, when each of the term χ_i^o or $1-\alpha^o,\forall i,o$ increases within its feasible domain, the objective function of (2.8) achieves a higher value. Second, when all of χ_i^o and $\alpha^o,\forall i,o$ are fixed, (2.8) becomes the feasibility checking problem of finding the solution of $\{\mathbf{v},\rho,\mathbf{p}\}$ that satisfies (2.8b)–(2.8d). Thus, we can employ the concept of monotonic optimization (Tuy *et al.*, 2005) to customize the BnB based algorithm in order to optimally solve (2.8). Let us denote $\Omega=\left[\chi^T,\tilde{\alpha}^T\right]^T$ as the set of variables χ and $1-\alpha^o,\forall o\in\mathscr{A}$. By following the definitions and facts in (Tervo *et al.*, 2015b, Section III-B, pp. 5577), we define $\mathscr{C}=\left\{\Omega\in\mathbb{R}^{F+2}|(2.8\text{b})-(2.8\text{d})\right\}$ as the normal compact set and $\mathscr{D}=\left[\underline{\Omega},\bar{\Omega}\right]$ as the box that contains all the feasible solutions related to Ω in (2.8). Obviously, the lower bound is given by $\underline{\Omega}=[0,\ldots,0,\varepsilon,\varepsilon]^T$, where $\varepsilon=1/W$. To compute the upper bound, we can simply consider that $\chi_i^o\leq\bar{\chi}_i^o=\log\left(1+\frac{\left|h_{ii}^o\right|^2 p_i^o}{\varepsilon W\sigma_0}\right)$ and $1-\alpha^o\leq 1-\varepsilon,\forall i,o$. Thus, the upper bound is given by $\bar{\Omega}=\left[\bar{\chi}_1^d,\ldots,\bar{\chi}_F^u,1-\varepsilon,1-\varepsilon\right]^T$. Given the value of $\Omega\in\mathscr{D}$, the problem of checking whether $\Omega\in\mathscr{C}$ or not becomes the feasibility

checking problem, which can be decoupled into two sub-problems as follow

$$(DL): \text{find} \left\{ \mathbf{v}, \mathbf{p}^d | (2.8\text{b}) - (2.8\text{d}), o = d \right\}$$
$$(UL): \text{find} \left\{ \rho, \mathbf{p}^u | (2.8\text{b}) - (2.8\text{d}), o = u \right\}.$$

Problem (DL) can be easily transformed into SOCP problem as in (Liu *et al.*, 2011b), which can be easily solved optimally. To solve problem (UL), we follow the framework in (Rashid-Farrokhi *et al.*, 1998b) to check the feasibility of problem (UL) with the given Ω. In particular, we can formulate the problem of power minimization as the work in (Rashid-Farrokhi *et al.*, 1998b) and apply the similar algorithm to achieve the optimal solution. Then, we can compare the achieved solution with constraint (2.7d) to check for the feasibility of problem (UL). Based on the above analysis, the reformulated problem (2.8) can be rewritten as $\max \left\{ \sum_{o \in \mathscr{A}} \sum_{i=1}^{F} \left(1 - \alpha^{\mathscr{A} \setminus o}\right) \chi_i^o | \Omega \in \mathscr{C} \subset \mathscr{D} \right\}$. To solve (2.8), we first check whether $\underline{\Omega}$ is feasible or not. If feasible, we run the BnB based algorithm that recursively branches the box $\mathscr{D}$ into smaller boxes, checks the feasibility of each new box, updates the new upper and lower bounds by the Box Reduction—Bound Computation process and disposes the boxes that do not contain the optimal solution. Unlike the case in (Tervo *et al.*, 2015b, Section III-B), our formulated boxes contain the variable $1 - \alpha^o, \forall o \in \mathscr{A}$ which only takes discrete value that satisfies the integer condition (2.7f). Therefore, we apply the modified box branching techniques as follows. Suppose that at the *n*th iteration, one of the boxes associated with the largest upper bound is selected to branch. The box is divided into two smaller boxes by a bisection rule that partitions along the longest edge of the box. Assume that j is the index of the longest edge of the chosen box $\mathscr{D} = \left[\underline{\Omega}, \bar{\Omega}\right]$. If j is the index of a continuous variable, i.e., not $1 - \alpha^o$, the two new boxes are computed as $\mathscr{D}^{(1)} = \left[\underline{\Omega}, \bar{\Omega} - \mathbf{e}_j \left(\bar{\Omega}_j - \underline{\Omega}_j\right)/2\right]$ and $\mathscr{D}^{(2)} = \left[\underline{\Omega} + \mathbf{e}_j \left(\bar{\Omega}_j - \underline{\Omega}_j\right)/2, \bar{\Omega}\right]$, where $\mathbf{e}_j$ represents a $(2F+2) \times 1$ vector that contains 1 at the jth element and 0 everywhere. If j is the index of a discrete variable such as $1 - \alpha^o$, we first compute $\Delta\Omega = \frac{\left\lceil W\left((\bar{\Omega}_j - \underline{\Omega}_j)/2\right)\right\rceil}{W}$. Then, the two new boxes are computed as $\mathscr{D}^{(1)} = \left[\underline{\Omega}, \bar{\Omega} - \mathbf{e}_j \Delta\Omega\right]$ and $\mathscr{D}^{(2)} = \left[\underline{\Omega} + \mathbf{e}_j \Delta\Omega, \bar{\Omega}\right]$. According to (Tuy *et al.*, 2005), the BnB algorithm exhaustively searches for all possible solutions and terminates after many iterations when the difference between the upper and lower

bounds is arbitrary small and the global optimal solution is determined. The details of the BnB algorithm are similar to Algorithm 1 in (Tervo *et al.*, 2015b, Section III-B), where explicit descriptions of the protocol of Box Branching, Box Reduction, Pruning and Bound Computation are presented.

2.4 Low Complexity Iterative Algorithm

The BnB algorithm is basically an exhaustive search method, which requires very high computational cost to yield the global optimal solution. Typically, this solution approach is not practical and only helpful for benchmark evaluation. In this section, we first propose to remove the integer constraints in (2.7f) and to solve the relaxed problem of (2.7). Then, we aim to develop a low complexity efficient algorithm to solve this new problem based on multiple equivalent transformations and convex approximation methods to achieve the solution. The solution obtained from this algorithm is then refined by choosing $\tilde{\alpha}^o$ close to α^o that satisfies $\tilde{\alpha}^o W = \lceil \alpha^o W \rceil$ and $(1-\tilde{\alpha}^o) W = \lfloor (1-\alpha^o) W \rfloor$. Note that $\tilde{\alpha}^o$ is a feasible solution of (2.7) which also satisfies constraint (2.7b) [1].

2.4.1 Equivalent transformations

We first deal with the non-concave objective function of the relaxation of (2.7). By introducing a set of slack variables $q_i^u, q_i^d \geq 0$, where $i \in \mathscr{F}$, we can simply reformulate the problem as

$$\max_{\substack{\alpha,\mathbf{v},\mathbf{q}\geq 0 \\ \rho,\mathbf{p}}} \sum_{i=1}^{F} \left(W q_i^d + W q_i^u \right) \tag{2.9a}$$

$$\text{s.t.} \quad \gamma_i^o \geq e^{q_i^o/(1-\alpha^{\mathscr{A}\setminus o})} - 1, \forall i \in \mathscr{F}, \tag{2.9b}$$

$$(2.7\text{b}) - (2.7\text{e}), (2.7\text{g}),$$

[1] The feasibility of $\tilde{\alpha}^o$ that satisfies (2.7b) after applying the rounding function can be checked by observing two monotonically increasing functions $f_1(x) = x \log\left(1 + \frac{a_1}{b_1+x}\right)$ and $f_2(y) = y \log\left(1 + \frac{a_2}{b_2+y}\right)$. If $\exists (\hat{x}, \hat{y}) : f_1(\hat{x}) \geq f_2(\hat{y})$, then we have $f_1(\lceil \hat{x} \rceil) \geq f_1(\hat{x}) \geq f_2(\hat{y}) \geq f_2(\lfloor \hat{y} \rfloor)$.

where $\mathbf{q} = \left[q_1^u, \ldots, q_F^u, q_1^d, \ldots, q_F^d\right]^T$ is a vector consisting of the newly introduced variables. Indeed, if we multiply the term $\left(1-\alpha^{\mathscr{A}\setminus o}\right)$ on both sides of (2.9b), the right side of (2.9b) is a convex function with respect to variables q_i^o and $\alpha^{\mathscr{A}\setminus o}$. However, the left side of (2.9b) is still a non-concave function. To ease later transformation tractability, we can further decompose this constraint into two new inequalities with the help of another set of new slack variables $\mathbf{t} = \left[t_1^u, \ldots, t_F^u, t_1^d, \ldots, t_F^d\right]^T$, where $t_i^u, t_i^d \geq 0, \forall i \in \mathscr{F}$, as

$$\left(1-\alpha^{\mathscr{A}\setminus o}\right)\gamma_i^o \geq t_i^o, \tag{2.10}$$

$$\left(1-\alpha^{\mathscr{A}\setminus o}\right)\left(e^{q_i^o/(1-\alpha^{\mathscr{A}\setminus o})}-1\right) \leq t_i^o, \tag{2.11}$$

$\forall i \in \mathscr{F}$. Note that the objective function now becomes linear with respect to the variables $\mathbf{q}$ and (2.11) admits the form of exponential cone programming, which is a convex constraint. We observe that the relaxation of (2.7) and (2.9) are equivalent since at optimality, all the constraints in (2.10), (2.11) are active. The proof of equivalence is briefly summarized in Appendix 1.1. Constraint (2.10) can be made more tractable by decomposing it into the following inequalities:

$$\left(1-\alpha^{\mathscr{A}\setminus o}\right)p_i^o \left|h_{ii}\right|^2 \geq t_i^o z_i^o, \tag{2.12a}$$

$$\sum_{j\neq i} p_j^o \left|h_{ij}\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right)W\sigma_0 \leq z_i^o, \tag{2.12b}$$

where $z_i^d \geq 0$ and $z_i^u \geq 0$ for $i \in \mathscr{F}$ are the newly introduced variables, which can be packed into a vector $\mathbf{z} = \left[z_1^d, \ldots, z_F^d, z_1^u, \ldots, z_F^u\right]^T$ for notational convenience. The equivalence between these transformations follow the same line of proof as shown in Appendix 1.2. Next, we turn our attention to another non-convex constraint (2.7b). In particular, we rewrite each inequality in (2.7b) into a set of three inequalities as

$$\alpha^o a_i^o \geq \left(1-\alpha^{\mathscr{A}\setminus o}\right)b_i^o, \tag{2.13a}$$

$$a_i^o \leq \log\left(1+\Gamma_i^o\right), \tag{2.13b}$$

$$b_i^o \geq \log\left(1+\gamma_i^o\right), \tag{2.13c}$$

where $a_i^o \geq 0$, $b_i^o \geq 0$ are the newly introduced variables which are, for brevity, denoted by a set of variables $\mathbf{a} = \left[a_1^d, \ldots, a_F^d, a_1^u, \ldots, a_F^u\right]^T$, $\mathbf{b} = \left[b_1^d, \ldots, b_F^d, b_1^u, \ldots, b_F^u\right]^T$ for later usage. The proof for the equivalence of these transformations in (2.13) is presented in Appendix 1.3.

At this point, we can reformulate (2.9) into a new equivalent form as in (2.14) given at the end of this subsection, where the constraints in (2.14f) and (2.14g) are the results from the equivalent decomposition from (2.13b) and (2.7c) with $o = d$. Note that in (2.14), we have introduced a set of new variables $\nu = [\nu_1, \ldots, \nu_{F+M}]^T$ with additional constraints (2.14f) and (2.14g). The equivalence between these can be proved by following similar steps used in the proof in Appendix 1. To leverage the presentation of all the constraints in (2.13b) and (2.7c) in a unified manner, we introduce an additional constraint (2.14k) to exhibit the similarity between (2.13b) and (2.7c). Similarly, we rewrite (2.13b) and (2.7c) with $o = u$ by (2.14i) and (2.14j), respectively, where $\eta_i, \forall i \in \mathscr{F}$ are the newly introduced variables and $\eta = [\eta_1, \ldots, \eta_F, \eta_{F+1}, \ldots, \eta_{F+M}]^T$. The reason for introducing η will be shown in the next subsection as it is helpful to reveal the non-convex property of constraint (2.14i) so that the convex approximation technique can be appropriately applied.

$$\max_{\substack{\alpha,\mathbf{q}\geq 0\\ \mathbf{t}\geq 0,\mathbf{z}>\mathbf{0}\\ \mathbf{a}\geq\mathbf{0},\mathbf{b}\geq\mathbf{0}\\ \mathbf{v},\overline{\nu},\boldsymbol{\rho},\bar{\mathbf{p}},\eta}} \quad \sum_{i=1}^{F}\left(Wq_i^u + Wq_i^d\right) \tag{2.14a}$$

$$\text{s.t.} \quad \left(1-\alpha^{\mathscr{A}\setminus o}\right) p_i^o \left|h_{ii}\right|^2 \geq t_i^o z_i^o, \forall i \in \mathscr{F}, \tag{2.14b}$$

$$\sum_{j\neq i} p_j^o \left|h_{ij}^o\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right) W\sigma_0 \leq z_i^o, \forall i \in \mathscr{F}, \tag{2.14c}$$

$$\left(1-\alpha^{\mathscr{A}\setminus o}\right)\left(e^{q_i^o/(1-\alpha^{\mathscr{A}\setminus o})} - 1\right) \leq t_i^o, \forall i \in \mathscr{F}, \tag{2.14d}$$

$$\alpha^o a_i^o \geq \left(1-\alpha^{\mathscr{A}\setminus o}\right) b_i^o, \forall i \in \mathscr{F}, \tag{2.14e}$$

$$\left|\mathbf{v}_l^H \mathbf{h}_l\right|^2 / \nu_l \geq e^{a_l^d} - 1, \forall l \in \mathscr{G}, \tag{2.14f}$$

$$\sum_{k\neq i} \left|\mathbf{v}_k^H \mathbf{h}_l\right|^2 + \alpha^o W\sigma_0 \leq \nu_l, \forall l \in \mathscr{G}, \tag{2.14g}$$

$$b_i^o \geq \log\left(1+\gamma_i^o\right), \forall i \in \mathscr{F}, \tag{2.14h}$$

$$\eta_i^2 \mathbf{h}_l^H \Sigma_l^{-1} \mathbf{h}_i \geq e^{a_l^u} - 1, \forall l \in G, \tag{2.14i}$$

$$\rho_l \geq \eta_l^2, \forall l \in \mathscr{G}, \tag{2.14j}$$

$$a_j^o \geq \mathsf{r}_{\min}^o / (W \alpha^o), \forall j \in \mathscr{M}, \tag{2.14k}$$

$$(2.7d), (2.7e), (2.7g), \tag{2.14l}$$

2.4.2 Problem approximations

We note that the transformed optimization problem (2.14) is still non-convex due to the non-convex constraints (2.14b), (2.14e), (2.14f), (2.14h), (2.14i). To deal with these constraints, we present various convex approximation methods to construct an approximated convex problem which can be used to develop a low complexity iterative algorithm. In addition, we rely on the result in (Marks & Wright) to associate the three conditions which can be employed to verify the application of the proposed convex approximation. The proposed convex approximation methods should follow these three conditions so that the developed iterative algorithm based on these approximation will converge to the result with a solution that satisfies the KKT conditions of problem (2.7). First, note that (2.14b), (2.14e) admit the same non-convex constraint form, which can be simply expressed as $xy \geq uv$. It is easy to see that the left and right sides of this inequality are neither convex nor concave functions with respect to all variables. To deal with this obstacle, we first employ the upper bound convex approximation of the right side by the following inequality as $uv \leq \frac{\xi^{(n)}}{2}u^2 + \frac{1}{2\xi^{(n)}}v^2$, for every $\xi^{(n)} > 0$ (Beck *et al.*, 2010). It is straightforward to show that $uv = \frac{\xi^{(n)}}{2}u^2 + \frac{1}{2\xi^{(n)}}v^2$ when $\xi^{(n)} = v/u$. In addition, at this value of $\xi^{(n)}$, their first derivative is also equal, e.g., $\nabla(uv) = \nabla\left(\frac{\xi^{(n)}}{2}u^2 + \frac{1}{2\xi^{(n)}}v^2\right)$. Thus, this convex approximation satisfies the three conditions in (Marks & Wright). By replacing the right side of $xy \geq uv$ with its upper bound convex approximation $\frac{\xi^{(n)}}{2}u^2 + \frac{1}{2\xi^{(n)}}v^2$ and rewriting $xy = \frac{(x+y)^2-(x-y)^2}{4}$, we can write the approximated convex constraint in the SOC form as

$$\frac{x+y}{2} \geq \left\| \sqrt{\frac{\xi^{(n)}}{2}}u, \sqrt{\frac{1}{2\xi^{(n)}}}v, \frac{x-y}{2} \right\|. \tag{2.15}$$

Next, we consider the non-convex constraint (2.14f). Observe that the right side of (2.14f) is convex with respect to a_l^d and the left side is jointly convex with respect to $\mathbf{v}_l$ and ν_l. Thus, we proceed to approximate the functions on the left side of these inequalities by their first-order approximation, denoted as $F\left(\mathbf{h}_l, \mathbf{v}_l, \nu_l, \mathbf{v}_l^{(n)}, \nu_l^{(n)}\right)$ around the point $\mathbf{v}_l^{(n)}, \nu_l^{(n)}$ for $l \in \mathscr{G}$ as follows:

$$F\left(\mathbf{h}_l, \mathbf{v}_l, \nu_l, \mathbf{v}_l^{(n)}, \nu_l^{(n)}\right) = \frac{2\mathscr{R}\left((\mathbf{v}_l^{(n)})^H \mathbf{H}_l \mathbf{v}_l\right)}{\nu_l^{(n)}} - \frac{(\mathbf{v}_l^{(n)})^H \mathbf{H}_l \mathbf{v}_l^{(n)}}{(\nu_l^{(n)})^2} \nu_l, \tag{2.16}$$

where $\mathbf{H}_l = \mathbf{h}_l \mathbf{h}_l^H$. Now, we turn our attention to the non-convex constraints (2.14h). By simple algebraic manipulation, these constraints can be rewritten as

$$b_i^o + \log\left(\sum_{j\neq i}^{F} p_j^o \left|h_{ij}^o\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right) W\sigma_0\right) \geq \log\left(\sum_{j=1}^{F} p_j^o \left|h_{ij}^o\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right) W\sigma_0\right), \tag{2.17}$$

where the left side and the right side of these inequalities are concave functions of the involved variables. By following the same framework in (Kha *et al.*, 2012), we approximate the function on the right side of these inequalities by their first-order approximation, denoted as $G_i^o\left(\mathbf{p}^o, \alpha^{\mathscr{A}\setminus o}, \mathbf{p}^{o(n)}, \alpha^{\mathscr{A}\setminus o(n)}\right)$ around the point $\mathbf{p}^{o(n)}, \alpha^{\mathscr{A}\setminus o(n)}$ as follow

$$G_i^o\left(\mathbf{p}^o, \alpha^{\mathscr{A}\setminus o}, \mathbf{p}^{o(n)}, \alpha^{\mathscr{A}\setminus o(n)}\right) = \log\left(g_i^o\left(\mathbf{p}^{o(n)}, \alpha^{\mathscr{A}\setminus o(n)}\right)\right) + \frac{g_i^d\left(\mathbf{p}^o, \alpha^{\mathscr{A}\setminus o}\right)}{g_i^o\left(\mathbf{p}^{o(n)}, \alpha^{\mathscr{A}\setminus o(n)}\right)}, \tag{2.18}$$

where $g_i^o\left(\mathbf{p}^o, \alpha^{\mathscr{A}\setminus o}\right) = \sum_{j=1}^{F} p_j^o \left|h_{ij}^o\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right) W\sigma_0$. Finally, we focus on the non-convex constraints (2.14i), which is equivalent to $e^{a_i^u} - 1 - \eta_i^2 \mathbf{h}_i^H \Sigma_i^{-1} \mathbf{h}_i \leq 0$. Note that function $h(\mathbf{h}, \eta_i, \rho_{\sim i}) = -\eta_i^2 \mathbf{h}_i^H \Sigma_i^{-1} \mathbf{h}_i$ is a jointly concave function of the corresponding variables (Boyd & Vandenberghe, 2004), where $\rho_{\sim i}$ denote the set of transmit power at the SAPs and MUEs to the MBS on the UL, except for the ith user. Thus, the convex upper bound of the term $h(\mathbf{h}, \eta_i, \rho_{\sim i})$ in (2.14i) can be found by its first-order approximation around the point $\eta_i^{(n)}, \rho_{\sim i}^{(n)}$ as follow

$$h(\mathbf{h}, \eta_i, \rho_{\sim i}) \leq H\left(\mathbf{h}, \eta_i, \rho_{\sim i}, \eta_i^{(n)}, \rho_{\sim i}^{(n)}\right) \tag{2.19}$$

$$= h\left(\eta_i^{(n)}, \rho_{\sim i}^{(n)}\right) - 2\eta_i^{(n)} \mathbf{h}_i^H \left(\Sigma_i^{(n)}\right)^{-1} \mathbf{h}_i \left(\eta_i - \eta_i^{(n)}\right)$$
$$+ \mathrm{Tr}\left[\left(\eta_i^{(n)}\right)^2 \left(\Sigma_i^{(n)}\right)^{-1} \mathbf{h}_i \mathbf{h}_i^H \left(\Sigma_i^{(n)}\right)^{-1} \left(\Sigma_i - \Sigma_i^{(n)}\right)\right] \tag{2.20}$$

It can be easily verified that the approximation employed in (2.16), (2.18) and (2.20) also satisfy the three conditions shown in (Marks & Wright) by following the same steps to verify the convex approximation on $xy \geq uv$. By applying the approximations (2.15), (2.16), (2.18) and (2.20) into their corresponding non-convex constraints in (2.14), the problem in (2.7) can be solved by iteratively solving the following approximated convex problem, which is formulated at the $n+1$ index as

$$\max_{\substack{\alpha,\mathbf{q}\geq 0 \\ \mathbf{t}\geq 0,\mathbf{z}>\mathbf{0} \\ \mathbf{a}\geq\mathbf{0},\mathbf{b}\geq\mathbf{0} \\ \mathbf{v},\nu\rho,\mathbf{p},\eta}} \sum_{i=1}^{F} \left(W q_i^u + W q_i^d\right) \tag{2.21a}$$

$$\text{s.t.} \quad \left\| \sqrt{\frac{\xi_i^{o(n)}}{2}} \frac{t_i^o}{|h_{ii}|}, \sqrt{\frac{1}{2\xi_i^{o(n)}}} \frac{z_i^o}{|h_{ii}|}, \frac{1-\alpha^{\mathscr{A}\setminus o} - p_i^o}{2} \right\| \leq \frac{\left(1-\alpha^{\mathscr{A}\setminus o}\right) + p_i^o}{2}, \forall i \in \mathscr{F}, \tag{2.21b}$$

$$\left\| \sqrt{\frac{\psi_i^{o(n)}}{2}} \left(1-\alpha^{\mathscr{A}\setminus o}\right), \sqrt{\frac{1}{2\psi_i^{o(n)}}} b_i^o, \frac{\alpha^o - a_i^o}{2} \right\| \leq \frac{\alpha^o + a_i^o}{2}, \forall i \in \mathscr{F}, \tag{2.21c}$$

$$F\left(\mathbf{h}_l, \mathbf{v}_l, \nu_l, \mathbf{v}_l^{(n)}, \nu_l^{(n)}\right) \geq e^{a_l^d} - 1, \forall l \in \mathscr{G}, \tag{2.21d}$$

$$b_i^o + \log\left(\sum_{j\neq i}^{F} p_j^o \left|h_{ij}^o\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right) W\sigma_0\right) \geq$$
$$G_i^o\left(\mathbf{p}^{\mathbf{o}}, 1-\alpha^{\mathscr{A}\setminus o}, \mathbf{p}^{o(n)}, 1-\alpha^{\mathscr{A}\setminus o(n)}\right), \forall i \in \mathscr{F}, \tag{2.21e}$$

$$H\left(\mathbf{h}, \eta_l, \rho_{\sim l}, \eta_l^{(n)}, \rho_{\sim l}^{(n)}\right) \leq 1 - e^{a_l^u}, \forall l \in \mathscr{G}, \tag{2.21f}$$

$$(2.14c), (2.14d), (2.14g), (2.14j), (2.14k), (2.7d), (2.7e), (2.7g). \tag{2.21g}$$

Note that $\alpha^{d(n)}, \alpha^{u(n)}, \xi_i^{o(n)}, \psi_i^{o(n)}, p_i^{o(n)}$ for $i \in \mathscr{F}, o \in \mathscr{A}$ and $\eta_l^{(n)}, \rho_l^{(n)}, \mathbf{v}_l^{(n)}, \nu_l^{(n)}$ for $l \in \mathscr{G}$ are not the variables of the optimization problem but the parameters that are iteratively updated by the optimum solution after each iteration. We denote $\xi^{(n)} = [\xi_1^{d(n)}, \ldots, \xi_F^{d(n)}, \xi_1^{u(n)}, \ldots, \xi_F^{u(n)}]^T$, $\psi^{(n)} = [\psi_1^{d(n)}, \ldots, \psi_F^{d(n)}, \psi_1^{u(n)}, \ldots, \psi_F^{u(n)}]^T$, $\eta^{(n)} = [\eta_1^{(n)}, \ldots, \eta_F^{(n)}]^T$ and $\nu^{(n)} = [\nu_1^{(n)}, \ldots, \nu_{F+M}^{(n)}]^T$

for the sake of brevity. Thus, the pseudo code presents the JBRAO algorithm to solve the optimization problem is summarized in the Algorithm 1. We rely on the proof in (Venkatraman *et al.*, 2016a, Appendix A) to provide the proof that Algorithm 1 converges at the KKT solution of (2.14). The detailed proof is omitted here.

Algorithm 1: Iterative JBRAO Algorithm

Initialize starting points of $\alpha^{(n)}, \xi^{(n)}, \psi^{(n)}, \eta^{(n)}, \rho^{(n)}, \mathbf{p}^{(n)}, \mathbf{v}^{(n)}, \nu^{(n)}$;
Set $n := 0$;
repeat
  Solve the convex problem in (2.21) to achieve the optimal solution
  $\alpha^\star, \mathbf{q}^\star, \mathbf{t}^\star, \mathbf{z}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \mathbf{v}^\star, \nu^\star, \rho^\star, \mathbf{p}^\star, \eta^\star$;
  Set $n := n+1$;
  Update $\alpha^{(n)} = \alpha^\star, \mathbf{p}^{(n)} = \mathbf{p}^\star, \eta^{(n)} = \eta^\star, \rho^{(n)} = \rho^\star, \mathbf{v}^{(n)} = \mathbf{v}^\star, \nu^{(n)} = \nu^\star, \xi_i^{o(n)} = z_i^{o\star}/t_i^{o\star}$,
  $\psi_i^{o(n)} = b_i^{o\star}/(1-\alpha^{\mathscr{A}\setminus o\star})$ for $i \in \mathscr{F}, o \in \mathscr{A}$;
until Convergence

2.4.3 Imperfect channel state information (CSI)

In practice, due to time-varying channel conditions and user mobility, CSI knowledge is imperfect. This subsequently leads to a few difficulties in system design considering imperfect CSI. At first sight, it is easy to see that we cannot in a straightforward manner apply the MMSE receive beamforming vector with only the knowledge of the estimated CSI to achieve a closed-form formula of the rate as in (2.5). More importantly, the solution of the optimization problem in (2.7) with the knowledge of estimated CSI might not lie in the feasible domain of the problem with perfect CSI. This is because this "imperfect" solution may probably lead to the violation of constraints (2.7b) and (2.7c). Motivated by these observations, we consider the robust design of resource allocation in the case of errorneous CSI with the worst-case design as in (Iserte *et al.*, 2006; Jeong *et al.*, 2011). In particular, the estimated channel values are known with the estimation errors lying in some known size bounded sets. For example, the estimated channel model can be given as in (Iserte *et al.*, 2006): $\mathbf{h}_i = \hat{\mathbf{h}}_i + \varepsilon_{\mathbf{h}}, \forall i \in \mathscr{G}$ and

$h^o_{ij} = \hat{h}^o_{ij} + \varepsilon_{h^o}, \forall o \in \mathscr{A}, \forall i, j \in \mathscr{F}$ where $\hat{\mathbf{h}}_i, \hat{h}^o_{ij}$ are the estimated CSI and $\|\varepsilon_{\mathbf{h}}\|^2 \leq \delta_{\mathbf{h}}$ and $|\varepsilon_{h^o}|^2 \leq \delta_{h^o}$ represent the estimation error which lies in a bounded set defined by $\delta_{\mathbf{h}}$ and δ_{h^o}, respectively. The robust optimization problem can be formulated as:

$$\max_{\substack{\alpha,\mathbf{v},\\ \rho,\mathbf{p}}} \max_{|\varepsilon_{h^o}|^2 \leq \delta_{h^o}} \sum_{i=1}^{F} \left(\mathsf{r}^d_i + \mathsf{r}^u_i \right) \tag{2.22a}$$

$$\text{s.t.} \min_{\|\varepsilon_{\mathbf{h}}\|^2 \leq \delta_{\mathbf{h}}} \mathsf{R}^o_i \geq \max_{|\varepsilon_{h^o}|^2 \leq \delta_{h^o}} \mathsf{r}^o_i, \forall i \in \mathscr{F}, \tag{2.22b}$$

$$\min_{\|\varepsilon_{\mathbf{h}}\|^2 \leq \delta_{\mathbf{h}}} \mathsf{R}^o_j \geq \mathsf{r}^o_{\min}, \forall j \in \mathscr{M}, \tag{2.22c}$$

$$(2.7\text{d}) - (2.7\text{g}), \tag{2.22d}$$

where in this problem, we only have the knowledge of estimated CSI, e.g., $\hat{\mathbf{h}}_i$ and $\hat{h}^o_{ij}$. For the sake of simplicity and tractability, we assume the MRC receive beamforming, $\mathbf{w}_i = \frac{\hat{\mathbf{h}}_i}{\|\hat{\mathbf{h}}_i\|}$ at the MBS to maintain a tractable formula of UL rate. By following the same analysis in (Jeong *et al.*, 2011), we can write each term of (2.22b) and (2.22c) as in (2.23)–(2.25).

$$\min_{\|\varepsilon_{\mathbf{h}}\|^2 \leq \delta_{\mathbf{h}}} \mathsf{R}^d_i = \alpha^d W \times \log \left(1 + \frac{\left| \hat{\mathbf{h}}_i^H \mathbf{v}_i \right|^2 - \delta_{\mathbf{h}} \|\mathbf{v}_i\|_2^2}{\sum_{j \neq i} \left(\left| \hat{\mathbf{h}}_i^H \mathbf{v}_j \right|^2 + \delta_{\mathbf{h}} \left\| \mathbf{v}_j \right\|_2^2 \right) + \alpha^d W \sigma_0} \right) \tag{2.23}$$

$$\min_{\|\varepsilon_{\mathbf{h}}\|^2 \leq \delta_{\mathbf{h}}} \mathsf{R}^u_i = \alpha^u W \log \left(1 + \frac{\rho_i \left(\left\| \hat{\mathbf{h}}_i \right\|^2 - \delta_{\mathbf{h}} \right)}{\sum_{j \neq i} \rho_j \left(\frac{\left| \hat{\mathbf{h}}_i^H \hat{\mathbf{h}}_j \right|^2}{\left\| \hat{\mathbf{h}}_i \right\|^2} + \delta_{\mathbf{h}} \right) + \alpha^u W \sigma_0} \right) \tag{2.24}$$

$$\max_{|\varepsilon_{h^o}|^2 \leq \delta_{h^o}} \mathsf{r}^o_i = \left(1 - \alpha^{\mathscr{A} \setminus o} \right) W \log \left(1 + \frac{\left(\left| \hat{h}^o_{ii} \right|^2 + \delta_{h^o} \right) p^o_i}{\sum_{j \neq i} \left(\left| \hat{h}^o_{ij} \right|^2 - \delta_{h^o} \right) p^o_j + \left(1 - \alpha^{\mathscr{A} \setminus o} \right) W \sigma_0} \right) \tag{2.25}$$

Towards this end, we follow the same steps of relaxation, transformation and approximation in Section 2.4 to arrive at the approximated convex problem, in which a similar iterative algorithm as Algorithm 1 can be applied.

Table 2.1 Summary of optimization variables

Notation	Used for problem	Description
$\alpha^d(\alpha^u)$	(2.8), (2.21)	Spectrum splitting factor in the DL (and UL) time slot
$\mathbf{a}^d(\mathbf{a}^u)$	(2.21)	Soft SE level on the DL (UL) from the MBS
$\mathbf{b}^d(\mathbf{b}^u)$	(2.21)	Soft SE level on the DL (UL) at the small cell via WAC
κ	(2.28)	Auxiliary variables for SOCP approximation
η	(2.21)	Auxiliary power variables to assist the extraction of (2.14i)
ρ	(2.8), (2.21)	UL transmit power from SAP and MUEs to MBS
$\mathbf{p}^d(\mathbf{p}^u)$	(2.8), (2.21)	DL (UL) transmit power from SAPs(SUEs) to SUEs(SAPs)
$\mathbf{q}^d(\mathbf{q}^u)$	(2.21)	Soft rate level at the small cells on the DL (UL) via WAC
$\mathbf{t}^d(\mathbf{t}^u)$	(2.21)	Auxiliary variables to assist the extraction of (2.9b)
ν	(2.21)	Soft interference level on the DL from MBS to SAPs or MUEs
$\mathbf{v}$	(2.8), (2.21)	DL beamforming vector from MBS to SAPs and MUEs
χ	(2.8)	Soft SE level at the small cells on the DL (UL) via WAC
$\mathbf{z}^d(\mathbf{z}^u)$	(2.21)	Soft interference level at the small cells on the DL (UL) via WAC

2.5 SOCP approximation of general exponential constraint

The problem formulation in (2.21) is categorized as a general nonlinear convex program (GNCP), which in general can be solved efficiently within a polynomial time. When a GNCP contains exponential cone constraints, its convexity can be recognized through the callback evaluation process in MATLAB and thus can be solved by a convex solver such as MATLAB's FMINCON. However, this solver often requires high computational time compared to other standard convex programs such as the SOCP. Therefore, this motivates us to approximate (2.21) into a SOCP problem within a desired controlled accuracy since all constraints in (2.21) satisfy the SOCP condition, except for (2.21d), (2.21e), (2.21f), (2.14d). By rewriting (2.21e) as

$$\sum_{j\neq i}^{F} p_j^o \left|h_{ji}\right|^2 + \left(1-\alpha^{\mathscr{A}\setminus o}\right) W\sigma_0 \geq e^{c_i^o}, \tag{2.26}$$

$$G_i^o\left(\mathbf{p}^{\mathbf{o}}, 1-\alpha^{\mathscr{A}\setminus o}, \mathbf{p}^{o(n)}, 1-\alpha^{\mathscr{A}\setminus o(n)}\right) - b_i^o \leq c_i^o, \tag{2.27}$$

we can observe that (2.21d), (2.21f), and (2.26) take the convex exponential cone constraint form of $y \geq e^x$. Moreover, we also remark that (2.14d) contains a more general form of exponential cone, which is $y \geq te^{x/t}$, where if $t = 1$, we obtain the special form $y \geq e^x$. Towards this end, we extend the result in (Ben-Tal & Nemirovski, 2001a) to provide a replacement for the

general exponential cone constraint by a set of newly introduced constraint, which is given by the following system of inequalities as

$$\begin{aligned}
\kappa_{m+4} &\leq y, \\
\left\| \left[2t + x/2^{m-1} \;\; t - \kappa_1 \right] \right\|_2 &\leq t + \kappa_1, \\
\left\| \left[5t/3 + x/2^{m} \;\; t - \kappa_2 \right] \right\|_2 &\leq t + \kappa_2, \\
\left\| \left[2\kappa_1 \;\; t - \kappa_3 \right] \right\|_2 &\leq t + \kappa_3, \\
19/72t + \kappa_2 + 1/24\kappa_3 &\leq \kappa_4, \\
\left\| \left[2\kappa_{i-1} \;\; t - \kappa_i \right] \right\|_2 &\leq t + \kappa_i, i = 5, 6, \ldots, m+3, \\
\left\| \left[2\kappa_{m+3} \;\; t - \kappa_{m+4} \right] \right\|_2 &\leq t + \kappa_{m+4},
\end{aligned} \tag{2.28}$$

where $\kappa_i, i = 1, \ldots, m+4$ are the newly introduced variables and m is the parameter that controls the accuracy of the approximation. The proof for the derivation of (2.28) is given in Appendix 2. Interestingly, with $t = 1$, we arrive at the special case of $y \geq e^x$ as in (Tervo *et al.*, 2015b). Here, for a given m, the system of constraints in (2.28) can approximate $y \geq te^{x/t}$ to a desired accuracy ε defined by m over a given interval $u = x/t \in [0, \bar{u}]$. From (Ben-Tal & Nemirovski, 2001a), when $x/t \in [0, \bar{u}]$ and $y \geq te^{x/t}$, there exists $\{y, x, t, \kappa_1, \ldots, \kappa_{m+4}\}$ that satisfies (2.28). In addition, if $x/t \in [0, \bar{u}]$ and (x, t, y) can be extended by some $\kappa_i, i = 1, \ldots, m+4$ to a solution of (2.28), then $te^{x/t} - \varepsilon \leq y \leq te^{x/t} + \varepsilon$.

Complexity analysis: We discuss the complexity of each proposed algorithm to solve (2.7) in this section. For the BnB based exhaustive search algorithm, since the number of boxes branched by the algorithm is unbounded, the complexity of this algorithm cannot be determined. On the other hand, it is well-known that both the GNCP and SOCP can solve a convex problem within a polynomial time. Thus, we provide the worst-case per-iteration complexity analysis of Algorithm 1 where the exponential cone constraint is applied to the approximation (2.28), which is recognized as SOCP, by following the result in (Lobo *et al.*, 1998). Since the structure of this problem is complicated, we assume that in the worst case, all the constraints have the same size of the most complicated SOC constraint in the problem, where the size

of a SOC constraint is the number of elements in the norm. With this assumption, there are $(N+3)(F+M)+12F+2+4F(m+4)$ variables in problem (2.21) that applies SOCP approximation (2.28). In addition, there are $10F+1$ SOC constraints in (2.21) of size $F+1$ and $4F(m+4)$SOC constraints in (2.28) of size 3. Thus, by omitting the small order, the complexity is given by $\mathcal{O}\left(N^2\left(F^2+M^2\right)F^2+F^4+F^3(m+4)^3\right)$.

2.6 Numerical Results

In this section, the network performance with the proposed algorithms is evaluated. We assume time-invariant and flat Rayleigh fading channels and the path-loss component is calculated as $\left(d_{ij}/d_0\right)^{-3.8}$, where d_{ij} is the distance between the ith transmitter and the jth receiver, and $d_0 = 10$ m is the reference distance. Note that all the numerical results in this section are obtained on the average of 100 channel realizations. In most results, we apply the scenario depicted in Fig. 2.3, where we assume a circular coverage of the macrocell with radius $20d_0$. The MBS is positioned at the center and $F = 4$ small cells are uniformly placed within the considered coverage. We assume that each small cell has a circular coverage of radius $2d_0$ with its SAP at the center and a SUE at the circumference of the each small cell coverage. In addition, we assume $M = 2$ MUEs randomly scattered across the macrocell coverage. Unless being mentioned elsewhere, we choose this scenario as the standard simulation mode, where the number of transmit antenna as $N = 4$ and the maximum transmit power as $P_m = 45$ dBm. The maximum transmit power at the SAPs and users are set as $\bar{\rho} = \bar{p}^o = 35$ dBm. To protect the QoS at each MUE, we choose the minimum rate requirement $\mathsf{r}^o_{\min} = 10^5$ nats/s. The noise power is $W\sigma_0 = -120$ dB and the bandwidth is $W = 10$ MHz.

In Fig. 2.4, we compare the convergence performance of different algorithms applied to solve (2.7) with $F = 2$ small cells, where its objective function is divided by W, which can be defined as sum spectral efficiency (SSE) (nats/s/Hz). Moreover, we also show the performance of the algorithm in (Wang *et al.*, 2016) (denoted by Algorithm in [20]) to solve our problem (2.7). Note that these algorithms are applied to each set of channel realizations to obtain the corresponding objective function value. The evolution of these objective values with respect to

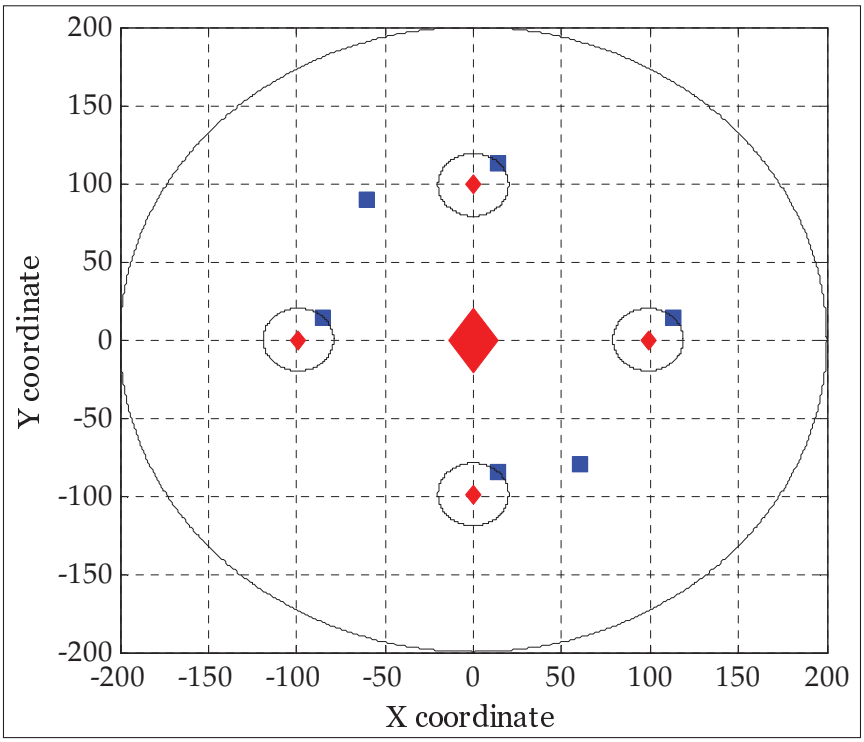

Figure 2.3 Spatial simulation setting

the iteration number are then summed and divided by the number of set of channel realizations to obtain Fig. 2.4. It is worth to mention that in (Wang *et al.*, 2016), this algorithm proposes to use fixed transmit power on the WAC. Based on that, it iteratively searches for the optimal spectrum splitting factor by a bi-section method, then fixes this factor value and solves a convex problem based on the given splitting factor. In our problem, since we use two factors of spectrum splitting, the attempt to search for these optimal values should follow the Box Branching protocol as proposed in BnB Algorithm in Section 2.3. As shown in the figure, the BnB Algorithm recursively updates its upper and lower bounds after several iterations and terminates when their gap is arbitrary small. On the other hand, our proposed algorithm requires much less iterations to converge to its stationary point. Moreover, by averaging out the performance over several channel realizations, our proposed algorithm is shown to converge to the same stationary point despite different initial points. Compared to the algorithm in (Wang *et al.*, 2016) after the same iteration number, the solution achieved from our proposed algorithm is shown to be much closer to the global optimal solution achieved by the BnB Algorithm. This is because

in (Wang *et al.*, 2016), the usage of transmit power on the WAC is not considered for a joint design and not optimally utilized.

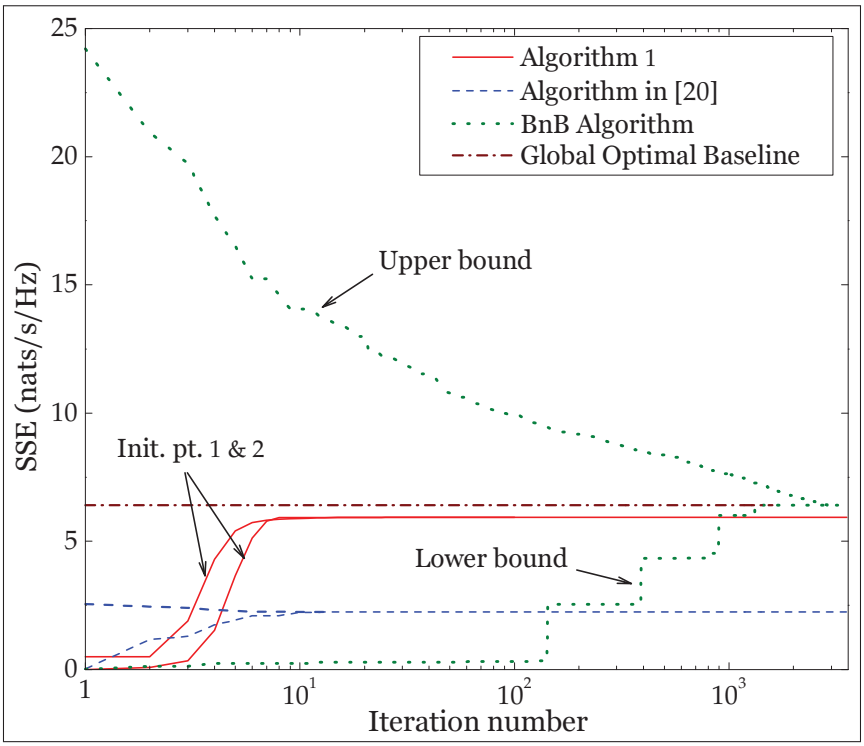

Figure 2.4 SSE convergence.

In Fig. 2.5 and Table 2.2, we compare the average running time and the value of the ratio objective function in (2.21) over W from different solvers over different channel realizations. For the GNCP problem in (2.21) with general exponential cone constraints, we use a general convex solver FMINCON to iteratively obtain the optimum solutions at each iteration of Algorithm 1. To implement the SOC approximations in (2.28) and show the improvement of running time, we use different convex solvers, namely SDPT3, SEDUMI and ECOS. Note that SDPT3 and SEDUMI (Sturm, 1999) are provided by YALMIP (Lofberg, 2004), where SEDUMI is well-known to solve SOCP faster and more efficient than SDPT3. Besides, we attempt to run a commercial SOCP solver, namely ECOS, to solve the approximated SOCP problem. As we can see from the figure, when the problem size grows, all the solvers require more time to achieve the stationary point solution. FMINCON requires a huge amount of time

to finally solve, whereas SEDUMI only requires approximately 3–100 times less running time to achieve the same results. This performance is further improved by the usage of ECOS, which again proves the power of applying SOCP approximation in solving GNCP. In Table 2.2, we compare the value of objective function associated with their average running time obtained at each solvers. These values are achieved at the termination of the algorithm when the difference between two iterations is around 10^{-4}.

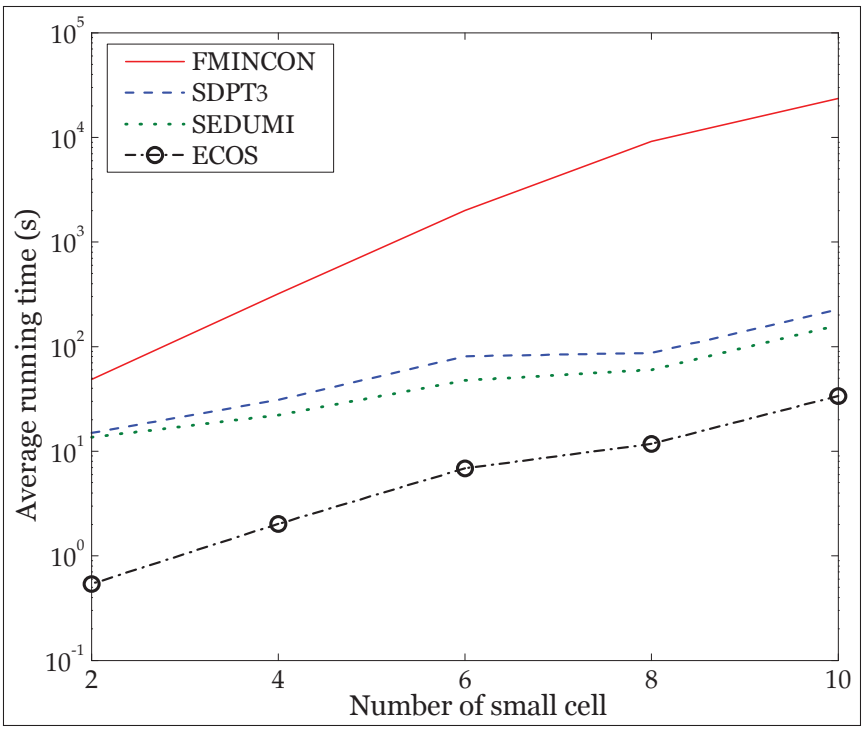

Figure 2.5 Running time of different solvers.

Table 2.2 Comparison of the objective function and running time at different solvers

Solver/F		2	4	6	8	10
FMINCON	Runn. Time	48.7153	319.9065	2×10^3	9.20×10^3	2.36×10^4
	Obj. func.	5.93255	10.33211	11.1548	11.1541	11.1545
SDPT3	Runn. Time	14.9915	30.9467	80.7447	87.4012	226.3562
	Obj. func.	5.93265	10.332	11.1551	11.1554	11.1551
SEDUMI	Runn. Time	13.6397	22.1418	47.6735	60.1245	160.3279
	Obj. func.	5.93265	10.3319	11.15505	11.1553	11.1554
ECOS	Runn. Time	0.5385	2.0193	6.8717	11.7978	33.6731
	Obj. func.	5.93265	10.33206	11.1553	11.1552	11.1552

In Fig. 2.6, we show the SSE on both UL and DL in the imperfect CSI case versus $\bar{p}_f^d$, with different values of error bound $\delta_{\mathbf{h}} = \delta_{h^o} = \delta = 0, 0.1, 0.3$ and $\mathsf{r}_{\min}^0 = 10^5$, 10^6 nats/s. Note that for simplicity, MMSE receive beamforming vector is applied at the MBS on the UL only in the perfect CSI case ($\delta = 0$). From the figure, we observe that MMSE provides better performance than the MRC. On the other hand, all the SSE regarding MMSE and MRC cases increase with $\bar{p}_f^d$ and finally saturates at high value of $\bar{p}_f^d$, e.g., at $\bar{p}_f^d = 50$ dBm. This is obvious because when SAP has higher individual power budget, they will choose to transmit at higher power to achieve more SSE. However, at the regime of high $\bar{p}_f^d$, due to the interference from other SAPs, one tends not to utilize its maximum power budget to attain the optimal sum rate of the network, which in turn leads to the saturation of the SSE. In addition, the SSE in the imperfect CSI is degraded at lower performance compared to the perfect CSI. This is because the optimization problem in (2.22) is solved with the knowledge of the estimated CSI and the obtained solution is deteriorated with this imperfect information. Further, the achieved SSE reduces when δ increases. Finally, when $\mathsf{r}_{\min}^0$ increases, SSE decreases accordingly. This can be explained as when the MUEs have larger minimum rate requirement, more spectrum should be drawn towards the macrocell access transmission to guarantee the constraints. Thus, the radio resource saved for the small cell access link becomes less, which results in lower small cell SSE.

In Fig. 2.7a, we compare the SSE on the UL and DL between four schemes: (i) Proposed scheme where the objective function of (2.7) contains two terms of UL and DL; (ii) UL scheme where (2.7a) only has one term of UL; (iii) DL scheme where (2.7a) has one term of DL and (iv) EQ scheme where (2.7) has one additional constraint $\alpha^d = \alpha^u$. Note that the EQ scheme is similar to the case where we only have one level of spectrum partitioning. To prevent the rate at each small cell to go to zero, we add a minimum rate constraint on each small cell on the UL and DL, respectively. In particular, in each optimization problem for each scheme, we add one more linear constraint $q_i^o \geq \bar{q}^o, \forall i, o$. In the simulation, we choose $\bar{q}^o = 0.1$ nats/s/Hz. Alternatively, the solutions of these schemes are obtained by applying Algorithm 1 to the corresponding problems. We compare the performance of four schemes by evaluating the

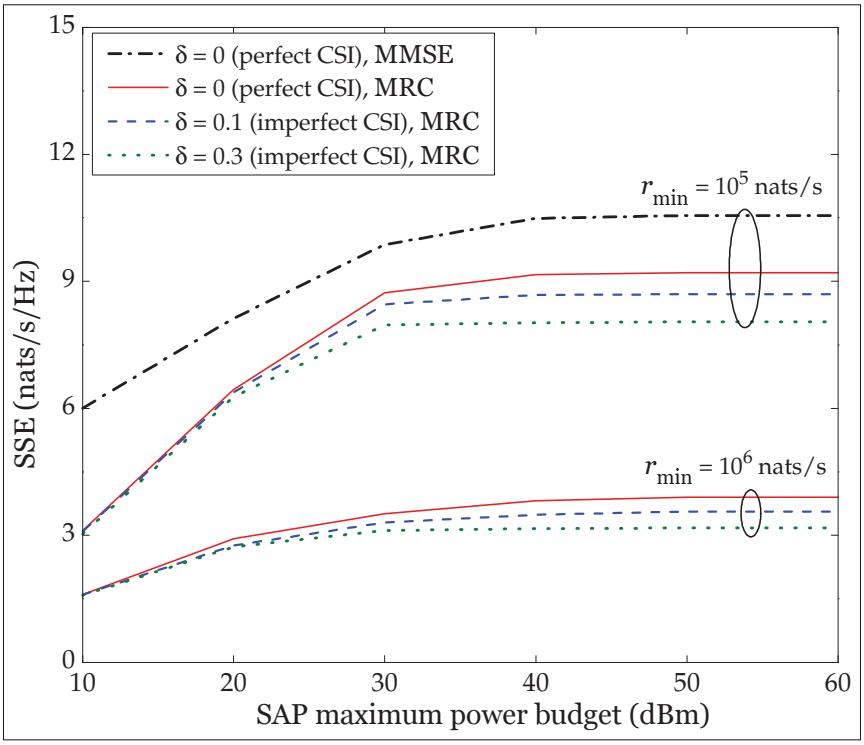

Figure 2.6 SSE of the network in the case of imperfect CSI.

SSE with the achieved solutions. From the figure, we note that the proposed scheme always achieves higher SSE compared to UL and DL schemes. The reason for this is that without considering the resource allocation on both the UL and DL, the allocator only contributes as much resource either to the UL or DL transmission as possible while maintaining enough quality for the other transmission side. Note that when P_m increases, the SSE of DL scheme increases while the SSE of UL scheme remains constant. Since P_m represents the power resource for the DL backhaul transmissions, thus, when it increases, only the DL communications benefit. In addition, we can also observe that the performance of the proposed scheme is always better than the EQ scheme. These results show the effectiveness of our proposed design compared to the traditional designs.

In order to get further insight, in Fig. 2.7b, we compare the SSE of the four schemes by varying $\bar{\rho}_f$. The SSE of the proposed scheme always outperforms the others. We also note that when $\bar{\rho}_f$ increases, the SSE of the UL scheme increases while the SSE of the DL scheme remains

constant. This can be explained because $\bar{\rho}_f$ represents the power resource for the UL backhaul transmissions, thus, when it increases, only the UL communications benefit.

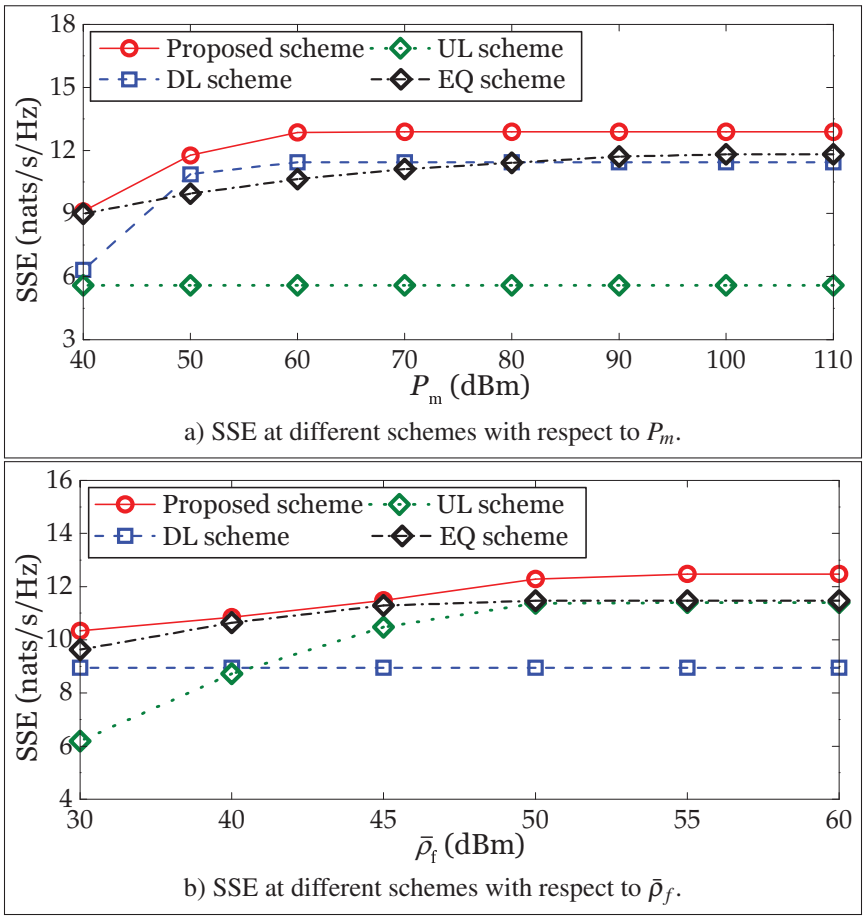

Figure 2.7 SSE of the network with respect to maximum power budget at the MBS and SAP.

Fig. 2.8a shows the level of bandwidth partitioning dedicated for the DL access transmission from the SAPs to SUEs, namely $1-\alpha^u$ when P_m increases. We observe that when P_m increases, the percentage of bandwidth used for the DL access transmission increases and finally saturates when P_m is sufficiently large. This is due to the fact that when more power budget is available for the WBC on the DL, it creates more resource to increase the wireless backhaul rate on the DL and allows the DL access transmissions to use more bandwidth to increase their achievable rates, thus increase $1-\alpha^u$. Compared to the usage of one level bandwidth partitioning in EQ

scheme, the proposed scheme exploits bandwidth partitioning on the UL and DL transmission differently and more efficiently to achieve a better SSE than the EQ scheme as in Fig. 2.7a.

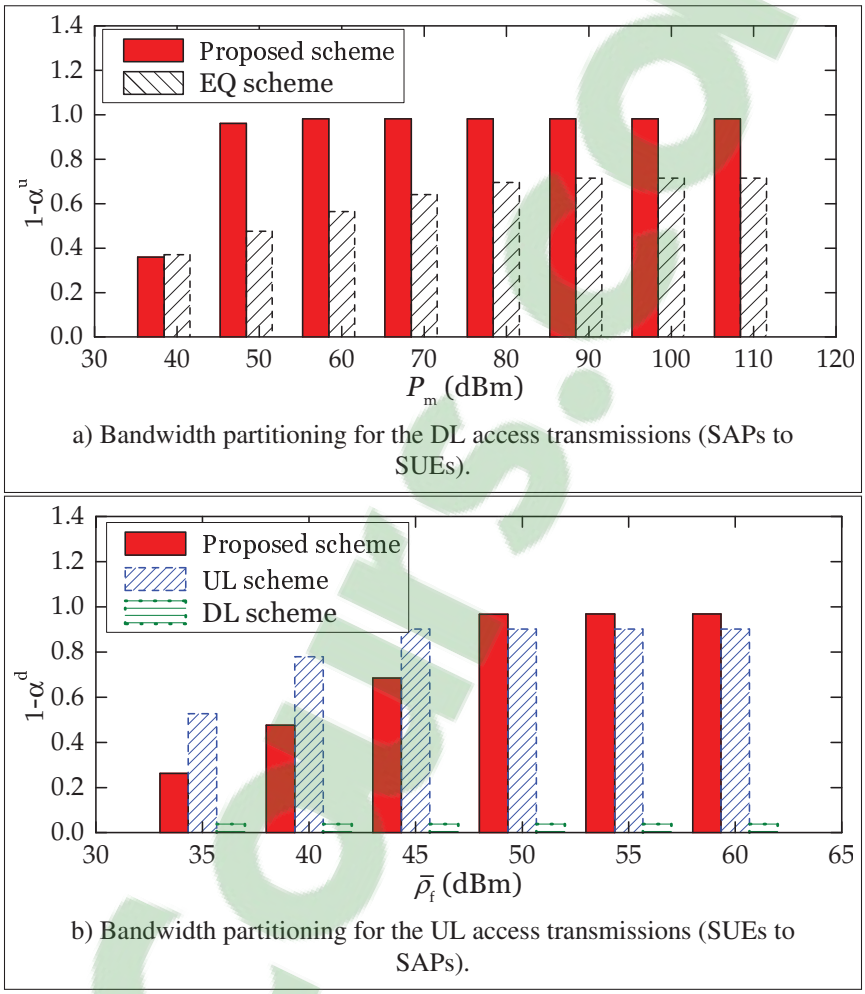

Figure 2.8 Bandwidth partitioning with respect to maximum power budget at the MBS and SAP.

In Fig. 2.8b, we show the level of bandwidth partitioning dedicated to the UL access transmission from the SUEs to the SAPs, namely $1-\alpha^d$, versus the maximum power budget at each SAP for the WBC $\bar{\rho}_f$. We can easily note that when $\bar{\rho}_f$ increases, the term $1-\alpha^d$ for the proposed scheme and UL scheme increases, while it remains constant for the DL scheme. This can be explained similarly to Fig. 2.7b since $\bar{\rho}_f$ only takes effect on the problem of maximizing the UL transmissions. When $\bar{\rho}_f$ increases, more power budget is available to support the WBC

on the UL. This in turn creates more resource for the UL, thus enforces the UL access transmissions to use more bandwidth to increase their achievable rates. This results in the increment of the term $1-\alpha^d$ in the proposed and UL scheme.

In Fig. 2.9a, we observe that when $\mathsf{r}^o_{\min}$ increases, all the SSE from four schemes decrease. This can be explained using similar analogy as in Fig. 2.6 when the MUEs require higher minimum rate for the QoS, more spectrum should be drawn towards the macrocell access transmission to guarantee the constraints, which subsequently leads to lower SSE achieved at the small cells. In addition, the proposed scheme always outperforms the others, which again validates the benefit of our design. Finally, we show the average achieved SSE by applying the proposed algorithm with respect to MBS maximum budget power for different numbers of small cell at $\mathsf{r}^o_{\min} = 10^3$ nats/s. From the figure, we can see that when P_m increases, the achieved SSE increases at different numbers of small cell. This can be explained similarly to Fig. 2.7a. Moreover, when the number of small cell increases, we obtain higher SSE. However, this increment of SSE is low at large number of small cell, e.g., $F = 50$. This is due to the effect of severe interference when there are more small cells deployed into the WB HCNs, which limits the achieved SSE.

2.7 Concluding Remarks

This paper studied the joint design of transmit beamforming, power allocation and bandwidth partitioning for the maximization of sum rate of small cells on both UL and DL transmissions in two-tier HCNs using WBC. By employing the RTDD system, we propose a strategy to partition the bandwidth in two consecutive time slots by two different levels and we formulate an optimization problem based on the proposed scheme. To solve the non-convex problem, we leverage the SPCA method to iteratively find the stationary point of the approximated problem. Moreover, we approximate the problem into SOCP and propose a fast converging algorithm to achieve the solution. To study the performance loss in the imperfect CSI case, we formulate a robust optimization problem and apply the proposed algorithm to solve it efficiently. Numerical results are conducted to demonstrate the network performance gain achieved by deploying the proposed algorithms.

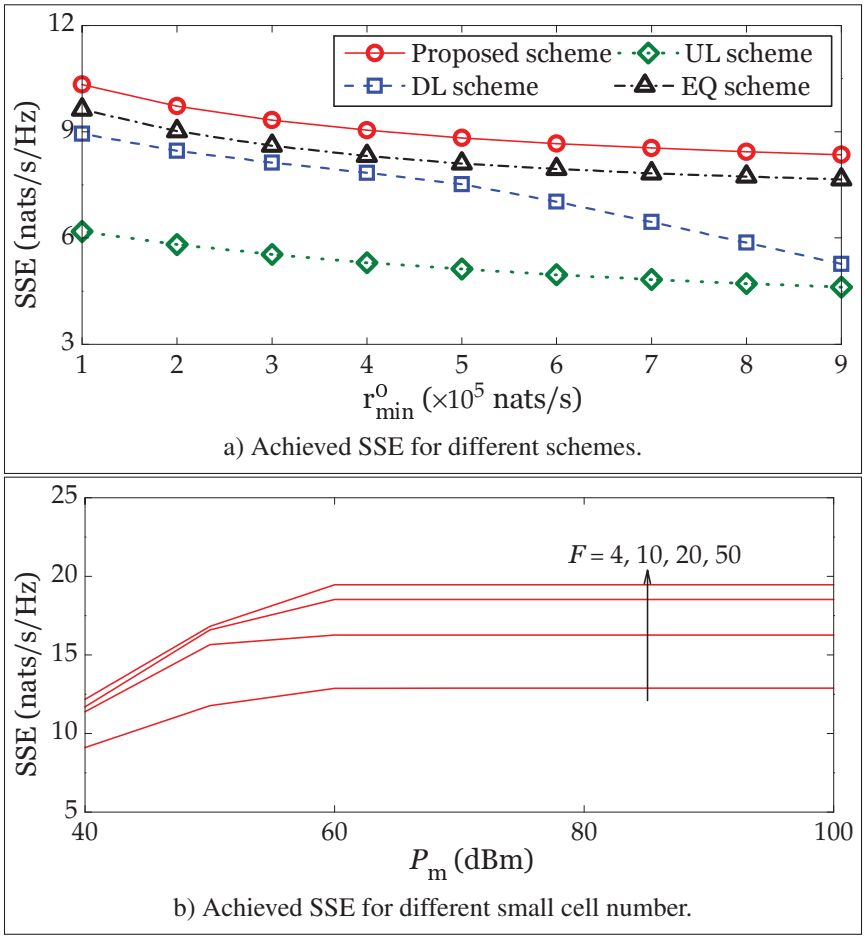

Figure 2.9 SSE of the network with respect to r^{o}_{min} and P_m.

CHAPTER 3

CENTRALIZED AND DISTRIBUTED ENERGY EFFICIENCY DESIGNS IN WIRELESS BACKHAUL HETNETS

Tri Minh Nguyen[1], Animesh Yadava[2], Wessam Ajib[2], Chadi Assi[3]

[1] Département de Génie Électrique, École de Technologie Supérieure,
[2] Département d'Informatique, Université de Québec à Montréal
[3] Concordia Institute for Information Systems Engineering, Concordia University

This article was published in *IEEE Transaction on Wireless Communications* in July 2017 (Nguyen *et al.*, 2017a).

3.1 Introduction

The evolution and widespread acceptance of 5G networks has strongly relied on its promise of providing thousand-fold enhancement of network capacity (Andrews *et al.*, 2014). Among others, network densification is considered as a key candidate technology for achieving the desired capacity(Bhushan *et al.*, 2014). Not surprisingly, achieving maximum throughput with minimal energy consumption has recently become an extremely attractive area of research(Hu & Qian, 2014). According to (Auer *et al.*, 2011), almost 80% of the total network energy is consumed at base station (BS) sites; thus, saving energy on such dense networks simply translates to greener and more economical communications. Motivated by the need for energy efficiency operation and deployment, which also helps in lowering operational cost for mobile network operators and contributes less CO_2 emission to the environment, maintaining the optimal resource management is equally vital to attain the best system energy efficiency (EE) (Rao & Fapojuwo, 2014).

Currently, deploying more BSs generally exposes the challenging cost problem of installing more fiber backhaul links. To overcome this burden, wireless backhaul (WB) (Ge *et al.*, 2014) emerged to present a simple and viable solution to solve the expensive backhaul architecture installation obstacle in dense networks. Unlike the conventional wired backhaul architecture in small cell networks where information is transported via fiber links, WB enables these small

cell access points (SAPs) to transmit and receive backhaul data over-the-air. According to (Siddique *et al.*, 2015a), it is sufficiently mature to activate the WB operation in the sub–6 GHz spectrum band with the available hardware. By replacing (or coexisting with) the fiber connection, the operators can confide WB to solve the problem of installation and difficulties for the backhaul deployment in urban and some rural areas. However, WB communication (WBC) must guarantee both high speed and reliable backhaul transmissions to maintain a certain level of quality of service (QoS). In light of these observations, it is imperative to consider an EE design for WB.

3.1.1 Related work

WB concept was first proposed in *IEEE* 802.16 mesh networks (Viswanathan & Mukherjee, 2006). By enabling WB transmission, the authors in (Viswanathan & Mukherjee, 2006) developed an algorithm based on linear optimization that maximizes the total network throughput to design an optimal routing and scheduling strategy for the medium access control layer. In the 5G context, WB technology was widely revisited in various network scenarios and, interestingly, extended to different spectrum bands (Siddique *et al.*, 2015a). Specifically, in the 60–80 GHz band, known as mm-wave, the authors of (Hur *et al.*, 2013b) presented a novel idea for small cells equipped with mm-wave transmitters to effectively align their transmit beam under wind induced impairments. In the sub–6 GHz band, some works proposed to reuse the available hardware and implementation to accommodate WBC concurrently with the wireless access communications (WAC). By employing the reverse time division duplex (RTDD) model, the authors in (Sanguinetti *et al.*, 2015; Wang *et al.*, 2016) considered the joint bandwidth allocation and user association that maximizes the achievable downlink (DL) sum log rate of small cells when the macrocell base station (MBS) is equipped with large antenna. In (Zhao *et al.*, 2015), the authors studied the admission control of SAPs in order to permit WB to serve as many SAPs as possible while guaranteeing QoS rates. The authors of (Niu *et al.*, 2017) were the first ones to consider the optimization of scheduling and power control that maximizes the

mm-wave WB EE communication. The authors of (Yang *et al.*, 2016b) considered the WB network and analyzed the EE of the small cells using stochastic geometry.

In general, achieving EE communications for green 5G heterogeneous networks (HetNets) has drawn significant research attention in recent years (Hu & Qian, 2014; I *et al.*, 2014; Soh *et al.*, 2013). Specifically, optimal beamforming and zero-forcing based beamforming design for maximizing the EE in multicell multiuser HetNets were studied in (He *et al.*, 2014) and (Xu *et al.*, 2014), respectively. In (Ng *et al.*, 2012), the authors addressed the joint power allocation and antenna design that maximizes the system EE in OFDMA massive MIMO networks. However, these works only aimed to develop centralized algorithms with full coordination between cells. This approach often requires high computational cost and signaling overhead since each node needs to have the full knowledge of the global channel state information (CSI) to compute the joint optimal solution. To overcome these difficulties, the authors in (Huang *et al.*, 2014) proposed a decentralized algorithm based on alternative optimization to design the beamforming that maximizes the system EE, where limited information exchange between uncoordinated BSs is assumed. The work in (Ngo *et al.*, 2014a) employs the framework of dual decomposition in (Palomar & Chiang, 2006a) to decompose the coupled problem into smaller problems to iteratively solve and update the involved parameters until convergence. To enhance the convergence behavior of the dual decomposition method, Kaleva *et al.* (2016) develops a decentralized algorithm based on alternative direction multiplier method (ADMM).

When there are more antennas at each small cell, the ratio of power for maintaining radio frequency chains or switching the BS ON/OFF to the total power consumption becomes more important since it is now comparable to the transmit power. This motivated the authors of (Zhou *et al.*) to consider BS sleep/active or in (Tervo *et al.*, 2015b) for antenna selection to refine the achieved EE in comparison with fixed circuit power. Moreover, at the SAP, the energy consumed in decoding the collected WB data from the MBS cannot be ignored. This is due to the fact that the SAPs have small range of operation and the energy consumed in the decoding process is comparable to the one used in transmitting the data (Rubio & Pascual-Iserte, 2014). For the DL of WB small cell HetNets, each SAP relates its backhaul rate proportionally to the

amount of power to decode the backhaul messages before forwarding them to its own users. Hence, this decoding power is fundamentally important (Rubio & Pascual-Iserte, 2014) and should be considered in the problem of EE optimization.

3.1.2 Contribution

In this paper, we consider the two-tier small cell HetNets that operates WBC in the sub–6 GHz band. In particular, WBC between multi-antenna MBS and single antenna SAPs are considered among conventional WAC of MBS-macrocell user equipments (MUEs) and SAPs-small cell users equipments (SUEs). To completely remove the self interference at the SAP, we apply the RTDD combined with equal spectrum splitting in each time slot to separate transmissions between WBC and WAC for interference mitigation. Based on that, we study the joint beamforming and power allocation optimization (JBPAO) design that maximizes a proposed metric, the access EE (AEE), which is defined by the ratio of the total wireless access spectral efficiency achieved at the MUEs and SUEs to the total power consumption. Unlike traditional works (e.g., (He *et al.*, 2014; Xu *et al.*, 2014; Venturino *et al.*, 2014; Tervo *et al.*, 2015b)) where a fixed value of decoding power is assigned at each BS, we emphasize the importance of the adaptive decoding power which is a result of considering WBC in our model. With this consideration in the formulation of AEE, the overall power consumption is more appropriately addressed, which leads to some improvement in the achieved EE. The contributions of the paper are summarized as follows:

- We formulate the problem of maximizing the access EE under the WB small cell HetNets. The model of total power consumption proposed here considers not only the transmit power but also the adaptive decoding power at each SAP as a function of its achievable rate, and therefore is more appropriate.
- The formulated optimization problem is non-convex and NP-hard. To solve it optimally, we present an exhaustive search based on customized branch-and-bound algorithm to achieve the global optimal solution for transmit beamforming and power allocation. The achieved result can be used as the network performance benchmark.

- To overcome the high computational complexity of the branch-and-bound algorithm, we propose a low complexity efficient algorithm based on first order Taylor convex approximation (FOTCA) which helps in achieving a local optimal solution by iteratively solving the convex approximated problem of the non-convex problem. The final solution is achieved through few iterations, which is much faster than the branch-and-bound algorithm.
- This EE problem formulation is further extended to consider the small cell selection that takes into account the impact of power to switch ON/OFF each SAP. By simply introducing binary variables to represent the active/idle status of each SAP, we follow the same step of the FOTCA based framework to drive the new problem into a mixed integer second order cone programming (MISOCP) approximated problem, where the developed algorithm can be applied to efficiently obtain the solution.
- We further propose to decompose the convex approximated problem into smaller sub-problems by the ADMM approach and develop a distributed algorithm to enable independent computation at each SAP and MBS. Specifically, each SAP and MBS will only have to solve its own sub-problem with local CSI by exchanging limited information with other BSs to update each cell parameters at each iteration until the algorithm converges.

The rest of the paper is organized as follows. Section 3.2 introduces the system model and the AEE. Section 3.3 formulates the optimization problem that maximizes the access EE on the DL of the WB HetNets and proposes two algorithms to solve it. In Section 3.4, a distributed algorithm is developed to solve the optimization problem. Section 3.5 extends the study of the previous AEE problem with SAP selection. Section 3.6 presents our numerical results under different simulation setups and discussions and the concluding remark are given in Section 3.7. A list of acronyms used in this paper is given in Table 3.1.

Table 3.1 List of acronyms.

Acronym	Description n
ADMM	Alternating direction method of multipliers
EE, AEE	Energy efficiency, Access EE
AO	Alternating optimization
AWGN	Additive white Gaussian noise
BnB	Branch and Bound
CSI	Channel state information
DL (UL)	Downlink (Uplink)
FOTCA	First order Taylor convex approximation
HetNets	Heterogeneous networks
JBPAO	Joint beamforming power allocation optimization
KKT	Karush-Kuhn-Tucker
MBS (BS)	Macrocell base station (Base station)
MISOCP	Mixed integer SOCP
MUE	Macrocell user equipment
QoS	Quality-of-service
RTDD	Reverse time division duplexing
SLNR	Signal-to-leakage-plus-noise ratio
SAP	Small cell access point
SUE	Small cell user equipment
WMMSE	Weighted minimum mean square error
WA (WAC)	Wireless access (WA communication)
WB (WBC)	Wireless backhaul (WB communication)

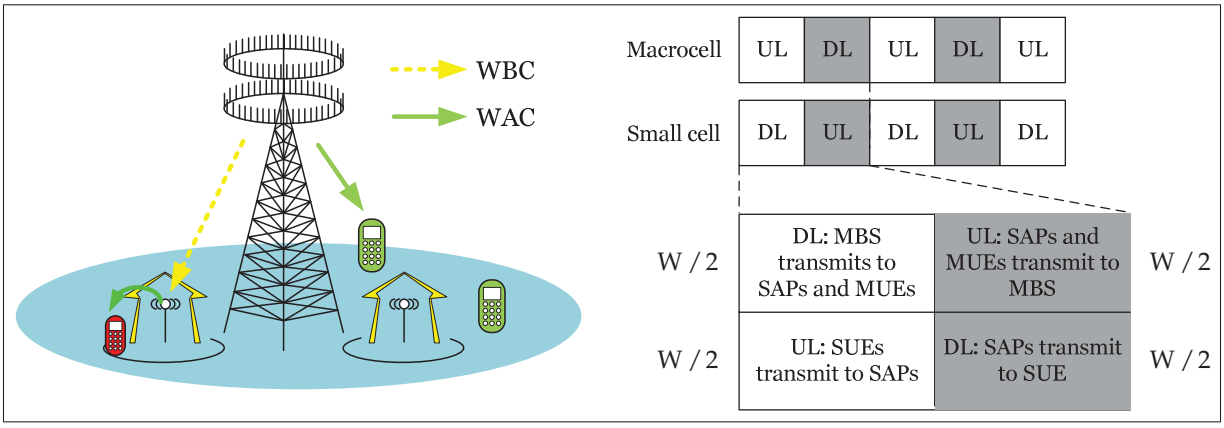

Figure 3.1 Model of two-tier DL wireless backhaul small cell HetNets.

3.2 System Model

3.2.1 Spatial Model

We consider the downlink of a two-tier HetNet consisting of one MBS of N antennas in the macrocell tier and S_a single antenna SAPs in the small cell tier, as depicted in Fig. 3.1. The MBS serves M MUEs and each SAP serves the same number of S_u SUEs within their coverage. The general case of different number of SUEs in different small cells can be intuitively studied by following the same framework. Here, the wireless access transmissions refer to the transmissions from MBS to MUEs and from SAPs to SUEs, whereas the wireless backhaul transmissions refer to the ones from the MBS to the SAPs. Both WA and WB transmissions operate concurrently in the same resource block. For brevity, we assume that the signal transmissions from the MBS to the SAPs is modeled similarly to the ones from the MBS to MUEs.

3.2.2 Reverse time division duplex (RTDD)

We assume that all the transmissions operate under the RTDD system (Sanguinetti *et al.*, 2015). Unlike traditional time division duplexing, RTDD employs a reverse setting of UL and DL time slots in two tiers and splits the available spectrum of W Hertz in two equal parts to accommodate different transmissions. As shown by Fig. 3.1, the MBS can now transmit its DL signals on an orthogonal time-spectrum resource block with the UL signals from the SUEs to their serving SAPs. Similarly, the MUEs and SAPs can transmit UL signals to the MBS orthogonally with the DL signals from the SAPs. With this interference management, the in-band half-duplex SAPs can avoid self-interference when WBC coexists with WAC as in (Sanguinetti *et al.*, 2015).

3.2.3 Signal model

We assume the channel is flat over the spectrum and time-invariant within the coherence time T_c where T_c is larger than the duration of one time slot. We first consider the time and frequency

block dedicated for the macrocell DL, where the MBS transmits signals to its MUEs and SAPs as shown in Fig. 3.1. By denoting $\mathscr{F} = \{\mathscr{S}_a, \mathscr{M}\} = \{\{1,\ldots,S_a\},\{S_a+1,\ldots,S_a+M\}\}$, the received signal within this time-frequency block at the jth receiver is

$$y_j = \mathbf{v}_j^H \mathbf{h}_j x_j + \sum_{k\in\mathscr{F}\setminus j} \mathbf{v}_k^H \mathbf{h}_j x_k + n_j, \tag{3.1}$$

where $\mathbf{h}_j \in \mathbb{C}^{N\times 1}$ is the channel state vector which includes fading gain and pathloss components and $\mathbf{v}_j \in \mathbb{C}^{N\times 1}$ is the beamforming vector from the MBS to the jth receiver. x_j is the message intended for the jth receiver with unit average power, e.g., $\mathbb{E}\left\{x_j x_j^*\right\} = 1$ and n_j is a circularly symmetric complex additive white Gaussian noise (AWGN) at the jth receiver, which is distributed according to a normal distribution $\mathscr{CN}(0,N_0)$, where N_0 is the noise power over the allocated spectrum. Treating interference as noise, the achievable rate at the jth receiver is

$$R_j = \log\left(1+\Gamma_j\right) = \log\left(1+\frac{\left|\mathbf{v}_j^H \mathbf{h}_j\right|^2}{\sum\limits_{k\in\mathscr{F}\setminus j}\left|\mathbf{v}_k^H \mathbf{h}_j\right|^2 + N_0}\right), \tag{3.2}$$

where we denote the set of transmit beamforming as $\mathbf{v} \triangleq \{\mathbf{v}_i, \forall i \in \mathscr{F}\}$ for later usages. In the other time-frequency block dedicated for small cell DL, each ith SAP transmits data to its S_u intended SUEs in the ith small cell. Let us denote p_{ij} as the transmit power from the SAP to the jth SUE in the ith small cell and h_{ikj} as the channel from the ith SAP to the jth SUE in the kth small cell; $\mathbf{p} = [\mathbf{p}_1,\ldots,\mathbf{p}_{S_a}]^T$, where $\mathbf{p}_i = [p_{i1},\ldots,p_{iS_u}]^T$ is the set of transmit powers of the SAPs. Hence, the received signal within this time-frequency block at the jth SUE in the ith small cell is

$$y_{ij} = h_{iij}\sqrt{p_{ij}}s_{ij} + \sum_{k\in\mathscr{S}_u\setminus j} h_{iij}\sqrt{p_{ik}}s_{ik} + \sum_{k\in\mathscr{S}_a\setminus i}\sum_{l\in\mathscr{S}_u} h_{kij}\sqrt{p_{kl}}s_{kl} + z_{ij}, \tag{3.3}$$

where s_{ij} is the message intended for the jth SUE from the SAP in the ith small cell with unit average power, e.g., $\mathbb{E}\left\{s_{ij}s_{ij}^*\right\} = 1$. Similarly, z_{ij} is a circularly symmetric complex AWGN at the SUE from the SAP in the ith small cell, distributed according to a normal distribution $\mathscr{CN}(0,N_0)$. Treating interference as noise, the achievable rate at the jth SUE in ith small cell

is $r_{ij} = \log\left(1+\gamma_{ij}\right)$, where

$$\gamma_{ij} = \frac{p_{ij}\left|h_{iij}\right|^2}{\sum_{k\in\mathscr{S}_u\setminus j} p_{ik}\left|h_{iij}\right|^2 + \sum_{k\in\mathscr{S}_a\setminus i}\sum_{l\in\mathscr{S}_u} p_{kl}\left|h_{kij}\right|^2 + N_0}. \tag{3.4}$$

3.2.4 Access Energy Efficiency (AEE)

In this subsection, we introduce an EE expression for the considered system. Let us define the overall network power consumption P_{tot} as

$$P_{\text{tot}} = \kappa_m \sum_{i\in\mathscr{F}} \|\mathbf{v}_i\|^2 + \underbrace{P_m^{\text{active}} + P_m^{\text{diss}} + P_m^{\text{dec}}}_{P_m^{\text{circ}}} + \kappa_s \sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} p_{ij} + \sum_{i\in\mathscr{S}_a}\left[\underbrace{\overbrace{\left(P_s^{\text{active}} + P_s^{\text{diss}}\right)}^{P_s^{\text{ad}}} + P_{s,i}^{\text{dec}}}_{P_{s,i}^{\text{circ}}}\right], \tag{3.5}$$

where $\kappa_m, \kappa_s > 1$ are the constants which account for the inefficiency of the power amplifier at the MBS and each SAP when they transmit on the DL. P_m^{circ} and $P_{s,i}^{\text{circ}}$ represent the circuit power at the MBS and the ith SAP, respectively. Specifically, P_m^{circ} includes the power that keeps the MBS active P_m^{active}, the power dissipation in the transmitting filter, mixer, frequency synthesizer and digital-to-analog converter P_m^{diss}, and the power necessary for decoding at the MBS P_m^{dec}. For simplicity, we assume that P_m^{active} and P_m^{diss} remain constant at all times. Moreover, since we consider the DL of the MBS, we also assume the MBS always decodes the same amount of information at all time slots so that P_m^{dec} is also constant. Thus, P_m^{circ} should take the same value for all considered time slots. On the other hand, the decoding term $P_{s,i}^{\text{dec}}$, should be considered more carefully in the small cell tier. Since each ith SAP receives backhaul data from the MBS via the wireless channel, it should decode the received signals before transmitting them to its SUEs. Thus, based on the consensus that the decoding power used for this process increases with the achievable rate (Rubio & Pascual-Iserte, 2014) R_i at each ith SAP, we apply the model in (Rubio & Pascual-Iserte, 2014) to formulate the decoding power as a linear function of the

achievable rate, e.g.,

$$P_{s,i}^{\text{dec}} = \alpha_i R_i, \forall i \in \mathscr{S}_a. \tag{3.6}$$

Next, we propose an expression for the computation of AEE, defined as the ratio between the sum rate at all the MUEs and SUEs and the total network power consumption. In particular, the AEE (computed in bits/Joule/Hz) can be expressed as

$$\eta = \frac{\sum_{k\in\mathscr{M}} R_k + \sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} r_{ij}}{P_{\text{tot}}}. \tag{3.7}$$

Discussion: This above definition of the EE metric is widely used in the literature as the major metric to evaluate green communications and is often proposed in a single cell cellular network where the behavior of one cell is of most interest. In the context of multicell HetNets, beside this definition, there also exists other metrics to evaluate the network EE, such as weighted sum EE η_1 or product of individual EE η_2(Lasaulce *et al.*, 2009; Miao *et al.*, 2011; Venturino & Buzzi, 2015; Xiong *et al.*, 2014; Venturino *et al.*, 2014) as follows:

$$\eta = \begin{cases} \eta_1 = \omega_m \frac{\sum_{k\in\mathscr{M}} R_k}{\kappa_m \sum_{i\in\mathscr{F}} \|\mathbf{v}_i\|^2 + P_m^{\text{circ}}} \\ + \sum_{i\in\mathscr{S}_a} \omega_{s,i} \frac{\sum_{j\in\mathscr{S}_u} r_{ij}}{\kappa_s \sum_{j\in\mathscr{S}_u} p_{ij} + P_{s,i}^{\text{circ}}}, \\ \eta_2 = \left(\frac{\sum_{k\in\mathscr{M}} R_k}{\kappa_m \sum_{i\in\mathscr{F}} \|\mathbf{v}_i\|^2 + P_m^{\text{circ}}}\right)^{\omega_m} \\ \times \prod_{i\in\mathscr{S}_a} \left(\frac{\sum_{j\in\mathscr{S}_u} r_{ij}}{\kappa_s \sum_{j\in\mathscr{S}_u} p_{ij} + P_{s,i}^{\text{circ}}}\right)^{\omega_{s,i}}. \end{cases} \tag{3.8}$$

These formulations offer more flexibility to directly prioritize the EE at an individual cell rather than relying on indirectly weighting the individual rate at each user from the definition in (3.7). With different minimum rate QoS constraints and disparate power budgets for each tier, setting higher weight for the potential BS might allow to retrieve better energy-efficient resource allocation solution for that individual BS and attain suitable optimal EE performance as expected from the given weight setting. To this end, we observe that regardless the choice of EE metrics, the following analytical and algorithmic frameworks developed next can also be applied with (3.8) with very simple manipulation.

3.3 AEE Optimization Problem

3.3.1 Problem formulation

The constrained optimization problem of maximizing the AEE can be formulated as follows:

$$(\mathscr{P}): \max_{\mathbf{v},\mathbf{p}\geq 0} \frac{\sum_{k\in\mathscr{M}} R_k + \sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} r_{ij}}{P_{\text{tot}}} \tag{3.9a}$$

$$\text{s.t. } R_i \geq \sum_{j\in\mathscr{S}_u} r_{ij}, \forall i \in \mathscr{S}_a, \tag{3.9b}$$

$$R_k \geq r^m_{\min}, \forall k \in \mathscr{M}; \tag{3.9c}$$

$$r_{ij} \geq r^s_{\min}, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.9d}$$

$$\sum_{k\in\mathscr{F}} \|\mathbf{v}_k\|^2 \leq P_{\max}, \tag{3.9e}$$

$$\sum_{j\in\mathscr{S}_u} p_{ij} \leq p_{\max}, \forall i \in \mathscr{S}_a. \tag{3.9f}$$

To convey the property of the WB networks into this optimization problem, constraint (3.9b) enforces that the backhaul rate of each SAP should always be larger than the corresponding access rate at each SAP on the DL. The backhaul rate in this case becomes a limiting factor for the access rate at the SAP. Even if there are good channel conditions on the DL of the SAP, without sufficiently good WB link, the corresponding SAP cannot achieve higher rates than that regulated by the backhaul rate, and thus restrains the optimal AEE. (3.9b) also renders a complicated coupled relationship between variables $\mathbf{v}$ and $\mathbf{p}$ by nonlinear non-convex functions, e.g., R_i, r_{ij}, which makes it more difficult to solve compared to conventional minimum constant SINR requirement constraint used in multiuser networks. The minimum rate requirements at the MUEs and SUEs are given by (3.9c). (3.9e)–(3.9f) imply that MBS and SAP cannot transmit more than their maximum power budget $P_{\max}$ and $p_{\max}$, respectively. The following proposition characterizes the relationship between the achievable rate of the WBC and WAC.

Proposition 1. *Given the AEE constrained optimization problem as in* (3.9)*, all the inequalities* (3.9b) *become equalities at optimality.*

Proof. We prove Proposition 1 by contradiction. Assuming that at the optimal solution $\mathbf{v}^\star, \mathbf{p}^\star$, (3.9b) occurs at strict inequalities at some index $i \in \mathscr{S}_a$. Then, we can find a new value $\tilde{\mathbf{v}}_i = \varepsilon_i \mathbf{v}_i^\star$, $\varepsilon_i < 1$ so that $R_i, \forall i \in \mathscr{S}_a$ will slightly decrease to occur at equality. On the other hand, note that the numerator of objective function only contains the MUE rate $R_k, \forall k \in \mathscr{M}$. Thus, the scaled down version of $\mathbf{v}_i^\star, \forall i \in \mathscr{S}_a$ causes less interference to each $R_k, k \in \mathscr{M}$ and leads to the increase of the numerator of (2.7a). In addition, the denominator of (3.9a) decreases with the scaled down version of $\mathbf{v}_i^\star, \forall i \in \mathscr{S}_a$ and decoding power proportionally decreases with $R_i, \forall i \in \mathscr{S}_a$. These improves the value of objective function, which contradicts the assumption of optimality. This completes the proof. □

It is easy to observe that (3.9) is a non-convex and an NP-hard optimization problem due to the existence of R_i and r_{ij} on the numerator of (3.9a) as the function variables $\mathbf{v}$ and $\mathbf{p}$. The NP-hardness proof is sketched in Appendix 5. Note that since we consider an interference term in each rate formula, R_i and r_{ij} are neither convex nor concave functions with respect to the relevant variables. The appearance of R_i in constraint (3.9b) and in the denominator of (3.9a) introduces an additional non-convex factor to the entire optimization problem so that conventional nonlinear fractional programming method for EE problem used in (He *et al.*, 2014; Xu *et al.*, 2014; Ng *et al.*, 2012; Tervo *et al.*, 2015b) cannot be straightforwardly applied here. Toward this end, we rely on the result in Proposition 1 that at optimality, $R_i = \sum_{j \in \mathscr{S}_u} r_{ij}, \forall i \in \mathscr{S}_a$ to instead compute $P_{s,i}^{\text{dec}}$ as a function of $\sum_{j \in \mathscr{S}_u} r_{ij}$, e.g., $P_{s,i}^{\text{dec}} = \alpha_i \sum_{j \in \mathscr{S}_u} r_{ij}$. Based on this, we propose another optimization problem as follows:

$$(\mathscr{P}_1) : \max_{\mathbf{v},\mathbf{p} \geq 0} \left\{ \frac{\sum_{k \in \mathscr{M}} R_k + \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} r_{ij}}{\tilde{P}_{\text{tot}}} \middle| (3.9\text{b}) - (3.9\text{f}) \right\}. \tag{3.10}$$

where

$$\tilde{P}_{\text{tot}} = \kappa_m \sum_{i \in \mathscr{F}} \|\mathbf{v}_i\|^2 + P_m^{\text{circ}} + \kappa_s \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} p_{ij} + \sum_{i \in \mathscr{S}_a} \left[\left(P_s^{\text{active}} + P_s^{\text{diss}} \right) + \alpha_i \sum_{j \in \mathscr{S}_u} r_{ij} \right].$$

The purpose of visiting this problem will prove helpful in the next section to achieve the global optimal solution through a branch-and-bound algorithm. Here, we observe that depending on

the given values of $r^m_{\min}, r^s_{\min}, P_{\max}, p_{\max}$, (3.10) can be infeasible due to the minimum rate and power constraints (3.9c)–(3.9f). Setting a power budget to be low may cause infeasibility to (3.10) when minimum rate requirements $r^m_{\min}, r^s_{\min}$ are high. Moreover, posing high $r^m_{\min}, r^s_{\min}$ in an interference–limited operational regime can also make (3.10) infeasible despite having high power budget. To facilitate this issue, we only consider the feasible problem with sufficiently small $r^m_{\min}, r^s_{\min}$ setting in this paper. Furthermore, to guarantee the convergence of the iterative algorithm developed later, we assume that all the feasible solutions of (3.10) are regular as in (Scutari *et al.*, 2017). The condition of a regular solution of an optimization problem is sketched in Appendix 4.1.

3.3.2 Branch-and-bound algorithm for global optimal solution

In this subsection, we optimally solve (3.10) using a branch-and-bound (BnB) based algorithm. First, we rewrite (3.9) into a more tractable form as

$$(\mathscr{P}_{bnb}): \max_{\substack{\mathbf{v},\mathbf{p}\geq 0,\mathbf{t}\geq 0,\\ \underline{\tau}\leq\tau\leq\bar{\tau}}} \ell(\mathbf{t},\tau) = \left(\sum_{k\in\mathscr{M}} t^m_k + \sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} t^s_{ij}\right)\cdot\tau \tag{3.11a}$$

$$\text{s.t. } \tilde{P}_{\text{tot}} \leq \frac{1}{\tau}, \tag{3.11b}$$

$$R_i \geq \sum_{j\in\mathscr{S}_u} t^s_{ij}, \forall i \in \mathscr{S}_a, \tag{3.11c}$$

$$R_k \geq t^m_k, \forall j \in \mathscr{M}, \tag{3.11d}$$

$$t^m_k \geq r^m_{\min}, \forall k \in \mathscr{M}, \tag{3.11e}$$

$$t^s_{ij} \geq r^s_{\min}, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.11f}$$

$$r_{ij} = t^s_{ij}, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.11g}$$

$$(3.9e) - (3.9f), \tag{3.11h}$$

where $t_i^m \geq 0, \forall i \in \mathscr{F}$, $t_{ij}^s \geq 0, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u$, and $\underline{\tau} \leq \tau \leq \bar{\tau}$ are the slack variables introduced to facilitate the tractability of the optimal solution. The definitions of $\underline{\tau}$ and $\bar{\tau}$ are:

$$\underline{\tau} = \frac{1}{\kappa_m P_{\max} + P_m^{\text{circ}} + \kappa_s S_a p_{\max} + S_a P_s^{\text{ad}} + \bar{t}_{ij}^s} \tag{3.12}$$

$$\bar{\tau} = \frac{1}{P_m^{\text{circ}} + S_a P_s^{\text{ad}}}, \tag{3.13}$$

where $\bar{t}_{ij}^s$ will be defined shortly later. For convenience, we denote $\mathbf{t} = \left[(\mathbf{t}^m)^T, (\mathbf{t}^s)^T\right]^T$, where $\mathbf{t}^m = \left[\left(\mathbf{t}^{mb}\right)^T, (\mathbf{t}^{ma})^T\right] = \left[\left[t_1^m, \ldots, t_{S_a}^m\right]^T, \left[t_{S_a+1}^m, \ldots, t_{S_a+M}^m\right]^T\right]^T$, and $\mathbf{t}^s = \left[\left(\mathbf{t}_1^s\right)^T, \ldots, \left(\mathbf{t}_{S_a}^s\right)^T\right]^T$, $\mathbf{t}_i^s = \left[t_{i1}^s, \ldots, t_{iS_u}^s\right]^T$. The equivalence between (3.11) and (3.9) can be shown by following the same line of proof for Proposition 1. Here, we note two important observations. Firstly, when each of $t_j^m, \forall j \in \mathscr{M}$, $t_{ij}^s, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u$, or τ increases within its feasible domain, (3.11) achieves a higher valued objective function. Secondly, when all of $t_j^m, \forall j \in \mathscr{M}$, $t_{ij}^s, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u$, and τ are fixed, (3.11) becomes a convex feasibility checking optimization problem. Thus, we can employ the concept of monotonic optimization (Tervo *et al.*, 2015b) to customize the BnB algorithm in order to optimally solve (3.11). Before providing the detailed algorithm, we use the results and concepts defined in (Tervo *et al.*, 2015b) to present some definitions and notations to support the presentation of the customized BnB algorithm.

Let $\Omega = [\mathbf{t}, \tau]^T$ denote the set of variables $\mathbf{t}, \tau$. We define the normal compact set $\mathscr{D} = \{\Omega | (3.11\text{b}) - (3.11\text{h})\}$ by following the same definition as in (Tervo *et al.*, 2015b). On the other hand, we denote $\mathscr{C} = \left[\underline{\Omega}, \bar{\Omega}\right]$ as the box that contains all the feasible solutions of Ω in (3.11). From (3.11e), we observe that $t_j^m \geq r_{\min}^m, \forall j \in \mathscr{M}$ and $t_{ij}^s \geq r_{\min}^s, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u$. Thus, the lower bound of Ω can be computed as $\underline{\Omega} = \left[r_{\min}^m, \ldots, r_{\min}^s, \underline{\tau}\right]^T$. In addition, we can simply compute the upper bound of Ω as $\bar{\Omega} = \left[\bar{t}_1^m, \ldots, \bar{t}_{S_a S_u}^s, \bar{\tau}\right]^T$, where

$$\begin{aligned} t_j^m &\leq \log\left(1 + \left\|\mathbf{v}_j\right\|^2 \left\|\mathbf{h}_j\right\|^2\right) \\ &\leq \log\left(1 + P_{\max}\left\|\mathbf{h}_j\right\|^2\right) = \bar{t}_j^m, \forall j \in \mathscr{M}, \\ t_{ij}^s &\leq \log\left(1 + p_{ij}\left|h_{iij}\right|^2\right) \end{aligned} \tag{3.14}$$

$$\leq \log\left(1+p_{\max}\left|h_{iij}\right|^2\right)=\bar{r}_{ij}^{s},\forall i\in\mathscr{S}_a,\forall j\in\mathscr{S}_u. \tag{3.15}$$

Now, given the value of Ω, the problem of checking whether $\Omega\in\mathscr{C}$ is reduced to a convex feasibility checking problem to find $\{\mathbf{v},\mathbf{p}\geq 0|(3.11\text{b})-(3.11\text{h})\}$. From these discussions, problem (3.11) can be rewritten as $\max\{\ell(\mathbf{t},\tau)\,|\Omega\in\mathscr{C}\subset\mathscr{D}\}$. First we check if $\underline{\Omega}$ is feasible or not. If feasible, we customize a BnB algorithm to solve (3.11) globally. This algorithm recursively branches the box $\mathscr{D}$ into smaller boxes, checks the feasibility of each new box, updates the new upper and lower bounds by the box reduction—bound computation process and disposes the boxes that do not contain the optimal solution. The algorithm terminates when the difference between the upper and lower bounds is arbitrarily small and the global optimal solution is determined. Since the protocol of the BnB algorithm is similar to the content in (Tervo *et al.*, 2015b), its detailed presentation is omitted here for brevity. The steps of box branching, box reduction can also be found in (Tervo *et al.*, 2015b). Note that for a given set $\mathscr{D}_i=\left[\underline{\Omega}_i,\bar{\Omega}_i\right]$, where $\underline{\Omega}_i=\left[\underline{\mathbf{t}}_i,\underline{\tau}_i\right]^T$ and $\bar{\Omega}_i=\left[\bar{\mathbf{t}}_i,\bar{\tau}_i\right]^T$, the lower and upper bounds computation are given by $L(\mathscr{D}_i)=\ell(\underline{\mathbf{t}}_i,\underline{\tau}_i)$ and $U(\mathscr{D}_i)=\ell\left(\bar{\mathbf{t}}_i,\bar{\tau}_i\right)$, respectively.

3.3.3 Low complexity FOTCA-based algorithm

The BnB algorithm basically requires very high computational cost to yield the global optimal solution. This solution approach is not practical and is only helpful for benchmark evaluation. In this section, we develop a low complexity and efficient algorithm to solve (3.10) based on multiple equivalent transformations and convex approximation methods. When applying the convex approximation, we rely on the result in (Marks & Wright) to associate the three conditions which can be employed to verify the application of the proposed convex approximation. These three conditions guarantee that the developed iterative algorithm based on these approximations will converge after much less iteration than the BnB algorithm. The following proposition is provided to transform (3.10) into a more equivalent tractable form to develop the low complexity algorithm to reach the solution.

Proposition 2. *By introducing slack variables* $t_i^m \geq 0, \forall i \in \mathscr{F}$, $s_{ij} \geq 0, t_{ij}^s \geq 0, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u$, $\tau \geq 0$, *and* $q \geq 0$, *problem* (3.10) *is equivalent to the following optimization problem* $\left(\mathscr{P}_{eq}\right)$

$$\max_{\Pi} \; q^2 \tag{3.16a}$$

$$s.t. \;\; \left(\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^s\right) \cdot \tau \geq q^2, \tag{3.16b}$$

$$\tilde{\tilde{P}}_{tot} \leq \frac{1}{\tau}, \tag{3.16c}$$

$$\log\left(1+\nu_i\right) \geq t_i^m, \forall i \in \mathscr{F}, \tag{3.16d}$$

$$t_i^m \geq \sum_{j \in \mathscr{S}_u} s_{ij}, \forall i \in \mathscr{S}_a, \tag{3.16e}$$

$$s_{ij} \geq \log\left(1+\gamma_{ij}\right) \geq t_{ij}^s, \forall \left(i,j\right) \in \left(\mathscr{S}_a, \mathscr{S}_u\right), \tag{3.16f}$$

$$\frac{\left|\mathbf{v}_i^H \mathbf{h}_i\right|^2}{\sum\limits_{k \in \mathscr{F} \setminus i} \left|\mathbf{v}_k^H \mathbf{h}_i\right|^2 + N_0} \geq \nu_i, \forall i \in \mathscr{F}, \tag{3.16g}$$

$$t_k^m \geq r_{\min}^m \forall k \in \mathscr{M}, \tag{3.16h}$$

$$t_{ij}^s \geq r_{\min}^s, \forall \left(i,j\right) \in \left(\mathscr{S}_a, \mathscr{S}_u\right), \tag{3.16i}$$

$$(3.9e) - (3.9f), \tag{3.16j}$$

where

$$\tilde{\tilde{P}}_{tot} = \kappa_m \sum_{i \in \mathscr{F}} \left\|\mathbf{v}_i\right\|^2 + P_m^{circ} + \kappa_s \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} p_{ij} + \sum_{i \in \mathscr{S}_a} \left[\left(P_s^{active} + P_s^{diss}\right) + \alpha_i \sum_{j \in \mathscr{S}_u} s_{ij}\right]$$

and $\Pi = \{\mathbf{v}, \mathbf{p} \geq 0, \nu \geq 0, \mathbf{s} \geq 0, \mathbf{t} \geq 0, \tau \geq 0, q \geq 0\}$. *In addition,* $\mathbf{t}$ *is denoted as in* (3.11)*;* $\mathbf{s} = \left[\left(\mathbf{s}_1\right)^T, \ldots, \left(\mathbf{s}_{S_a}\right)^T\right]^T$, *where each* $\mathbf{s}_i = \left[s_{i1}, \ldots, s_{iS_u}\right]^T$*; and* $\nu = \left[\nu_1, \ldots, \nu_{S_a}, \nu_{S_a+1}, \ldots, \nu_{S_a+M}\right]^T$.

Proof. The proof is given in Appendix 3. □

We observe that the current optimization problem (3.16) is a maximization of a convex objective function, which is also a non-convex problem. However, the maximization of q^2 is equivalent to maximization of q. Since q is a linear function, the objective function of (3.16a)

is now a concave function. Thus, problem (3.16) can be equivalently rewritten as

$$\max_{\Pi} \{q | (3.16b) - (3.16j)\}. \tag{3.17}$$

Moreover, (3.16b) can be rewritten into a second order cone (SOC) constraint as

$$\frac{\tau + \left(\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^{s}\right)}{2} \geq \sqrt{q^2 + \frac{\left(\tau - \left(\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^{s}\right)\right)^2}{4}}. \tag{3.18}$$

Next, by taking a closer investigation at the constraints (3.16c) and (3.16g), we find that the factors that cause the non-convexity of (3.16c) and (3.16g) are on the greater side of the inequalities and share the same form of function $f(y,x) = \frac{|y|^2}{x}, \forall y \in \mathbb{C}^N, x \in \mathbb{R}^+$. In particular, we can rewrite (3.16g) and (3.16c) as

$$f\left(\mathbf{v}_i^H \mathbf{h}_i, v_i\right) \geq \sum_{k \in \mathscr{F} \setminus i} \left|\mathbf{v}_k^H \mathbf{h}_i\right|^2 + N_0, \forall i \in \mathscr{F}, \tag{3.19}$$

$$f(1, \tau) \geq \tilde{\tilde{P}}_{\text{tot}}. \tag{3.20}$$

At this point, we apply a generic first order Taylor approximation to approximate all the $f(\cdot,\cdot)$ function in (3.16c) and (3.16g). For example, the function $f(y,x)$ can be approximated around the point of $y^{(n)}, x^{(n)}$ by

$$F\left(y, x, y^{(n)}, x^{(n)}\right) = \frac{2\mathscr{R}\left(y^{(n)} y\right)}{x^{(n)}} - \frac{\left|y^{(n)}\right|^2}{(x^{(n)})^2} x. \tag{3.21}$$

Finally, we return at the non-convex constraint (3.16f). First we rewrite this constraint into two separate constraints as

$$s_{ij} + \log\left(I_{ij} + N_0\right) \geq \log\left(p_{ij} \left|h_{iij}\right|^2 + I_{ij} + N_0\right), \tag{3.22a}$$

$$t_{ij} + \log\left(I_{ij} + N_0\right) \leq \log\left(p_{ij} \left|h_{iij}\right|^2 + I_{ij} + N_0\right), \tag{3.22b}$$

It is obvious to state that both (3.22a) and (3.22b) are non-convex constraints. Thus, by denoting $g_{ij}^{SI}(\mathbf{p}) = p_{ij}\left|h_{iij}\right|^2 + I_{ij} + N_0$ and $g_{ij}^{I}(\mathbf{p}) = I_{ij} + N_0$, where

$$I_{ij} = \sum_{k \in \mathscr{S}_u \setminus j} p_{ik}\left|h_{iij}\right|^2 + \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} p_{kl}\left|h_{kij}\right|^2,$$

we apply the first order Taylor approximation to approximate the functions $\log\left(g_{ij}^{SI}(\mathbf{p})\right)$ and $\log\left(g_{ij}^{I}(\mathbf{p})\right)$ around the point $\mathbf{p}^{(n)}$ as

$$G_{ij}^{SI}\left(\mathbf{p}, \mathbf{p}^{(n)}\right) = \log\left(g_{ij}^{SI}\left(\mathbf{p}^{(n)}\right)\right) + \frac{g_{ij}^{SI}(\mathbf{p}) - g_{ij}^{SI}\left(\mathbf{p}^{(n)}\right)}{g_{ij}^{SI}\left(\mathbf{p}^{(n)}\right)}, \tag{3.23a}$$

$$G_{ij}^{I}\left(\mathbf{p}, \mathbf{p}^{(\mathbf{n})}\right) = \log\left(g_{ij}^{I}\left(\mathbf{p}^{(n)}\right)\right) + \frac{g_{ij}^{I}(\mathbf{p}) - g_{ij}^{I}\left(\mathbf{p}^{(n)}\right)}{g_{ij}^{I}\left(\mathbf{p}^{(n)}\right)}, \tag{3.23b}$$

By applying the approximations (3.21) and (3.23) into their corresponding non-convex constraints in (3.16c), (3.16e) and (3.16g), the problem in (3.17) can be solved by iteratively solving the following approximated convex problem $\left(\tilde{\mathscr{P}}^{(n+1)}\right)$, which is formulated at the $n+1$ index as

$$\max_{\Pi} \; q \tag{3.24a}$$

$$\text{s.t.} \;\; \frac{\tau + \mathbf{1}^T\mathbf{t}^{ma} + \mathbf{1}^T\mathbf{t}^{s}}{2} \geq \left\| q, \frac{\tau - \left(\mathbf{1}^T\mathbf{t}^{ma} + \mathbf{1}^T\mathbf{t}^{s}\right)}{2} \right\| \tag{3.24b}$$

$$F\left(1, \tau, \tau^{(n)}\right) \geq \tilde{\tilde{P}}_{\text{tot}}, \tag{3.24c}$$

$$F\left(\mathbf{v}_i^H\mathbf{h}_i, \nu_i, \mathbf{v}_i^{(n)}, \nu_i^{(n)}\right) \geq \sum_{k \in \mathscr{F} \setminus i} \left|\mathbf{v}_k^H\mathbf{h}_i\right|^2 + N_0, \forall i \in \mathscr{F}, \tag{3.24d}$$

$$s_{ij} + \log\left(I_{ij} + N_0\right) \geq G_{ij}^{SI}\left(\mathbf{p}, \mathbf{p}^{(n)}\right), \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.24e}$$

$$\log\left(p_{ij}\left|h_{iij}\right|^2 + I_{ij} + N_0\right) - t_{ij} \geq G_{ij}^{I}\left(\mathbf{p}, \mathbf{p}^{(\mathbf{n})}\right), \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.24f}$$

$$(3.16\text{d}), (3.16\text{e}), (3.16\text{h}), (3.16\text{j}). \tag{3.24g}$$

Note that $\mathbf{v}^{(n)}, \nu^{(n)}, \mathbf{p}^{(n)}, \tau^{(n)}$ are not the variables of the optimization problem but the parameters that are iteratively updated by the optimal solution after each iteration. Algorithm 2 presents the pseudocode to solve the optimization problem (3.17). The proof that Algorithm 2 converges is given in Appendix 4.2. Here, we only present the convergent property of the objective function. The convergent proof of the optimization variables to a limiting point satisfying the KKT conditions is more elaborated in (Venkatraman *et al.*, 2016a, Appendix A), which is omitted here due to space constraint.

3.4 Distributed FOTCA-based Algorithm

In the previous section, we developed a centralized algorithm to solve for the joint beamforming and power allocation at the MBS and SAPs. In practice, each SAP or MBS only knows its local CSI, e.g., the MBS only knows $\mathbf{h}_j, \forall j \in \mathscr{F}$ and the ith SAP only knows $h_{ijk}, \forall j \in \mathscr{S}_a, k \in \mathscr{S}_u$. Therefore, the centralized solution is not practical since there is insufficient CSI at each node to compute the joint solution. In this section, by assuming that there are limited message exchanges between MBS and SAPs, we develop a distributed algorithm based on the framework of decomposition and ADMM. In particular, we decompose the main convex approximated problem (3.24) into smaller sub-problems. Then, each SAP (and MBS) iteratively solves its local problem based on the local CSI and exchanges few information bits to update local parameters until convergence. To simplify the decomposition process, we first replace constraints (3.24b) and (3.24c) by

$$\frac{\tau + z_1}{2} \geq \sqrt{q^2 + \frac{(\tau - z_1)^2}{4}}, \tag{3.25}$$

$$\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^{s} \geq z_1, \tag{3.26}$$

$$F\left(1, \tau, \tau^{(n)}\right) \geq \kappa_m \sum_{i \in \mathscr{F}} \|\mathbf{v}_i\|^2 + P_m^{\text{circ}} + z_2, \tag{3.27}$$

$$z_2 \geq \kappa_s \mathbf{1}^T \mathbf{p} + \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} s_{ij}, \tag{3.28}$$

where $z_1 \geq 0$ and $z_2 \geq 0$ are the newly introduced variables. Similarly, by introducing $J_{ij}^{\text{loc}} \geq 0$ and $J_{ij}^{\text{ext}} \geq 0$ as the local and external interference terms at the jth SUE in the ith small cell, we

can replace (3.24e) and (3.24f) by

$$s_{ij}+\log\left(J_{ij}^{\text{loc}}+J_{ij}^{\text{ext}}+N_0\right)\geq G_{ij}^{SI}\left(p_{ij},J_{ij}^{\text{loc}},J_{ij}^{\text{ext}},\mathbf{p}^{(n)}\right)$$

and

$$\log\left(p_{ij}\left|h_{iij}\right|^2+J_{ij}^{\text{loc}}+J_{ij}^{\text{ext}}+N_0\right)\geq G_{ij}^{I}\left(J_{ij}^{\text{loc}},J_{ij}^{\text{ext}},\mathbf{p}^{(n)}\right)+t_{ij},$$

respectively, with additional constraints

$$J_{ij}^{\text{loc}} \geq \sum_{k\in\mathscr{S}_u\setminus j} p_{ik}\left|h_{iij}\right|^2, \tag{3.29}$$

$$J_{ij}^{\text{ext}} \geq \sum_{k\in\mathscr{S}_a\setminus i}\sum_{l\in\mathscr{S}_u} p_{kl}\left|h_{kij}\right|^2, \tag{3.30}$$

Then, let us introduce the Lagrange multipliers $\lambda=[\lambda_1\lambda_2]^T,\mu=[\mu_1,\ldots,\mu_{S_a}]^T$, and $\xi=[\xi_{11},\ldots,\xi_{S_aS_u}]^T$, and proceed to write the dual problem as

$$\min_{\lambda\geq 0,\mu\geq 0,\xi\geq 0} g\left(\lambda,\mu,\xi\right). \tag{3.31}$$

where $g\left(\lambda,\mu,\xi\right)$ is determined by the following optimization problem

$$\max_{\Pi^c}\ \mathscr{L}\left(\mathbf{p},q,\mathbf{s},\mathbf{t},\mathbf{z},\mathbf{J},\lambda,\mu,\xi\right) \tag{3.32a}$$

$$\text{s.t. } \frac{\tau+z_1}{2}\geq\sqrt{q^2+\frac{\left(\tau-z_1\right)^2}{4}}, \tag{3.32b}$$

$$F\left(1,\tau,\tau^{(n)}\right)\geq\kappa_m\sum_{i\in\mathscr{F}}\left\|\mathbf{v}_i\right\|^2+P_m^{\text{circ}}+z_2, \tag{3.32c}$$

$$s_{ij}+\log\left(J_{ij}^{\text{loc}}+J_{ij}^{\text{ext}}+N_0\right)\geq G_{ij}^{SI}\left(p_{ij},J_{ij}^{\text{loc}},J_{ij}^{\text{ext}},\mathbf{p}^{(n)}\right),\forall i\in\mathscr{S}_a,\forall j\in\mathscr{S}_u, \tag{3.32d}$$

$$\log\left(p_{ij}\left|h_{iij}\right|^2+J_{ij}^{\text{loc}}+J_{ij}^{\text{ext}}+N_0\right)-t_{ij}\geq G_{ij}^{I}\left(J_{ij}^{\text{loc}},J_{ij}^{\text{ext}},\mathbf{p}^{(n)}\right),\forall i\in\mathscr{S}_a,\forall j\in\mathscr{S}_u, \tag{3.32e}$$

$$J_{ij}^{\text{loc}}\geq\sum_{k\in\mathscr{S}_u\setminus j}p_{ik}\left|h_{iij}\right|^2,\forall i\in\mathscr{S}_a,\forall j\in\mathscr{S}_u, \tag{3.32f}$$

$$(3.24\text{d}),(3.16\text{e}),(3.16\text{h}),(3.16\text{i}),(3.16\text{j}), \tag{3.32g}$$

where $\Pi^c = \{\Pi, \mathbf{z} \geq 0, \mathbf{J} \geq 0\}$ and the objective function $\mathscr{L}(\mathbf{p}, q, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J}, \lambda, \mu, \xi)$ is the *partial augmented* Lagrangian function of (3.24), which can be written as

$$\mathscr{L}(\mathbf{p}, q, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J}, \lambda, \mu, \xi) = q + \lambda_1 \left(\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^s - z_1\right) + \lambda_2 \left(z_2 - \kappa_s \mathbf{1}^T \mathbf{p} - \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} s_{ij}\right)$$
$$+ \sum_{i \in \mathscr{S}_a} \mu_i \left(t_i^m - \sum_{j \in \mathscr{S}_u} s_{ij}\right) + \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} \xi_{ij} \left(J_{ij}^{\text{ext}} - \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} p_{kl} \left|h_{kij}\right|^2\right) + \frac{\rho}{2} \mathscr{P}(\mathbf{p}, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J}), \tag{3.33}$$

where

$$\mathscr{P}(\mathbf{p}, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J}) = -\left(\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^s - z_1\right)^2$$
$$\underbrace{- \sum_{i \in \mathscr{S}_a} \left(t_i^m - \sum_{j \in \mathscr{S}_u} s_{ij}\right)^2 - \left(z_2 - \kappa_s \mathbf{1}^T \mathbf{p} - \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} s_{ij}\right)^2}_{\mathscr{P}_1(\mathbf{t}^m, \mathbf{s}, z_2, \mathbf{p})}$$
$$- \sum_{i \in \mathscr{S}_a} \sum_{j \in \mathscr{S}_u} \left(J_{ij}^{\text{ext}} - \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} p_{kl} \left|h_{kij}\right|^2\right)^2, \tag{3.34}$$

is the penalizing term, which is added to support the convergence speed of the distributed algorithm. Without (3.34), when λ, μ, ξ are fixed, (3.32) can be decomposed into the one macrocell and multiple small cell sub-problems by the dual decomposition method (Ngo *et al.*, 2014a). However, the appearance of $\mathscr{P}(\mathbf{p}, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J})$ in (3.33) has coupled all the variables in the objective function in (3.32a), so that the methods proposed in (Ngo *et al.*, 2014a) cannot be directly applied in here. At this point, we observe that when λ, μ, ξ are fixed, (3.32) is still convex. Thus, if we apply Alternating Optimization (AO) to iteratively solve each block of variables, we can actually decompose the problem into smaller sub-problems. In particular, by considering $\Pi^m = \{q \geq 0, \mathbf{t}^m \geq 0, \tau \geq 0, \mathbf{v}, \nu \geq 0, \mathbf{z} \geq 0\}$ as variables and fixing the values of the other variables from the previous iteration as parameters, the macrocell sub-problem can

be casted as

$$\max_{\Pi^m} \; q + \lambda_1 \left(\mathbf{1}^T \mathbf{t}^{ma} - z_1\right) + \lambda_2 z_2 + \sum_{i \in \mathscr{S}_a} \mu_i t_i^m - \left(\mathbf{1}^T \mathbf{t}^{ma} - \mathbf{1}^T \mathbf{t}^{s(t)} - z_1\right)^2 + \mathscr{P}_1 \left(\mathbf{t}^m, \mathbf{s}^{(t)}, z_2, \mathbf{p}^{(t)}\right) \tag{3.35a}$$

$$\text{s.t. } \frac{\tau + z_1}{2} \geq \sqrt{q^2 + \frac{\left(\tau - z_1\right)^2}{4}}, \tag{3.35b}$$

$$F\left(1, \tau, \tau^{(n)}\right) \geq \kappa_m \sum_{i \in \mathscr{F}} \|\mathbf{v}_i\|^2 + P_m^{\text{circ}} + z_2, \tag{3.35c}$$

$$(3.24\text{d}), (3.16\text{e}), (3.16\text{h}), (3.9\text{e}), \tag{3.35d}$$

and by considering $\Pi_i^s = \{\mathbf{p}_i \geq 0, \mathbf{s}_i \geq 0, \mathbf{t}_i^s \geq 0, \mathbf{J}_i \geq 0\}$ as variables and fixing the values of other variables from the previous iteration as parameter, the ith small cell sub-problem can be written as

$$\begin{aligned} \max_{\Pi_i^s} \; & \lambda_1 \mathbf{1}^T \mathbf{t}_i^s - \lambda_2 \kappa_s \mathbf{1}^T \mathbf{p}_i - \mu_i \sum_{j \in \mathscr{S}_u} s_{ij} + \left(\sum_{j \in \mathscr{S}_u} \xi_{ij} J_{ij}^{\text{ext}} - \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} \xi_{kl} p_{il} \left|h_{ikj}\right|^2 \right) \\ & + \mathscr{P}\left(\mathbf{p}_i, \mathbf{p}_{\sim i}^{(t)}, \mathbf{s}_i, \mathbf{s}_{\sim i}^{(t)}, \mathbf{t}_i^s, \mathbf{t}^{m(t)}, \mathbf{t}_{\sim i}^{s(t)}, \mathbf{z}^{(t)}, \mathbf{J}_i, \mathbf{J}_{\sim i}^{(t)}\right) \end{aligned} \tag{3.36a}$$

$$\text{s.t.: } s_{ij} + \log\left(J_{ij}^{\text{loc}} + J_{ij}^{\text{ext}} + N_0\right) \geq G_{ij}^{SI}\left(p_{ij}, J_{ij}^{\text{loc}}, J_{ij}^{\text{ext}}, \mathbf{p}^{(n)}\right), \forall j \in \mathscr{S}_u, \tag{3.36b}$$

$$\log\left(p_{ij}\left|h_{iij}\right|^2 + J_{ij}^{\text{loc}} + J_{ij}^{\text{ext}} + N_0\right) - t_{ij} \geq G_{ij}^{I}\left(J_{ij}^{\text{loc}}, J_{ij}^{\text{ext}}, \mathbf{p}^{(n)}\right), \forall j \in \mathscr{S}_u, \tag{3.36c}$$

$$J_{ij}^{\text{loc}} \geq \sum_{k \in \mathscr{S}_u \setminus j} p_{ik}\left|h_{iij}\right|^2, \forall j \in \mathscr{S}_u, \tag{3.36d}$$

$$(3.16\text{i}), (3.9\text{f}), \tag{3.36e}$$

where $\mathscr{P}\left(\mathbf{p}_i, \mathbf{p}_{\sim i}^{(t)}, \mathbf{s}_i, \mathbf{s}_{\sim i}^{(t)}, \mathbf{t}_i^s, \mathbf{t}_{\sim i}^{s(t)}, \mathbf{z}^{(t)}, \mathbf{J}_i, \mathbf{J}_{\sim i}^{(t)}\right)$ is the function $\mathscr{P}\left(\mathbf{p}, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J}\right)$ in which $\mathbf{p}_{\sim i}, \mathbf{s}_{\sim i}$, $\mathbf{t}_{\sim i}^s, \mathbf{z}, \mathbf{J}_{\sim i}$ are fixed with values $\mathbf{p}_{\sim i}^{(t)}, \mathbf{s}_{\sim i}^{(t)}$, $\mathbf{t}_{\sim i}^{s(t)}, \mathbf{z}^{(t)}, \mathbf{J}_{\sim i}^{(t)}$. Note that when λ, μ, ξ are fixed, the macrocell sub-problem (3.35) and each ith small cell sub-problem (3.36) are convex problems with respect to all the relevant variables. By allowing each SAP and MBS solve its local problem and updating the parameters with the exchanged information computed from optimal solution, we can obtain the joint optimal solution of these sub-problems at convergence. However, the optimal values obtained from solving these sub-problems may not be the global

optimal solution to (3.24) because the fixed set of dual variables λ, μ, ξ is not the optimal dual variables. To obtain the optimal λ, μ, ξ, we apply the gradient method to solve the master dual problem (3.31) by iteratively updating these dual variables so that they will approach to the optimal dual variables that achieve global optimality. The dual variables can be updated following these rules

$$\begin{aligned}
\lambda_1^{(n_3+1)} &= \left[\lambda_1^{(n_3)} - \varepsilon_\lambda \left(\left(\mathbf{1}^T \mathbf{t}^{ma} + \mathbf{1}^T \mathbf{t}^s\right) - z_1\right)\right]^+, \\
\lambda_2^{(n_3+1)} &= \left[\lambda_2^{(n_3)} - \varepsilon_\lambda \left(z_2 - \kappa_s \mathbf{1}^T \mathbf{p} - 1^T \mathbf{s}\right)\right]^+, \\
\mu_i^{(n_3+1)} &= \left[\mu_i^{(n_3)} - \varepsilon_\mu \left(t_i^m - \mathbf{1}^T \mathbf{s}_i\right)\right]^+, \\
\xi_{ij}^{(n_3+1)} &= \left[\xi_{ij}^{(n_3)} - \varepsilon_\xi (J_{ij}^{\text{ext}} - \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} p_{kl} \left|h_{kij}\right|^2)\right]^+.
\end{aligned} \tag{3.37}$$

The decentralized algorithm is summarized in Algorithm 3. Note that in Step 4, each ith SAP can compute the term $\tilde{J}_{ij}^{\text{ext}(n_2)} = \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} p_{kl}^{(n_2)} \left|h_{kij}\right|^2$ without the need to obtain the out-of-cell CSI, e.g., $h_{kij}, \forall k \neq i$ by the following process. Each ith SAP first request its jth SUE to report its received power at the n_2th step, e.g., $\sum_{k \in \mathscr{S}_a} \sum_{l \in \mathscr{S}_u} p_{kl}^{(n_2)} \left|h_{kij}\right|^2$. Then, it computes $\tilde{J}_{ij}^{\text{ext}(n_2)} = \sum_{k \in \mathscr{S}_a} \sum_{l \in \mathscr{S}_u} p_{kl}^{(n_2)} \left|h_{kij}\right|^2 - \sum_{l \in \mathscr{S}_u} p_{il}^{(n_2)} \left|h_{iij}\right|^2$, where $p_{il}^{(n_2)}$ and $\left|h_{iij}\right|^2$ are available at the ith SAP.

Algorithm 2: Iterative Centralized JBPAO Algorithm

1: Initialize starting points of $\mathbf{v}^{(n)}, \nu^{(n)}, \mathbf{p}^{(n)}, \tau^{(n)}$;
2: Set $n := 0$;
3: **repeat**
4: Solve the convex problem in (3.24) to achieve the optimal solution $\mathbf{v}^\star, \mathbf{p}^\star, \nu^\star, \mathbf{s}^\star, \mathbf{t}^\star, \tau^\star, q^\star$;
5: Set $n := n + 1$;
6: Update $\mathbf{v}^{(n)} = \mathbf{v}^\star, \nu^{(n)} = \nu^\star, \mathbf{p}^{(n)} = \mathbf{p}^\star, \tau^{(n)} = \tau^\star$;
7: **until** Convergence of the objective (3.24a).

Remark 1. *Problem* (3.32) *is the result of the equivalent transformation from the convex problem* (3.24) *and the relaxation of constraints* (3.16e)*,* (3.26)*,* (3.28) *and* (3.30) *into the objective*

(3.32a). *With this relaxation, the solution achieved at each step 6–7 of Algorithm 3 does not guarantee the feasibility of constraints* (3.16e)*,* (3.26)*,* (3.28) *and* (3.30) *as if they are not relaxed into the objective function. Because of this infeasibility, solution at each step 6–7 of Algorithm 3 does not guarantee the feasibility of the minimum rate constraints* (3.9c) *as well. When* $\lambda^{(n_3)}, \mu^{(n_3)}, \xi^{(n_3)}$ *converge at step 12 of Algorithm 2, the constraints* (3.16e)*,* (3.26)*,* (3.28) *and* (3.30) *occur at equality and the minimum rate constraint* (3.9c) *is satisfied.*

In (3.33)*, we note that without the penalizing term, the sub-problems in* (3.35) *and* (3.36) *can be optimally solved at one-shot to achieve the optimal solution with given* $\lambda^{(n_3)}, \mu^{(n_3)}, \xi^{(n_3)}$*. We refer to this approach as dual decomposition method. Indeed, by not considering the AO points, we can simply apply Algorithm 3 to achieve similar results as with the ADMM method. However, we will show later in our simulation results that dual decomposition method will converge slower than ADMM.*

Algorithm 3: Iterative Decentralized JBPAO Algorithm

1: Initialize FOTCA points of $\mathbf{v}^{(n_1)}, \nu^{(n_1)}, \mathbf{p}^{(n_1)}, \tau^{(n_1)}$; AO points $\mathbf{p}^{(n_2)}, q^{(n_2)}, \mathbf{s}^{(n_2)}, \mathbf{t}^{(n_2)}, \mathbf{z}^{(n_2)}, \mathbf{J}^{(n_2)}$; dual variables $\lambda^{(n_3)}, \mu^{(n_3)}, \xi^{(n_3)}$;
2: Set $n_1 := 0$; $n_2 = 0$; $n_3 = 0$;
3: **repeat**
4: **repeat**
5: The ith SAP compute $\tilde{J}_{ij}^{\text{ext}(n_2)}, \forall j \in \mathscr{S}_u$, then signal $\tilde{\mathbf{J}}^{\text{ext}(n_2)}$ to the other SAPs and MBS to feed into their own $\mathscr{P}(\mathbf{p}, \mathbf{s}, \mathbf{t}, \mathbf{z}, \mathbf{J})$;
6: The ith SAP solves (3.36), then signal Π_i^s to the other SAPs and MBS;
7: The MBS solves (3.35), then signal Π_i^s to the other SAPs;
8: $n_2 := n_2 + 1$;
9: **until** Convergence of $\mathbf{p}^{(n_2)}, q^{(n_2)}, \mathbf{s}^{(n_2)}, \mathbf{t}^{(n_2)}, \mathbf{z}^{(n_2)}, \mathbf{J}^{(n_2)}$;
10: Update dual variables using (3.37);
11: $n_3 := n_3 + 1$;
12: **until** Convergence of $\lambda^{(n_3)}, \mu^{(n_3)}, \xi^{(n_3)}$;
13: $n_1 := n_1 + 1$;
14: The MBS updates $\mathbf{v}^{(n_1)}, \nu^{(n_1)}, \mathbf{p}^{(n_1)}, \tau^{(n_1)}$ and repeat Step 3–12 until convergence of the objective (3.32a).

3.5 AEE with adaptive decoding power and SAP selection

In this section, we aim at improving the achieved AEE by conserving more power contributing to $\tilde{P}_{\text{tot}}$ with SAP selection strategy. In particular, we can decide to switch OFF the SAP to save more power so that the achieved AEE can result in a better performance. This improvement of AEE can be intuitively explained as follows. First, with fewer SAPs concurrently receiving and transmitting, SUEs and MUEs suffer from less interference so that the numerator of AEE increases. Second, fewer SAPs switched ON means that less total power is consumed, which in turn reduces the denominator of AEE. Being motivated by this intuition and taking advantage of the developed framework in Section 3.3.3, we introduce binary variables $b_i \in \{0,1\}, \forall i \in \mathscr{S}_a$ as the selection variables for the SAP, where $b_i = 1$ means that the ith SAP is active to receive backhaul data and transmit to its SUEs and $b_i = 0$ means that the ith SAP is switched OFF. The new formula of $P_{s,i}^{\text{circ}}, \forall i \in \mathscr{S}_a$ can be presented as

$$\hat{P}_{s,i}^{\text{circ}} = b_i \left(P_s^{\text{active}} + P_s^{\text{diss}} \right) + P_{s,i}^{\text{dec}}. \tag{3.38}$$

Then, we must ensure the relationship that when $b_i = 0$, $\mathbf{v}_i = 0$ and $p_i = 0$. This can be enforced since when the ith SAP is turned OFF, there should be no beamforming transmitted towards it and it also should not transmit any signal to its SUEs. This condition can be easily met by introducing the following two convex constraints: $\|\mathbf{v}_i\|^2 \leq b_i P_{\max}$ and $p_{ij} \leq b_i p_{\max}$. Thus, the AEE optimization problem with SAP selection can be formulated as

$$\max_{\mathbf{v},\mathbf{p}\geq 0,\mathbf{b}} \quad \frac{\sum_{i\in\mathscr{M}} R_i + \sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} r_{ij}}{\hat{P}_{\text{tot}}} \tag{3.39a}$$

$$\text{s.t.} \quad \|\mathbf{v}_i\|^2 \leq b_i P_{\max}, \forall i \in \mathscr{S}_a, \tag{3.39b}$$

$$p_{ij} \leq b_i p_{\max}, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.39c}$$

$$r_{ij} \geq b_i r_{\min}^s, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, \tag{3.39d}$$

$$b_i \in \{0,1\}, \forall i \in \mathscr{S}_a, \tag{3.39e}$$

$$(3.9\text{b}) - (3.9\text{f}). \tag{3.39f}$$

where

$$\hat{P}_{\text{tot}} = \kappa_m \sum_{i\in\mathscr{F}} \|\mathbf{v}_i\|^2 + P_m^{\text{circ}} + \kappa_s \sum_{i\in\mathscr{S}_a} \sum_{j\in\mathscr{S}_u} p_{ij} + \sum_{i\in\mathscr{S}_a} \hat{P}_{s,i}^{\text{circ}}.$$

For this non-convex combinatorial problem, we follow the same steps in Section 3.3.3 to arrive at the combinatorial approximated problem

$$\max_{\Pi^b} \{q|(3.39\text{b})-(3.39\text{e}),(3.24\text{b})-(3.24\text{g})\}, \tag{3.40}$$

where $\Pi^b = \{\Pi, \mathbf{b}\}$. Algorithm 2 can be applied to iteratively solve (3.40) and update relevant parameters until convergence. Note that without constraints (3.16d), (3.24e), and (3.24f), (3.40) is indeed a MISOCP problem. This is because these convex constraints contain the form $\log(x) \geq y$, which can be easily rewritten into $x \geq e^y$ and reveals (3.40) as a general mixed integer nonlinear convex programming. Towards this end, we are motivated to approximate (3.16d), (3.24e), and (3.24f) with corresponding SOC constraints so that (3.40) becomes a MISOCP and can be solved optimally and much more efficiently by a dedicated MISOCP solver like MOSEK. It is important to state that the accuracy of this approximation can be achieved within a controlled error since we aim to approximate $x \geq e^y$ by (Ben-Tal & Nemirovski, 2001a)

$$\begin{aligned}
\left\| \left(2^{-m}y + 5/3\right)/2, \sqrt{19/72} \right\| &\leq \kappa_1, \\
\left\| 2\left(1+2^{-m}y\right), (\kappa_2 - 1) \right\| &\leq \kappa_2 + 1, \\
\left\| \kappa_1, \kappa_2/\sqrt{24} \right\| &\leq \kappa_3, \\
\|2\kappa_{2+i}, (\kappa_{3+i} - 1)\| &\leq \kappa_{3+i} + 1, \, i = 1, \ldots, m, \\
\kappa_{3+m} + 1 &\leq x,
\end{aligned} \tag{3.41}$$

where $\kappa_i, \forall i = 1, \ldots, 3+m$ are the newly introduced variables and m is the given parameter that controls the accuracy of the approximation. This implies that for a given sufficiently large m, the system of constraint (3.41) can approximate the hypograph of $\left\{(x,y) \in \mathbb{R}^2 | x \geq e^y\right\}$ to an arbitrary accuracy ε. From (Ben-Tal & Nemirovski, 2001a), when $x \in [0, \bar{x}]$ and $x \geq e^y$, there exists $\{x, y, \kappa_1, \ldots, \kappa_{3+m}\}$ that satisfies (3.41). In addition, if $x \in [0, \bar{x}]$ and the solution

set $\left\{(x,y) \in \mathbb{R}^2 | x \geq e^y\right\}$ is able to be extended by $\{x, y, \kappa_1, \ldots, \kappa_{3+m}\}$ for some m that satisfies (3.41), then $e^y - \varepsilon \leq x \leq e^y + \varepsilon$.

Complexity analysis: We discuss the computational complexity of each proposed algorithm. We note that the convex approximated problem (2.21) is a general nonlinear convex programming because it contains the exponential cone constraints (3.16d), (3.24e), and (3.24f). By following the result in (Ben-Tal & Nemirovski, 2001b) and realizing that there are $S_a(3S_u + N + 2) + M(N+2) + 2$ continuous variables in (3.24), the worst-case per-iteration complexity analysis of Algorithm 1 for (3.24) after omitting the small order is given by $\mathscr{O}\left(S_a^4 S_u^4 + S_a^4 N^4 + M^4 N^4\right)$. Similarly, we derive the complexity analysis for Algorithm 2 for problem (3.40). With SOC approximation (3.41), (3.40) is a MI-SOCP, which is a combinatorial optimization problem. There are S_a binary variables b_i in (3.40), which results in 2^{S_a} combinations of all binary variables. Given the fixed value of $b_i, \forall i \in \mathscr{S}_a$, (3.40) is in the form of SOCP. Similar to the analysis in (Nguyen *et al.*, 2016b), we note that there are approximately $S_a(3S_u + N + 2) + M(N+2) + 2 + (S_a + M + 2S_aS_u)(m+4)$ continuous variables in (3.40) with 2 constraints of dimension $N(S_a + M)$, $2S_a + 2M + S_aS_u + 2$ constraints of dimension $S_a + M$ and $(S_a + M + 2S_aS_u)(m+4)$ SOC constraints of dimension 3. Thus, the worst-case per-iteration complexity of Algorithm 2 on problem (3.40) is given by $\mathscr{O}\left(2^{S_a}\left(S_a^4 S_u^3 (M^3 + N^3 + m^3)\right) + M^4\left(N^3 + m^3\right)\right)$.

Alternatively, we note that Algorithm 3 decouples the joint problem (3.24) into smaller subproblems at the MBS and each SAP. Thus, its complexity should rely on the individual complexity that solves the macrocell subproblem (3.35) and the small cell subproblem (3.36). By applying similar analysis and omitting small order, the worst-case per-iteration complexity analysis for solving (3.35), which contains $4 + (S_a + M)(N+2)$ continuous variables is $\mathscr{O}\left(\left(S_a^4 + M^4\right)N^4\right)$ and for solving (3.36), which contains $5S_u$ continuous variables is $\mathscr{O}\left(S_u^4\right)$.

Signaling information analysis (Shi *et al.*, 2014): For centralized approach in Algorithm 2, we assume that all computations are executed at the MBS. The MBS has the knowledge of $\mathbf{h}_i, \forall i \in \mathscr{F}$. It receives S_aS_u channel coefficients from each SAP, and broadcasts S_aS_u coefficients that represents the value of variables $\mathbf{p}^\star$ to all the SAPs. For the decentralized approach in

Algorithm 3, under the assumption that the MBS and each SAP only know their local CSI, the message passing in Algorithm 3 at each transmitter in the considered network can be listed as follows:

- MBS has the knowledge of $\mathbf{h}_i, \forall i \in \mathscr{F}$, broadcasts $2+(2S_a+M)$ coefficients to all the SAPs and receives $5S_u$ coefficients from each SAP.
- Each ith SAP has the knowledge of $h_{ijk}, \forall j \in \mathscr{S}_a, k \in \mathscr{S}_u$, broadcasts $5S_u$ coefficients to all neighbor SAPs and MBS and receives $2+(2S_a+M)$ coefficients from the MBS and $5S_u$ coefficients from other SAP.

Note that the decentralized based on dual decomposition method also has the same local knowledge of CSI and signaling information as the ADMM method.

3.6 Numerical Results

In this section, we evaluate the performance of the algorithms proposed in the previous sections. We assume time-invariant and flat Rayleigh fading channels and the pathloss component is calculated as $\left(d_{ij}/d_0\right)^{-3.8}$, where d_{ij} is the distance between the ith transmitter and the jth receiver, and $d_0 = 100$ m is the reference distance. We apply a circular coverage of macrocell with radius $10d_0$. The MBS is positioned at the center, where there are $S = 4$ small cells uniformly positioned in the macrocell coverage. We assume that each small cell has a circular coverage of radius d_0 with its SAP at the center and one SUE at the circumference of each small cell coverage. In addition, we assume that there are $M = 2$ MUEs uniformly scattered across the macrocell coverage. Unless being mentioned elsewhere, we choose this scenario as the standard model for the numerical simulation, where the number of transmit antenna as $N = 4$ and the maximum transmit power is $P_{\max} = 50$ dBm, while at the small cell, we choose the maximum transmit power at the SAP $p_{\max} = 30$ dBm. To protect the QoS of each MUE and SUE, we choose the minimum rate requirement $r^m_{\min} = r^s_{\min} = 0.01$ bits/s/Hz. The noise power is set $N_0 = -120$ dB, and the bandwidth is $W = 10$ MHz. Furthermore, we choose $P^{\text{circ}}_m = 30$ dBm, $P^{\text{ad}}_{s,i} = 30$ dBm, $\kappa_m = \kappa_s = 5$ and $\alpha_i = 1, \forall i \in \mathscr{S}_a$ (Ng *et al.*, 2012; Tervo *et al.*, 2015b). Throughout the numerical section, we compare the network performance between the

following schemes, Scheme A and B: maximization of AEE using adaptive decoding power model as in (3.6) and fixed decoding power, respectively; Scheme C: maximization of AEE considering adaptive decoding power model as in (3.6) and SAP selection as in Section 3.5; Scheme D: maximization of access sum rate (ASR) whose objective function is only the numerator of the objective function in (3.10). The solution of the ASR maximization problem is achieved through Algorithm 2 and then used to compute the achieved AEE by plugging it into the formula (3.7). It is worth to mention that Scheme B with fixed decoding power assignment and no SAP selection is similar to the traditional works in the literature (He *et al.*, 2014; Xu *et al.*, 2014; Ng *et al.*, 2012; Tervo *et al.*, 2015b).

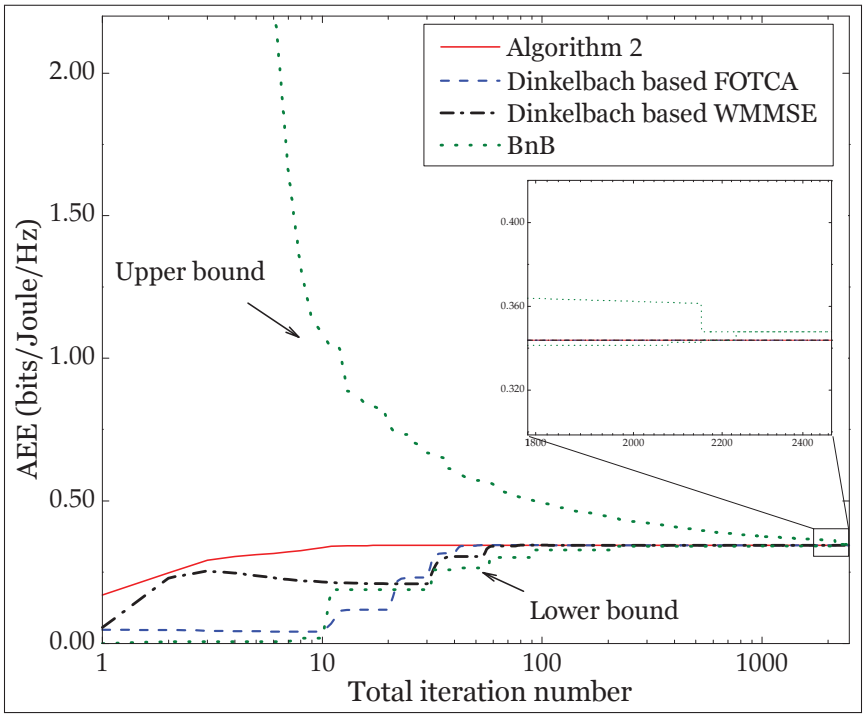

Figure 3.2 Convergence of proposed BnB, Algorithm 2, and other algorithms at $S_a = 2$.

In Fig. 3.2–3.4, we show the convergence of the objective function in (3.10) by applying the BnB algorithm mentioned in Section 3.3.2 and Algorithm 2 (Centralized JBPAO) with $S_a = 2, 3, 4$ amd $P_m^{\text{circ}} = 10$ dBm, $P_{s,i}^{\text{ad}} = 10$ dBm. In addition, we also compare these algorithms with the traditional approaches using the Dinkelbach method (Ng *et al.*, 2012) combined with

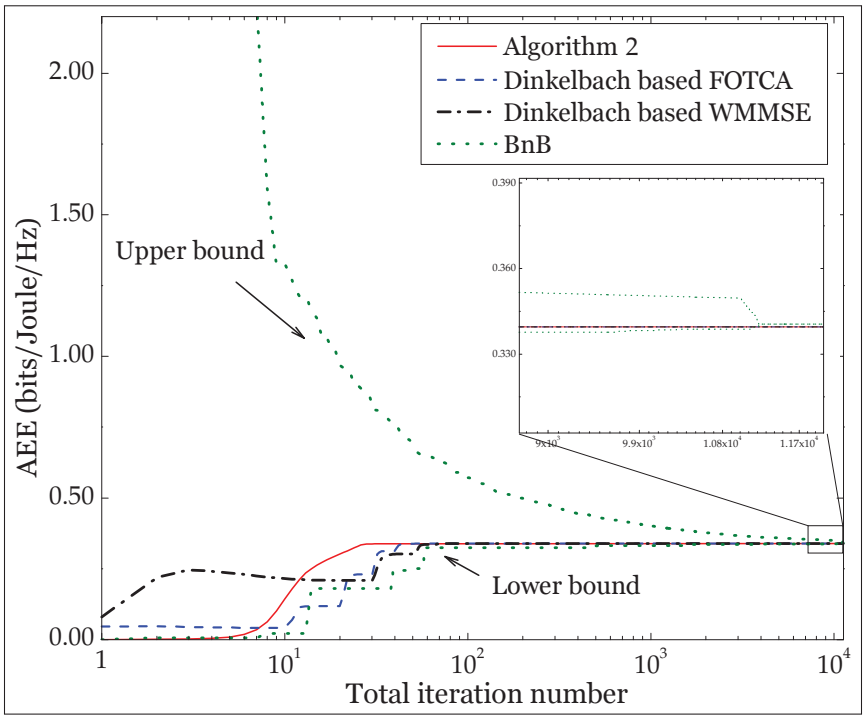

Figure 3.3 Convergence of proposed BnB, Algorithm 2, and other algorithms at $S_a = 3$.

the FOTCA technique and with the weighted minimum mean square error (WMMSE) (He *et al.*, 2014) to solve the fractional nonlinear problem. Note that with the Dinkelbach method, the fractional objective function in (3.10) is transformed into a subtractive form with additional fixed parameter $\theta^{(t)}$ (Ng *et al.*, 2012). The process of solving the new problem in subtractive form can be briefly described in two folds: first, we fix $\theta^{(t)}$ and apply the convex approximation in Section 3.3.3 or WMMSE in (He *et al.*, 2014) to iteratively solve the approximated problem and update corresponding parameters; second, we update $\theta^{(t)}$ and repeat the first step. In Fig. 3.2–3.4, we observe that the result of the objective function from Dinkelbach-based FOTCA and Dinkelbach-based WMMSE approaches and Algorithm 2 converge after few iterations. However, both the Dinkelbach-based approaches require more iterations to converge to the same solution as with Algorithm 2, starting with the same initial point. The reason is that in our proposed algorithm, we aim to directly tackle the fractional objective function without introducing new parameters, which saves more time to achieve the converged result. On the

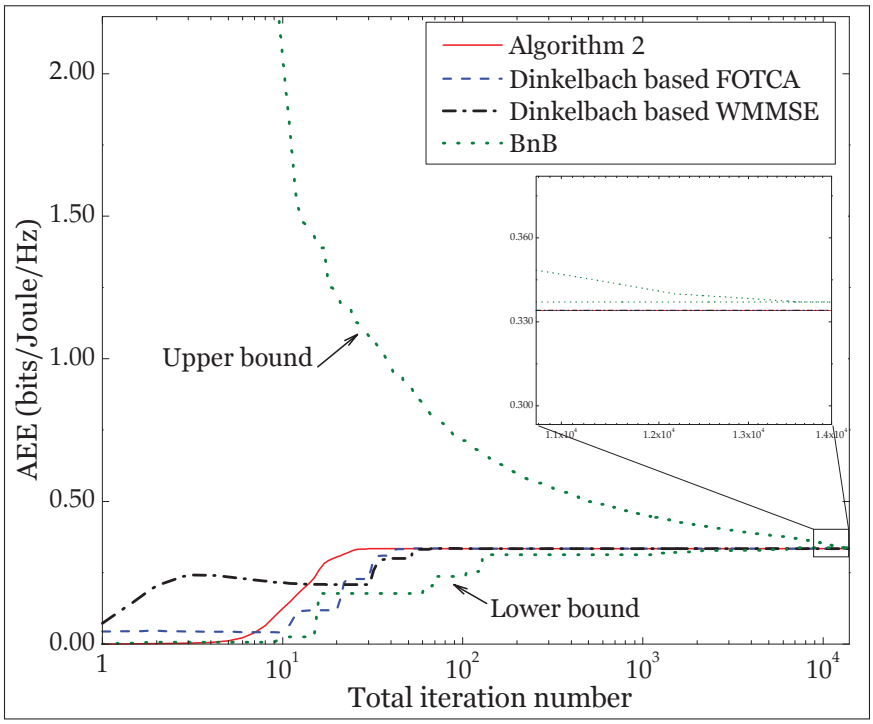

Figure 3.4 Convergence of proposed BnB, Algorithm 2, and other algorithms at $S_a = 4$.

other hand, the BnB algorithm requires many more iterations to update the upper and lower bounds of the objective function and finally converges after several iterations, where at larger S_a, the iteration number for convergence increases rapidly. We observe that the converged values of Algorithm 2 are very close to the baselines created by the converged values of the BnB algorithm at $S_a = 2, 3, 4$, which again shows the efficiency of our proposed method to achieve a close to optimal solution with much less computational complexity.

In Fig. 3.5, we show the convergence of Algorithm 2 (Centralized JBPAO) and Algorithm 3 (Decentralized JBPAO). Furthermore, we also compare the convergence speed of Algorithm 3 based on two different approaches: ADMM and Dual Decomposition (Dual Dec.). Here, the performance is computed with respect to the total number of iterations necessary for the algorithm to eventually converge. We observe that the decentralized Algorithm 3 based ADMM and Dual Dec. finally converge to the same value at the convergence of the centralized Algorithm 2 by iteratively solving the dual problem, applying the dual update and updating the

initial FOTCA point. Compared to the ADMM algorithm with roughly 70 iterations for convergence at the same initial points, the Dual Dec. algorithm requires much more iterations to search for the optimal dual variables through gradient method with respect to fixed FOTCA points.

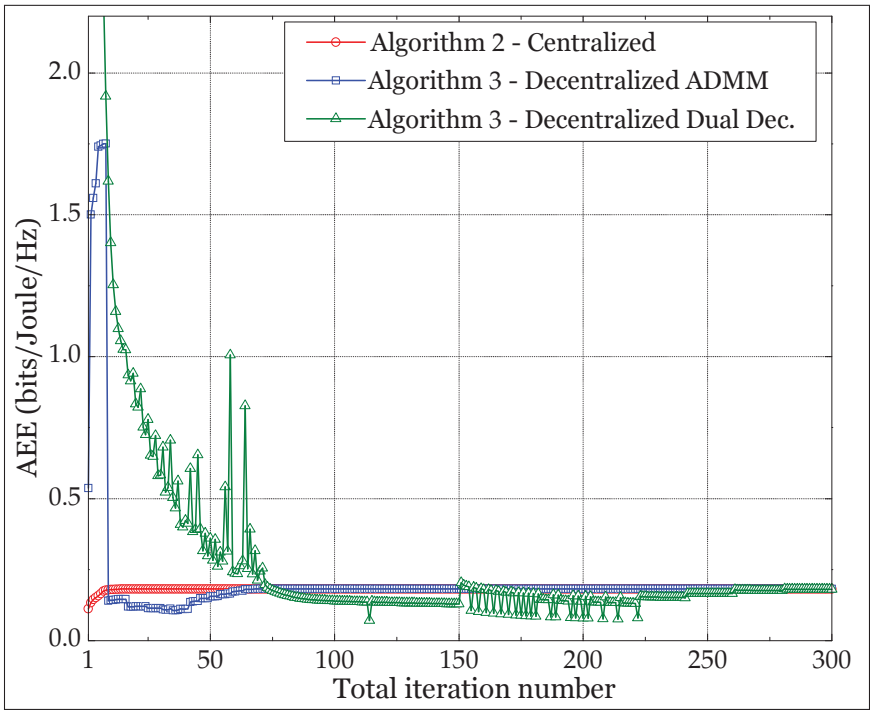

Figure 3.5 Convergence comparison of centralized and decentralized JBPAO algorithms.

In Fig. 3.6, we show the achieved AEE with respect to $P_{\max}$ by applying Algorithm 2 for Schemes A, B, C, and D. In addition, we also compare the performance of Algorithm 2 with the the signal-to-leakage-plus-noise ratio (SLNR) scheme as used in (Venturino & Buzzi, 2015). In SLNR scheme, transmit beamforming is chosen by $\mathbf{v}_i = c_i \left(\sum_{j \neq i} \mathbf{h}_j \mathbf{h}_j^H + N_0 \mathbf{I}_M \right)^{-1} \mathbf{h}_i, \forall i \in \mathscr{F}$, where c_i is the normalized factor and we optimize the joint power allocation at the MBS and SAPs. We choose $r_{\min}^m = r_{\min}^s = 0$. From the figure, we observe that when $P_{\max}$ increases, AEE values of Schemes A, B, and C and SLNR increase and then saturate. In fact, a low power is required to maximize the AEE; therefore, increasing MBS power does not help in improving the AEE. The intuition of this remark is that when the BS tries to allocate more

power with more power budget, their achievable rates are also limited due to more interference from other simultaneous transmissions. However, this limit does not suffice to compensate the cost of using more $\tilde{P}_{\text{tot}}$ and results in the decrease of AEE. Another observation is that the achieved AEE from Scheme C outperforms other scenarios. This is because when there are some small cells achieving low rate while they still need to maintain power $P_{s,i}^{\text{circ}}$ to stay on, it is better to switch off these SAPs to save circuit power so that less total power is consumed to achieve high AEE. The solution from Scheme A provides better AEE than Scheme B. This can be explained as when the SAP is unaware of the amount of power needed for the decoding process, it should apply redundant power to make sure sufficient information will be decoded. Hence, this results in power usage inefficiency and degrades the achieved AEE. In addition, compared to the scenario with fixed SLNR beamforming, Scheme A also outperforms SLNR scheme since it is more beneficial to optimally use antenna array to jointly design transmit beamforming rather than predetermining the beamformed direction and limit the design with only power allocation optimization. We observe that the achieved AEE of Scheme D shows a very low performance compared to others. This is because the MBS and SAPs attempt to use as much power as possible to attain the maximum ASR; thus when $P_{\max}$ becomes large, AEE will start to decrease.

In Fig. 3.7, we compare the performance of Algorithm 2 for Scheme A in the limited CSI scenario, where the CSI acquired at the transmitter is erroneous because of some source of errors such as imperfect estimation, quantization or limited feedback. For simplicity, we assume that only estimated CSI is known at the transmitters with the estimation errors lying in some bounded set. For example, the estimated channel model can be given by $\mathbf{h}_i = \hat{\mathbf{h}}_i + \boldsymbol{\varepsilon}_{\mathbf{h}}, \forall i \in \mathscr{F}$ and $h_{ijk} = \hat{h}_{ijk} + \varepsilon_h, \forall i, j \in \mathscr{S}_a, k \in \mathscr{S}_u$, where $\hat{\mathbf{h}}_i, \hat{h}_{ijk}$ are the estimated CSI and $\|\boldsymbol{\varepsilon}_{\mathbf{h}}\|^2 \leq \delta_{\mathbf{h}}$ and $|\varepsilon_h|^2 \leq \delta_h$ represents the estimation error which lies in a bounded set defined by $\delta_{\mathbf{h}}$ and δ_h, respectively. It is important that the solution of the optimization problem in (3.10) with the knowledge of estimated CSI might not lie in the feasible domain of the problem with perfect CSI. This is because this "imperfect" solution may probably lead to the violation of constraints (3.9b)–(3.9c). Thus, a robust optimization design in the case of erroneous CSI with

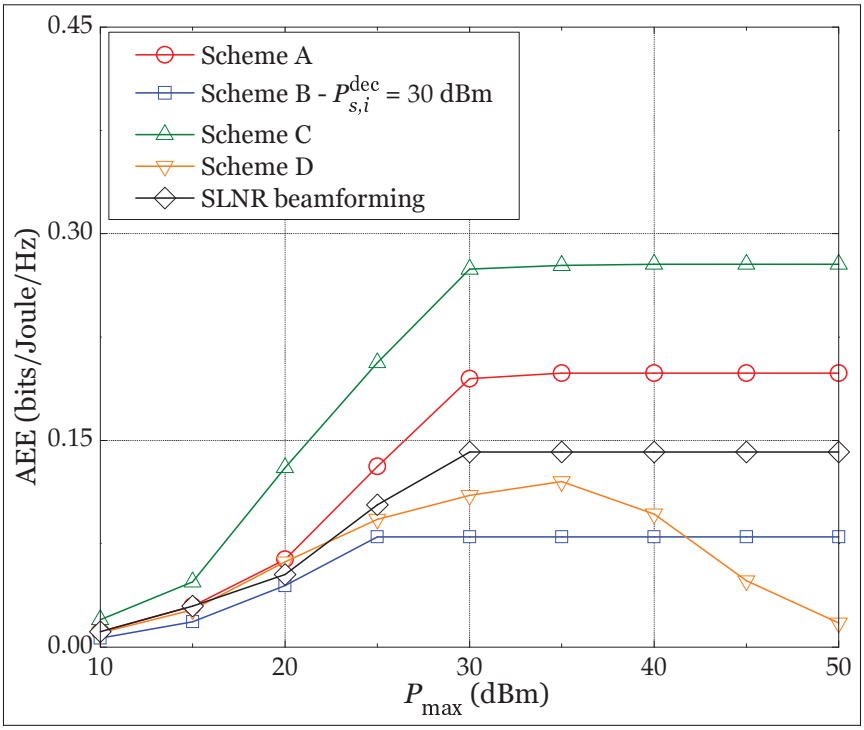

Figure 3.6 Comparison of achieved AEE with respect to $P_{\max}$ at different scenarios.

the worst-case design as in (Nguyen *et al.*, 2016b) must be considered. As in Fig. 3.7, we show the achieved AEE at different values of error bound $\delta_{\mathbf{h}} = \delta_h = \delta = 0, 0.2, 0.3, 0.4$ and $r_{\min}^m = r_{\min}^s = 0.01$, where $\delta = 0$ is equivalent to the perfect CSI case. We observe that the achieved AEE in the imperfect CSI is degraded at lower performance compared to the perfect CSI and the achieved AEE reduces when δ increases. This is because the respective robust optimization problem is solved with the estimated CSI knowledge and the obtained solution is deteriorated with this imperfect CSI.

To further investigate the behavior of overall power consumption, in Fig. 3.8, we evaluate P_{tot} with the solution achieved by applying Algorithm 2 to solve the problems with the same setup as in Fig. 3.6, e.g., Schemes A, B, C, and D. The result is shown at two levels of $P_{\max} = 30, 50$ dBm. By observing Fig. 3.8, we see that Scheme C always consumes less power than Scheme A, B and D. This is due to the fact that in Scheme C, the inefficient SAPs are switched off to save not only transmission power but also the circuit power, decoding power and the transmit

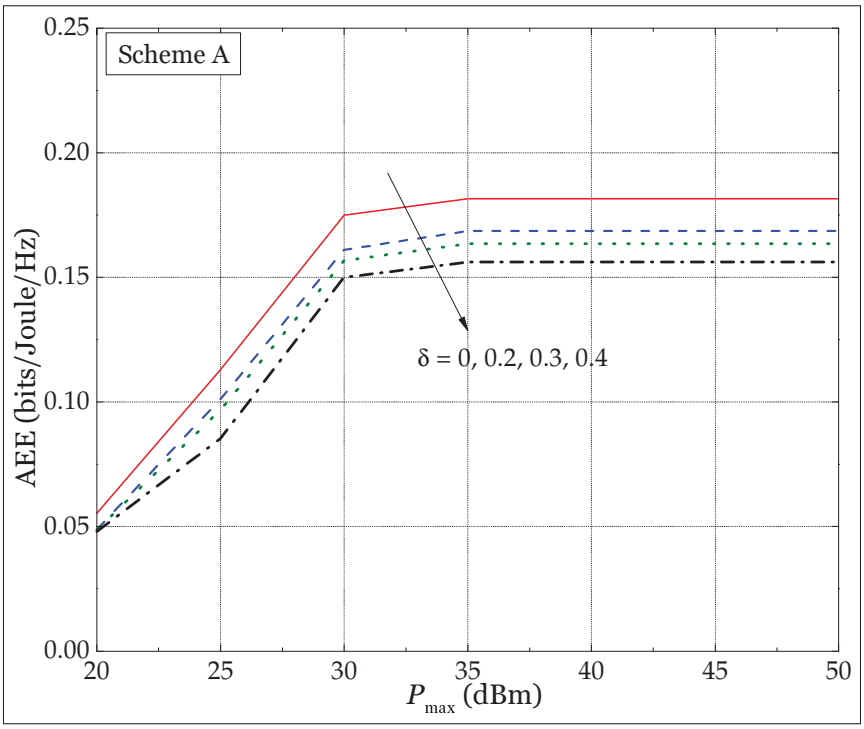

Figure 3.7 Achieved AEE with respect to $P_{\max}$ in limited CSI.

beamforming power from the MBS towards that SAP for WBC. It is worth to note that at high value of $P_{\max} = 50$ dBm, Scheme D uses most of the maximum budget power at the MBS $P_{\max}$ and each SAP $p_{\max}$ to maximize the ASR. This explains the reason that the achieved AEE of Scheme D is very low in Fig. 3.6.

In Fig. 3.9, we show the achieved AEE with respect to $r^s_{\min}$ at two different $P_{\max} = 30, 50$ dBm. The AEE is evaluated for Scheme A, B, and D. Again, we observe that Scheme A outperforms Scheme B and D. This can be explained similarly to Fig. 3.6. From the figure, we see that when $r^s_{\min}$ increases, the achieved AEE in Scheme A and B initially remain constant at low $r^s_{\min}$ and decrease when $r^s_{\min}$ grows sufficiently large. This is because as $r^s_{\min}$ increases, more transmit power from each SAP for WAC transmission and corresponding transmit beamforming power from MBS towards that SAP for WBC are required to maintain the minimum rate requirement constraints. In addition, the decoding power at each SAP also increases, which finally results in the decrease of AEE.

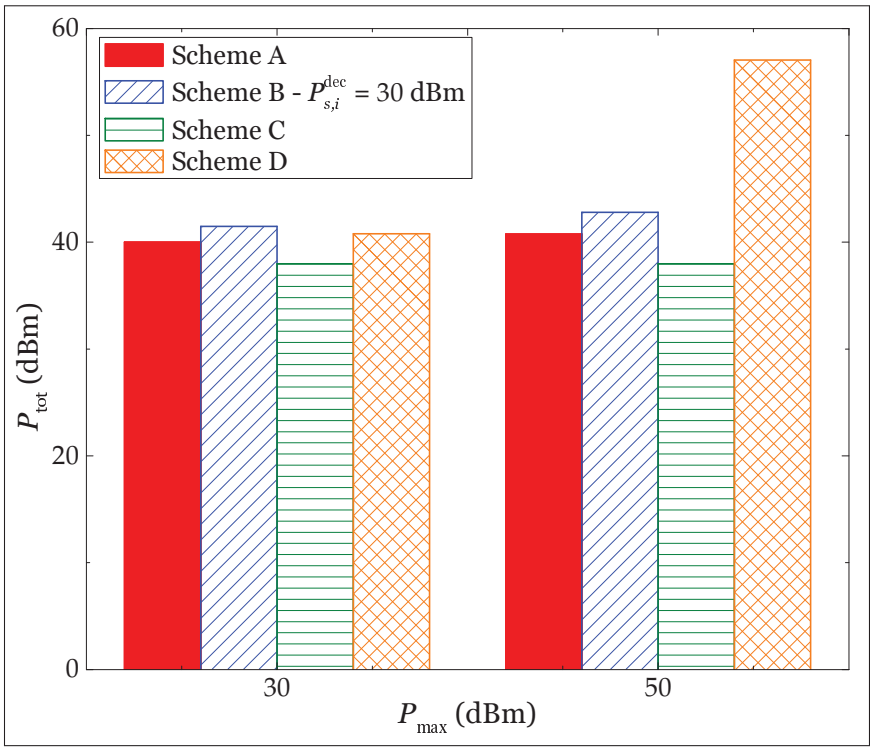

Figure 3.8 Achieved P_{tot} with respect to P_{max}.

In Fig. 3.10, we show the achieved AEE obtained from Algorithm 2 in Scheme A with respect to the number of small cells S_a at different values $P_m^{\text{circ}} = 25, 30, 35, 40$ dBm. We first see that when P_m^{circ} increases, more power at the MBS is required to maintain the operation of the WAC and WBC, which in turn reduces the achieved AEE. Moreover, when the number of small cell increases, all the achieved AEE decreases accordingly. This is obvious because when there are more small cells serving their SUEs, the total rate achieved at the small cells also increases accordingly. However, this increase is not sufficient to compensate the cost of consuming more total power. Specifically, the MBS needs to consume more power to beamform the backhaul data to the small cells. In addition, more transmit and circuit power at each SAP also contributes to the increment of P_{tot}, which increases the overall power consumption. Therefore, the AEE reduces when the number of small cell becomes higher.

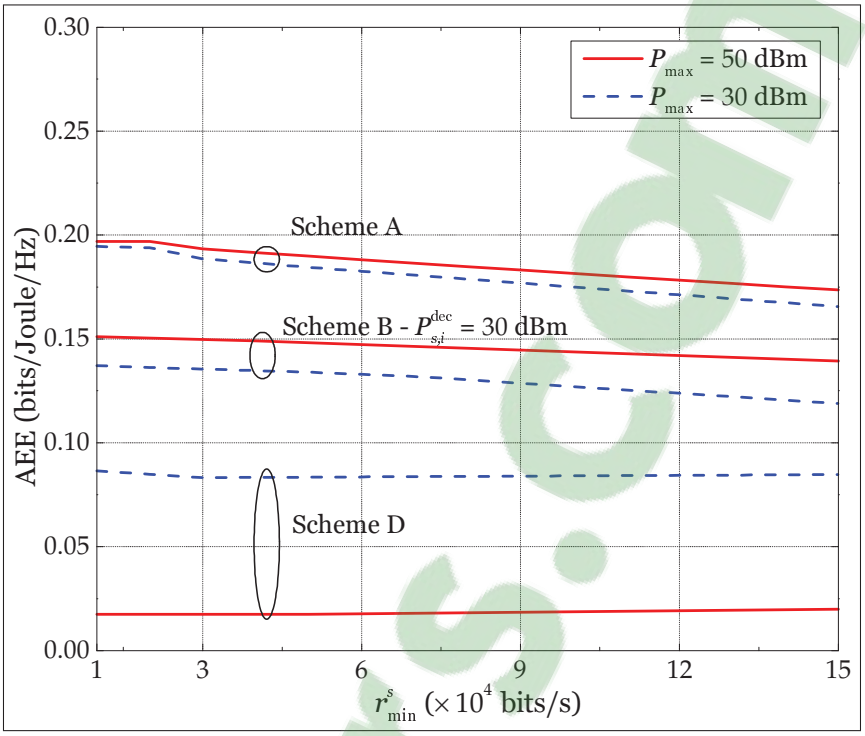

Figure 3.9 Achieved AEE with respect to $r^s_{\min}$.

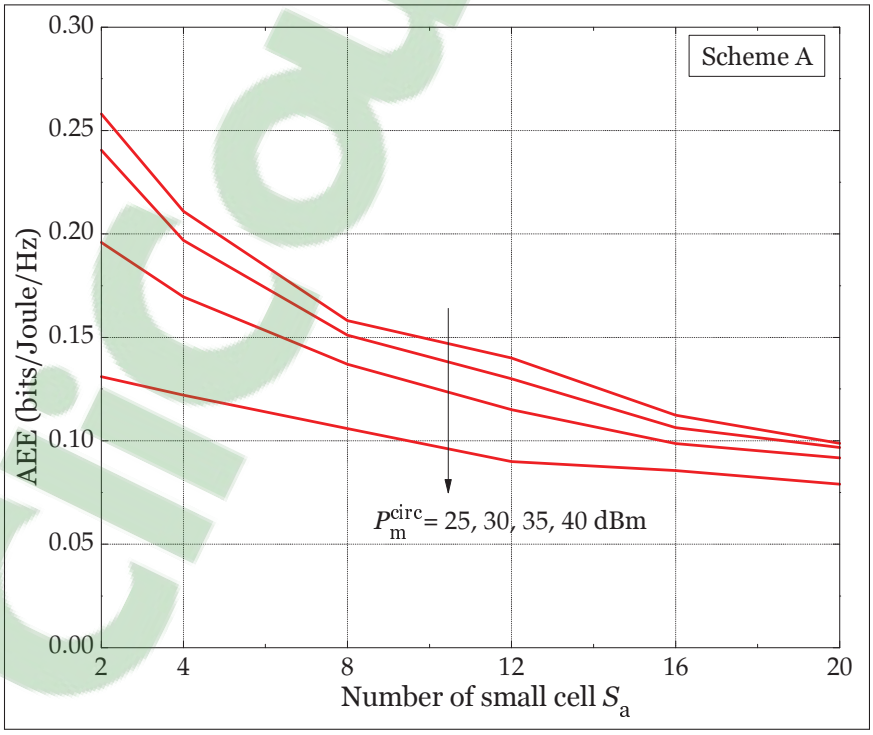

Figure 3.10 Achieved AEE with respect to S_a.

3.7 Concluding Remarks

This paper studied the joint design of transmit beamforming and power allocation that maximizes the AEE on the DL of the two-tier WB small cell HetNets. An important impact of decoding power at each SAP that is proportional to the achievable backhaul rate at each small cell is considered. In the first centralized algorithm, a global optimal solution is achieved by a very high complexity algorithm based on BnB approach. In the second centralized algorithm, a much lower complexity algorithm based on FOTCA is proposed, in which an approximated problem is iteratively solved until convergence. Through numerical results, we see that the solution achieved from the low complexity algorithm performs very close to the global optimal solution, which proves the efficiency of our proposed algorithm. Compared to traditional designs criteria using fixed decoding power, our proposed EE model achieves a better AEE since more power usage is conserved under the proposed adaptive decoding power. These result are far more improved in the extended AEE problem that jointly considers SAP selection. To approach a more practical algorithmic implementation, we develop a distributed algorithm by conducting the framework of ADMM on the convex approximated problem, where MBS and each SAP now only have to solve their own problems with local CSI and limited exchange information. Results have shown that our proposed distributed method significantly improves the convergence speed compared to previous works based on dual decomposition method.

CHAPTER 4

DESIGNING WIRELESS BACKHAUL HETEROGENEOUS NETWORKS WITH SMALL CELL BUFFERING

Tri Minh Nguyen[1], Wessam Ajib[2], Chadi Assi[3]

[1] Département de Génie Électrique, École de Technologie Supérieure,
[2] Département d'Informatique, Université de Québec à Montréal
[3] Concordia Institute for Information Systems Engineering, Concordia University

This article was published in *IEEE Transaction on Communications* in October 2018 (Nguyen *et al.*, 2018a).

4.1 Introduction

Network densification has recently become one essential catalyst for the evolution of existing heterogeneous networks (HetNets) towards the emerging 5G networks (Andrews *et al.*, 2014; Kela *et al.*). By densely deploying more short-range, low-power small cell access points (SAPs) in multi-tier HetNets (Ghosh *et al.*, 2012; Chandrasekhar *et al.*, 2009; Nguyen *et al.*, 2013), operators are able to deliver seamless connectivity to users and significantly enhance the area spectral efficiency. However, this also means that a denser and more expensive backhaul connection establishment with the core network is required.

Recently, wireless backhaul (WB) is presented as a viable solution to overcome the installation difficulty and expensive cost of wired backhaul architecture in legacy small cell networks (Siddique *et al.*, 2015a; Ge *et al.*, 2014; Hur *et al.*, 2013a). WB enables the SAPs to wirelessly transmit and receive backhaul data to/from the macrocell base station (MBS) concurrently with the wireless access (WA) communications. The establishment problem creates two fundamental bottlenecks which limit the overall achievable rate performance. First, deploying WB communications concurrently with WA communications renders the interference more severe. Compared to wired backhaul with high and stable rate, WB rate are prone to the wireless channel variation. Thus, allocating power resources between competing nodes to maintain a

sustainable operation is more challenging. Second, each SAP should relay the received backhaul data from the core network to its users (Siddique *et al.*, 2015a). Therefore, it is essential to impose that the WB rate achieved at a SAP exceeds the WA rate. Indeed, the WB highly demands an appropriate radio resource design that caters to the stringent WB and WA communications constraints in small cell HetNets to achieve its promising superior performance (Zhao *et al.*, 2015; Wang *et al.*, 2016). These challenges are known as the limiting factors of WB networks. Recently, deploying a local cache at each small cell is shown to be effective, c.f. (Golrezaei *et al.*, 2012; Liu & Lau, 2014, 2013), since it can offload wireless backhaul transmission task by prefetching data into the cache prior to the user request. However, small cell cache storage is often size-limited. Thus, the dual roles of limited wireless backhaul and small cell caching are very essential to satisfy the scope of 5G requirement for ubiquity and high data rate. Several work designs the resource allocation under the joint consideration of cache and limited fronthaul-capacity in cloud radio access network (C-RAN) as in (Tao *et al.*, 2016; Chen *et al.*, 2017; Tran *et al.*, 2017), or limited backhaul in small cell networks as in (Liao *et al.*, 2017; Hamidouche *et al.*, 2017). However, jointly designing WBl and cache is complicated regarding the discussed challenges above, so that the existing work of WB HetNets with resource allocation in (Wang *et al.*, 2016; Zhao *et al.*, 2015; Hu *et al.*, 2016) have been separated from the cache management for the sake of tractability.

4.1.1 Related Work

WB communications were first introduced as a standard technology to route backhaul data over-the-air in IEEE 802.16 wireless mesh networks in (Viswanathan & Mukherjee, 2006) where the authors developed a joint routing and scheduling strategy that maximizes the total network throughput. Towards the evolution of 5G, WB technology was proposed as a viable solution in sub–6 GHz spectrum band to overcome the deployment cost challenge of the wired backhaul by allowing the WB and WA communications to operate concurrently. In (Hoydis *et al.*, 2013; Sanguinetti *et al.*, 2015), a reverse time division duplexing (RTDD) scheme is proposed to mitigate severe interference from the newly introduced communications. In (Wang

et al., 2016), the authors developed a joint bandwidth partitioning and user association strategy that maximizes the sum log rate of the downlink small cell using RTDD. Moreover, Zhao *et al.* (2015) aimed at intuitively minimizing the operational cost of the WB architecture by achieving the least number of admitted WB small cells in the overall network under quality-of-service (QoS) constraints. In (Hu *et al.*, 2016), the authors studied the design of wireless backhaul transmissions in C-RAN. A joint beamforming for the WB links and the user-centric clustering is formulated to maximize the weighted sum rate.

The model of WB system in two-tier HetNets (Wang *et al.*, 2016) is constructed similarly to the multiple point–to–point three–node relays system, where in the special case of one small cell, wireless backhaul system reduces to the traditional two-hop relay system (Zlatanov *et al.*). Among relaying related work, buffer-aided relay has been considered a prominent technique to improve the overall system performance. However, this relay-buffer concept has not been effectively utilized in the existing WB networks such as in (Wang *et al.*, 2016; Sanguinetti *et al.*, 2015; Zhao *et al.*, 2015; Hu *et al.*, 2016), where each SAP is treated as a relay that receives backhaul data from the source MBS and then forwards to its small cell user (SUE). Here, each instantaneous WB rate at the SAP is generally different from the WA rate because of the independence of the fading of these two links (Xia *et al.*, 2008). Without buffering at the intermediate small cell (or relay) node, we observe two important insights from the conventional constraint, where the backhaul rate is considered larger than or equal to the access rate (Wang *et al.*, 2016). First, even when the WA link has a good channel condition, its performance is limited by the WB rate bottleneck with a possibility of poor channel condition (Samarakoon *et al.*, 2013). Second, when the backhaul link has a good channel, it is designed to transmit at a rate at most equal to the WA rate to avoid data drop occurrence (Xia *et al.*, 2008). To enhance the performance beyond these limitations, a buffer utilization, inspired by the buffer-aided relay framework (Xia *et al.*, 2008), can be regarded at each SAP to increase the upper bound of the access rate. This upper bound is now defined as the WB rate with the buffered data in the queue, rather than defined as only the WB rate like in (Wang *et al.*, 2016; Sanguinetti *et al.*, 2015; Zhao *et al.*, 2015; Hu *et al.*, 2016). According to (Xia *et al.*, 2008),

the WB communication can transmit strictly faster than its respective WA communication, and the data which is not forwarded to the SUE can be stored in the buffer of the SAP and be transmitted later in the subsequent time slots. Therefore, exploiting buffer-aided small cell is promising to improve wireless backhaul network.

Due to the aforementioned similarity between WB HetNets and 3-node relay system, designing buffer usage in 3-node relay networks deserves a thorough survey. In the literature, engineering of buffer usage in a three-node relay system has been an attractive research area, e.g., (Nomikos *et al.*, 2016; Zlatanov *et al.*; Xia *et al.*, 2008; Ikhlef *et al.*, 2012; Krikidis *et al.*, 2012; Zlatanov & Schober, 2013; Qiao *et al.*, 2013). These work invoke a framework of optimization to design the power allocation at the source and relay sides subject to various delay constraints such as: infinite delay (Krikidis *et al.*, 2012), average delay with adaptive relaying strategy (Zlatanov & Schober, 2013) or statistical delay with fixed relaying (Qiao *et al.*, 2013). These delay constraints can be interpreted as the requirement that the stored data should not exceed the buffer size. It is noteworthy that in these work, analytical results on queuing theory are applied at the buffer so that asymptotic or probabilistic behavior of the delay is precisely characterized and queuing state information (QSI) and channel state information (CSI) at all time slots are known to achieve the optimal solution. However, considering multiple three-node relay links that takes into account interference is under–explored in the literature so that jointly integrating this buffer-aided relay concept and managing the radio resource under interference-limited regime in the WB small cell HetNets should not be trivially treated. There are also some work sharing this point of view by considering buffer design in their interference-limited network. Peng *et al.* (2016) considered the explicit operation of buffer queue in heterogeneous C-RAN and took into account the expected queue stability constraints under optimizing the network objective using Lyapunov stochastic optimization and weighted minimum mean square error algorithms. In (Lau & Cui, 2010), the authors designed a joint power and sub-carrier allocation for orthogonal frequency division multiple access (OFDMA) system to minimize the system delay by solving a Bellman-equation-based dynamic programming problem and proposed a low complexity online algorithm to achieve the same optimal solution. Guo *et al.*

(2017) presented a dynamic resource allocation for delay-aware application in two-tier small cell networks which aims at maximizing the time-average sum capacity of small cell users subject to the user's delay constraint.

4.1.2 Motivation and Contribution

Motivated by the prominent benefits that buffer can provide on buffer-assisted 3-node relay or large cellular network, in this paper, we propose a novel model of two-tier WB small cell HetNets that considers the finite buffering capability at the SAPs. The novelty of our work lies in the proposal of employing small cell buffer in the two-tier WB small cell HetNets and how to formulate a problem to jointly optimize the usage of small cell buffer together with radio resource allocation to improve the network performance compared to conventional designs (Wang *et al.*, 2016; Nguyen *et al.*, 2016c; Zhao *et al.*, 2015). To avoid cross-tier and self-interference arising from the new half-duplex WB transmissions and cochannel interference from other WA transmissions, we are inspired by the cochannel time division duplexing scheme in (Hoydis *et al.*, 2013) to propose the time–spectrum allocation for the downlink transmissions on top of jointly designing transmit beamforming and power allocation at the MBS and SAPs. Compared to the previous work where the authors investigated the relationship between the WB and WA communications without exploiting the buffering ((Wang *et al.*, 2016; Nguyen *et al.*, 2016c; Zhao *et al.*, 2015)), our work appropriately exploits the SAP buffer by imposing a proper mathematical relationship based on the practical system between WB, WA, and buffered data management to achieve higher WA rate. In particular, we impose a new constraint in our optimization problem which requires that the sum of backhaul transmitted data at the current time slot and the buffered data from the previous time slots is greater than or equal to the access transmitted data. Based on this constraint, when the WB data is higher than the WA data evaluated at the achieved beamforming and power solution, the difference between WB and WA data is stored in the SAP buffer. These buffered data can be transmitted to SUEs in the subsequent time slots when the corresponding WA link has better channel conditions. Although solutions of buffer usage were well explored via 3-node relay system

(Nomikos *et al.*, 2016; Zlatanov *et al.*; Xia *et al.*, 2008; Ikhlef *et al.*, 2012; Krikidis *et al.*, 2012; Zlatanov & Schober, 2013; Qiao *et al.*, 2013), optimally allocating multi-dimensional radio resource to harness the best buffer usage in interference-limited environment, like our considered system, has not been well-investigated in the literature. It is worth to mention that introducing buffer into WB HetNets increases more challenges in the resource allocation tasks. In addition, our work differs from existing back-pressure power control framework for infinite buffer capacity in which power allocation is optimized to minimize the buffered data and stabilize the queue (Peng *et al.*, 2016; Venkatraman *et al.*, 2016b). Rather than stabilizing a finite buffer, we aim at using the buffer queue in a proper manner to maximize the WA sum rate. Our contributions are summarized as follows:

- We examine the impact of our proposed buffer strategy while assuming the availability of instantaneous CSI in all current and future time slots. Hence, we formulate an offline problem of joint design of beamforming and power allocation in a number of time slots that maximizes the WA sum rate. This problem's solution serves as a benchmark against the later–developed online algorithm. The formulated problem is non-convex and NP-hard. Hence, we develop a low-complexity algorithm to solve it via the following framework: (i) transform it into an equivalent and more tractable form; (ii) convexify the transformed problem by proper convex approximation techniques to the respective non-convex constraints.

- Assuming that the transmitter only knows the CSI at the current time slot, we reformulate our problem of interest as an online problem of beamforming and power allocation at each time slot. We also propose to maximize the WB rate at each time slot so that the data transmitted over WB link is higher than that over the WA link. Then, we employ similar steps from the offline approach to develop a low complexity algorithm that solves the online problem.

- We provide theoretical proofs which characterize the superiority of the proposed solutions for offline and online problems compared to the conventional work in (Wang *et al.*, 2016; Nguyen *et al.*, 2016c).

- Extensive numerical results show that our proposed solutions properly take advantage of buffering in WB small cell HetNets and hence improve the performance of the system.

The rest of this chapter is organized as follows. Section 4.2 introduces the system model. In Section 4.3, we formulate the offline and online optimization problems, and present the respective solution approaches to solve them. Section 4.6 presents and discusses our numerical results under different simulation setups. Finally, the concluding remarks are given in Section 4.7.

4.2 System Model

4.2.1 Spatial Model

We consider the downlink of a two-tier WB HetNet consisting of one MBS in the macrocell tier and S_a SAPs in the small cell tier as depicted in Fig. 4.1. The MBS is equipped with $N > 1$ antennas to communicate with its intended receivers, which includes S_a SAPs and M macrocell users (MUEs). We assume each small cell contains S_u SUEs. In this work, the WB communications between the MBS and SAPs operate concurrently with other WB communications and with WA communications from the MBS to its MUEs. Each SAP is equipped with a single antenna to serve S_u SUEs within its small cell coverage. For simplicity, we assume that the number of SUEs in each small cell is the same and the SUEs and MUEs are equipped with a single antenna each.

4.2.2 Time and Spectrum Allocation for Downlink Transmissions

To facilitate the radio resource allocation design presented later for half-duplex transmitters and receivers, we employ the time–spectrum allocation (similar to the cochannel time division duplexing scheme in (Sanguinetti *et al.*, 2015; Hoydis *et al.*, 2013)) to accommodate the WB and WA communications on the channel resources. The time domain is divided into multiple time slots. We assume that the downlink and uplink transmissions are assigned the odd and

even time slots, respectively. In each time slot dedicated for downlink, the spectrum band is split by a factor of α to accommodate the bakchaul downlink transmissions orthogonally to the access ones. The proposed time–spectrum allocation is illustrated in Fig. 4.1.

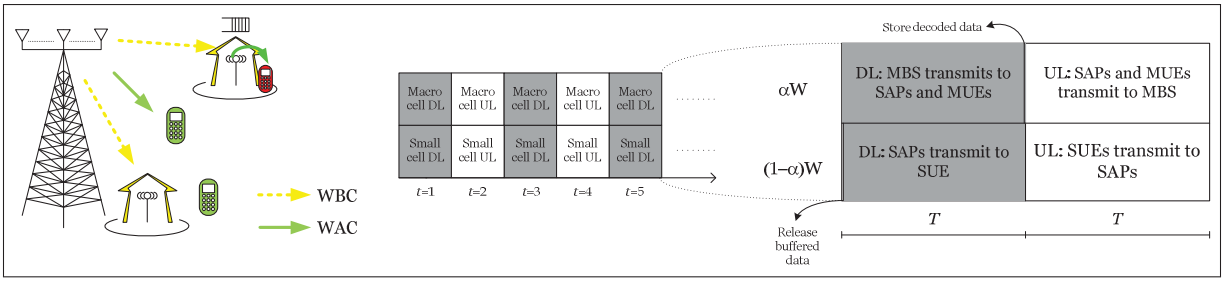

Figure 4.1 Two-tier WB HetNets with small cell buffering with the time–spectrum setting.

4.2.3 Signal Model

Let us consider first the resource blocks dedicated for macrocell downlink transmissions. Here, we assume that one resource block has T seconds and αW Hertz. For convenience, we denote $\mathscr{F} = \{\mathscr{S}_a, \mathscr{M}\} = \{\{1,\ldots,S_a\},\{S_a+1,\ldots,S_a+M\}\}$, where the SAP indices set is $\mathscr{S}_a$ and the MUE indices set is $\mathscr{M}$. By referring the interval $[t,t+1)$ as the t^{th} time slot, the received signal within this resource block at the j^{th} receiver at the t^{th} time slot is

$$y_j(t) = \mathbf{v}_j^H(t)\mathbf{h}_j(t)x_j(t) + \sum_{k\in\mathscr{F}\setminus j} \mathbf{v}_k^H(t)\mathbf{h}_j(t)x_k(t) + n_j(t), \tag{4.1}$$

where $\mathbf{h}_j(t) \in \mathbb{C}^{N\times 1}$ is the channel state vector at the t^{th} time slot which includes fading gain and path-loss components, $\mathbf{v}_j(t) \in \mathbb{C}^{N\times 1}$ is the beamforming vector at the t^{th} time slot from MBS to the j^{th} receiver, $x_j(t)$ is the message at the t^{th} time slot intended for the j^{th} receiver with unit average power, e.g., $\mathbb{E}\left\{x_j(t)\left(x_j(t)\right)^*\right\} = 1$ and $n_j(t)$ is a circularly symmetric complex additive white Gaussian noise (AWGN) at the j^{th} receiver, which is distributed according to a normal distribution $\mathscr{CN}(0,N_{01})$, where N_{01} is the noise power over the allocated spectrum. Treating interference as noise, the achievable rate $R_j(\mathscr{V}(t))$, computed in bits/s/Hz, and the

amount of information $D_j(\mathscr{V}(t))$, computed in bits, at the j^{th} receiver are

$$R_j(\mathscr{V}(t)) = \log\left(1 + \frac{\left|\mathbf{v}_j^H(t)\mathbf{h}_j(t)\right|^2}{\sum\limits_{k\in\mathscr{F}\setminus j}\left|\mathbf{v}_k^H(t)\mathbf{h}_j(t)\right|^2 + N_{01}}\right) \tag{4.2}$$

$$D_j(\mathscr{V}(t)) = \alpha TWR_j(\mathscr{V}(t)), \tag{4.3}$$

where we denote $\mathscr{V}(t) = \{\mathbf{v}_i(t), \forall i \in \mathscr{F}\}$, $\mathscr{V} = \{\mathscr{V}(t), \forall t = 1,3,5,\ldots,t_e\}$, and t_e is the considered number of time slots. $N_{01} = \alpha W\sigma_0$ is the noise power over the αW band and σ_0 is the noise power spectral density. In the other resource block dedicated for the small cell downlink, each i^{th} SAP transmits data to its intended SUEs. Let us denote $p_{ij}(t)$ as the transmit power from the SAP to the j^{th} SUE in the i^{th} small cell and $h_{ikj}(t)$ as the channel from the SAP of the i^{th} small cell to the j^{th} SUE in the another small cell, indexed by k ($k \neq i$) at the t^{th} time slot; $\mathscr{P}(t) = \{p_{ij}(t), \forall (i,j) \in (\mathscr{S}_a, \mathscr{S}_u)\}$, and $\mathscr{P} = \{\mathscr{P}(t), \forall t = 1,3,5,\ldots,t_e\}$; hence, the received signal within this resource block at the j^{th} SUE in the i^{th} small cell is

$$y_{ij}(t) = h_{iij}(t)\sqrt{p_{ij}(t)}s_{ij}(t) + \sum_{k\in\mathscr{S}_u\setminus j} h_{iij}(t)\sqrt{p_{ik}(t)}s_{ik}(t) + \sum_{k\in\mathscr{S}_a\setminus i}\sum_{l\in\mathscr{S}_u} h_{kij}(t)\sqrt{p_{kl}(t)}s_{kl}(t) + z_{ij}(t), \tag{4.4}$$

where $s_{ij}(t)$ is the message at the t^{th} time slot intended for the j^{th} SUE from the SAP in the i^{th} small cell with unit average power, e.g., $\mathbb{E}\left\{s_{ij}(t)s_{ij}^*(t)\right\} = 1$. Similarly, $z_{ij}(t)$ is a circularly symmetric complex AWGN at the SUE from the SAP in the i^{th} small cell, distributed according to a normal distribution $\mathscr{CN}(0, N_{02})$. Treating interference as noise, the achievable rate $r_{ij}(\mathscr{P}(t))$, computed in bits/s/Hz, and the amount of information data $d_{ij}(\mathscr{P}(t))$, computed in bits, at the j^{th} SUE in the i^{th} small cell are

$$r_{ij}(\mathscr{P}(t)) = \log\left(1 + \frac{p_{ij}(t)\left|h_{iij}(t)\right|^2}{I_{ij}(\mathscr{P}(t))}\right) \tag{4.5}$$

$$d_{ij}(\mathscr{P}(t)) = (1-\alpha)TWr_{ij}(\mathscr{P}(t)), \tag{4.6}$$

where

$$I_{ij}(\mathscr{P}(t)) = \sum_{k \in \mathscr{S}_u \setminus j} p_{ik}(t) \left| h_{iij}(t) \right|^2 + \sum_{k \in \mathscr{S}_a \setminus i} \sum_{l \in \mathscr{S}_u} p_{kl}(t) \left| h_{kij}(t) \right|^2 + N_{02}$$

and $N_{02} = (1-\alpha) W \sigma_0$.

4.2.4 Buffering Strategy

We assume that each i^{th} SAP contains a finite buffer that can store the decoded data received from the MBS at the t^{th} time slot. More specifically, at each channel realization, the MBS allocates a value of beamforming to transmit data to the SAP, and the SAP subsequently stores the decoded data in the buffer. In the resource block related to the transmissions from the SAP to its SUEs, the SAP first releases some of these data from the buffer and then transmits. Here, we assume the periods to store data in the buffer and release data out of the buffer are small and can be negligible compared to the transmission period. Note that the relationship between the backhaul and access rate when considering the small cell buffer requires that the combination of backhaul data received from the MBS and the data stored in the buffer should always be greater than or equal to the related access transmitted data. Assuming that with the applied beamforming and power allocation, the sum of backhaul data and amount of buffered data is strictly greater than the related access transmitted data, the SAP only releases sufficient data amount to transmit to its SUE at the current time slot and leaves the remaining data in the buffer. The remaining data will be transmitted in future time slots along with new decoded data received from the MBS. Compared to a conventional approach (in (Wang *et al.*, 2016; Nguyen *et al.*, 2016c; Zhao *et al.*, 2015)) which do not consider this proposed buffer strategy and relies solely on maximizing the small cell rate, the proposed scheme achieves a better performance in terms of sum small cell WA rate by improving the upper limit of the WA rate at each time slot. To better clarify this benefit, the transmission scenario given in Fig. 4.2a compares the transmission scheme of the conventional approach and the proposed scheme.

Fig. 4.2a and 4.2b show how our proposed Scheme I takes advantage of the buffering strategy, and hence the amount of data stored in the buffer in time slot 1 can be accumulated to the

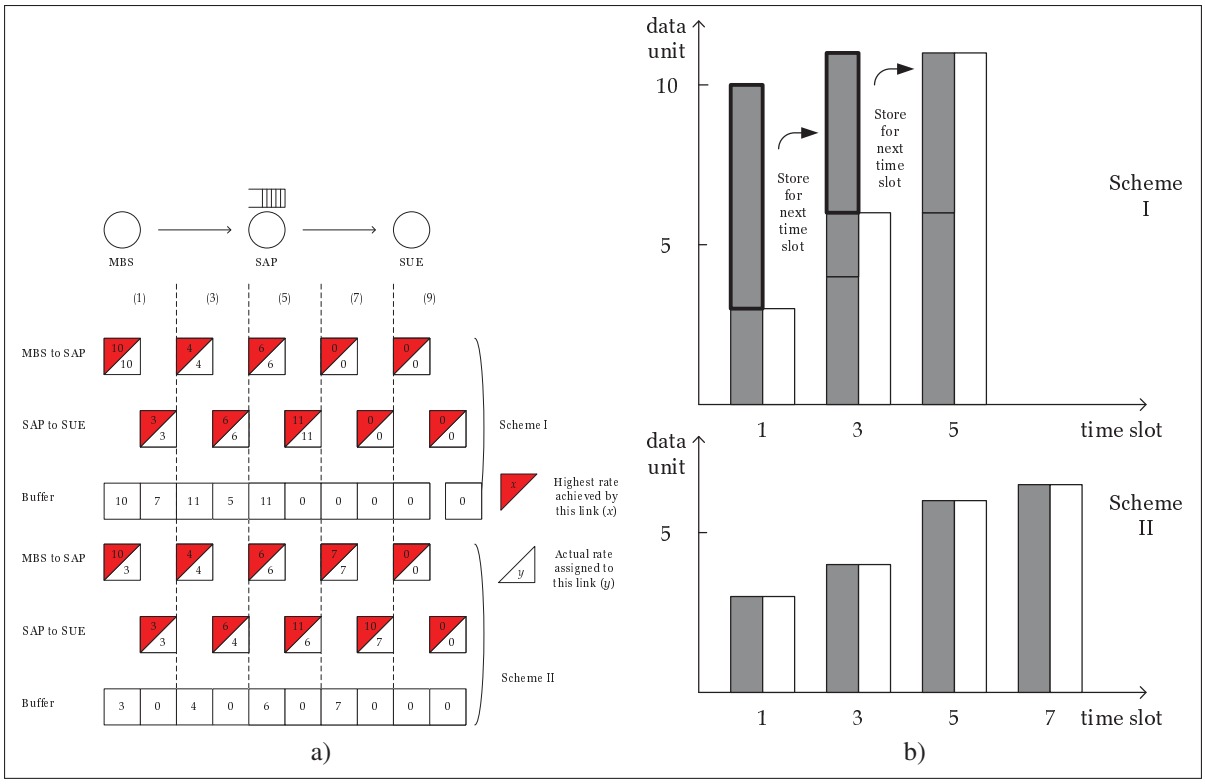

Figure 4.2 (a): An example illustrating the buffering protocol; (b): Buffer data appending protocol. Scheme I refers to the proposed buffering scheme. Scheme II refers to the conventional schemes (Wang *et al.*, 2016; Zhao *et al.*, 2015; Nguyen *et al.*, 2016c).

received data in time slot 2. Consequently, at time slot 3, the WA rate can exploit its good channel to transmit at the rate equal to the highest rate (e.g., 6 data units), since SAP has 11 data units in its buffer. We can observe that up to time slot 5, the total amount of transmitted data of Scheme I and Scheme II are 20 data units and 13 data units, respectively. In Scheme II, the buffer is not fully exploited and there is no data left in the buffer to be combined with the new coming backhaul data in the next time slot. This is because the conventional approach only aims at maximizing the WA rate and the WB rate only needs to reach the necessary WA rate. Consequently, Scheme II needs 8 time slots to complete 20 data units. Hence, we can observe that there is **a throughput improvement of 32%** from Scheme I compared to Scheme II.

4.3 Resource Allocation Optimization Problem

4.3.1 Review of Conventional Design

In this subsection, we motivate our novel problem by briefly describing the effect of the conventional design in (Wang *et al.*, 2016; Nguyen *et al.*, 2016c; Zhao *et al.*, 2015) on our problem of interest. Here, we assume that the transmitters know only the CSI of the current time slot. The CSI varies independently each consecutive time slot. In the context of wireless backhaul HetNets, the authors in (Wang *et al.*, 2016; Nguyen *et al.*, 2016c) proposed the backhaul-access constraint where the backhaul rate is greater than or equal to the access rate in each time slot. Based on this, the resulted optimization problem, denoted herein as Conventional Scheme, which jointly designs the transmit beamforming and power allocation at the MBS and SAPs to maximize the WA sum small cell data in all time slots is

$$(\mathscr{P}_{\text{con}}): \max_{\mathscr{V},\mathscr{P}} \quad \sum_{t=1,3,\ldots}^{t_e} \sum_{i\in\mathscr{S}_a} \sum_{j\in\mathscr{S}_u} d_{ij}(\mathscr{P}(t)) \tag{4.7a}$$

$$\text{s.t.} \quad D_i(\mathscr{V}(t)) \geq \sum_{j\in\mathscr{S}_u} d_{ij}(\mathscr{P}(t)), \forall i \in \mathscr{S}_a, \tag{4.7b}$$

$$D_i(\mathscr{V}(t)) \leq C_{\max}, \forall i \in \mathscr{S}_a, \tag{4.7c}$$

$$R_k(\mathscr{V}(t)) \geq R_{\min}, \forall k \in \mathscr{M}, \tag{4.7d}$$

$$r_{i_1 j_1}(\mathscr{P}(t)) \geq r_{\min}, (i_1, j_1) \in (\mathscr{S}_a^{\text{sub}}, \mathscr{S}_u^{\text{sub}}) \tag{4.7e}$$

$$\sum_{k\in\mathscr{F}} \|\mathbf{v}_k(t)\|^2 \leq P_m, \sum_{j\in\mathscr{S}_u} p_{ij}(t) \leq p_m, \tag{4.7f}$$

$$\forall t = 1, 3 \ldots, t_e, \tag{4.7g}$$

where constraint (4.7b) reflects the design property of Conventional Scheme which requires that the amount of WB transmitted data that is greater than or equal to WA transmitted data of the i^{th} SAP in each time slot. This is obvious because the SAP cannot transmit, via WA channel to the SUE, greater data than the amount it received from the MBS via WB channel. In some cases, the achieved WB rate is less than the WA rate when we apply some value of $\mathscr{V}, \mathscr{P}$.

These values are indeed infeasible to constraint (4.7b). In these cases, via solving problem (4.7) with constraint (4.7b), the power allocation of the WA transmissions is coordinated to decrease to achieve the WA rate no greater than the WB rate. Without (4.7b), if the WA data evaluated at the solved resource allocation is greater than the WB data, the system will be unreasonable since the available backhaul data is insufficient to support the WA transmission. (4.7c) requires that the amount of data stored in the i^{th} buffer in the t^{th} time slot should not exceed the maximum buffer storage $C_{\max}$. (4.7d) and (4.7e) respectively guarantee the QoS minimum rate requirement for each MUE and for some SUE indices belong to the set $\mathscr{S}_a^{\text{sub}}, \mathscr{S}_u^{\text{sub}}$, where $\mathscr{S}_a^{\text{sub}} \subseteq \mathscr{S}_a, \mathscr{S}_u^{\text{sub}} \subseteq \mathscr{S}_u$. Finally, (4.7f) are the maximum power budget at the MBS and SAP, respectively. We remark that at optimality, (4.7b) can occur at strict inequality at some SAPs. This implies that MBS has transmitted more backhaul data than what those SAP can deliver to their SUEs. Although the SAP can transmit some of the received WB data to its SUEs, there exists some data which cannot be sent so that packet dropping and data retransmissions can possibly occur. Besides, Section 4.2.4 has thoroughly discussed how the WA sum small cell rate can be improved by properly taking advantage of the buffered data. Hence, we are motivated to propose a problem with a redesign of the backhaul-access rate constraint (4.7b).

4.3.2 Proposed Offline Problem Formulation

To study the improvement resulted from the proposed scheme, let us first visit an ideal offline design of transmit beamforming and power allocation at the MBS and SAPs. In this scenario, we assume that the transmitters know the CSI for the current and future time slots. For convenience, we denote this as Proposed Offline Scheme. The discussion in Section 4.2.4 implies that the backhaul-access constraint must enforce the data received at the SUE at the current time slot be less than or equal to the data left in the buffer from the previous time slots combined with the received data at the SAP via WB channel. Therefore, the problem of maximizing the WA sum small cell data during all time slots in this scheme is given by

$$(\mathscr{P}_{\text{off}}) : \max_{\mathscr{V},\mathscr{P}} \quad \sum_{t=1,3,\ldots}^{t_e} \sum_{i\in\mathscr{S}_a} \sum_{j\in\mathscr{S}_u} d_{ij}(\mathscr{P}(t)) \tag{4.8a}$$

$$\text{s.t.} \quad D_i(\mathscr{V}(t)) + Q_i(\mathscr{V}(t-2), \mathscr{P}(t-2)) \geq \sum_{j \in \mathscr{S}_u} d_{ij}(\mathscr{P}(t)), \forall i \in \mathscr{S}_a \tag{4.8b}$$

$$D_i(\mathscr{V}(t)) + Q_i(\mathscr{V}(t-2), \mathscr{P}(t-2)) \leq C_{\max}, \forall i \in \mathscr{S}_a \tag{4.8c}$$

$$(4.7\text{d}) - (4.7\text{g}) \tag{4.8d}$$

where we denote

$$Q_i(\mathscr{V}(t-2), \mathscr{P}(t-2)) = \sum_{l=1,3,\ldots,}^{t-2} D_i(\mathscr{V}(l)) - \sum_{l=1,3,\ldots,}^{t-2} \sum_{j \in \mathscr{S}_u} d_{ij}(\mathscr{P}(l))$$

as the amount of data remaining in the buffer collected from all the time slots before time slot t. Unlike (4.7b), (4.8b) shows that the data received at each SUE in the i^{th} small cell at the t^{th} time slot should be upper bounded by the buffered data at the $(t-2)^{\text{th}}$ time slot plus the data received at the i^{th} SAP at the t^{th} time slot. Interestingly, the solution $\mathscr{V}, \mathscr{P}$ allows the WB rate to be lower than the WA rate as long as computed WB data plus the buffered data is greater than the computed WA data at the solution $\mathscr{V}, \mathscr{P}$. In addition, the data stored in the i^{th} buffer should be upper bounded by $C_{\max}$ as in (4.8c). The following proposition characterizes the superiority of the Proposed Offline Scheme over the Conventional Scheme.

Proposition 3. *The sum of optimal objective in* (4.7) *(Conventional Scheme) over t_e time slots is always less than or equal to the sum of optimal objective achieved in* (4.8) *(Proposed Offline Scheme) over t_e time slots.*

Proof. The proof is in Appendix 7. □

The formulated optimization problem (4.8) is non-convex due to the existence of the non-convex non-concave rate formula as a function of variables $\mathscr{V}$ and $\mathscr{P}$. In general, solving (4.8) optimally is difficult since this problem is categorized as NP-hard problem (Liu *et al.*, 2011a). Thus, developing a polynomial-time algorithm to achieve suboptimal solution is more realistic. In the following, we rely on the framework of equivalent transformations and convex approximation method (Marks & Wright) to develop an offline algorithm to solve (4.8) and to evaluate the performance of the Proposed Offline Scheme.

Remark 2. *Problem* (4.8) *is not always feasible due to the minimum rate constraints* (4.7d) *and* (4.7e). *We note that each considered transmitter, e.g., SAP or MBS, has a limited power budget, so that having low P_m and p_m could make* (4.8) *infeasible. Moreover, our system operates under the interference-limited regime, and hence, imposing high value of $R_{\min}$ and $r_{\min}$ could also lead to the infeasibility of* (4.8). *To facilitate these issues and direct our concentration on the low-complexity design, we only consider the feasible problem with sufficiently low $R^{\min}$ and $r_{\min}$. In case* (4.8) *is infeasible, we can reschedule to admit a smaller set of users which satisfy the minimum rate requirement and employ our developed framework to find solution.*

4.4 Evaluation of Offline Approach

In this section, we jointly solve the problem (4.8) for all variables $\mathscr{V}$ and $\mathscr{P}$ in all time slots. First, by introducing the slack variables $\mathscr{U} = \{u_i(t) \geq 0, \forall i \in \mathscr{S}_a, t = 1,3,5\ldots,t_e\}$, $\mathscr{W} = \{w_i(t) \geq 0, \forall i \in \mathscr{S}_a, t = 1,3,5\ldots,t_e\}$, $\mathscr{X} = \{\mu_{ij}(t) \geq 0, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, t = 1,3,5,\ldots,t_e\}$, and $\mathscr{Y} = \{\nu_{ij}(t) \geq 0, \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u, t = 1,3,5,\ldots,t_e\}$, we can equivalently rewrite (2.7) as

$$\max_{\substack{\mathscr{V},\mathscr{P},\mathscr{U},\\ \mathscr{W},\mathscr{X},\mathscr{Y}}} \quad \sum_{t=1,3,\ldots}^{t_e} \sum_{i\in\mathscr{S}_a} \sum_{j\in\mathscr{S}_u} \nu_{ij}(t) \tag{4.9a}$$

$$\text{s.t.} \quad w_i(t) \leq R_i(\mathscr{V}(t)) \leq u_i(t), \forall i \in \mathscr{S}_a \tag{4.9b}$$

$$\nu_{ij}(t) \leq r_{ij}(\mathscr{P}(t)) \leq \mu_{ij}(t), \forall i \in \mathscr{S}_a, \forall j \in \mathscr{S}_u \tag{4.9c}$$

$$\alpha \sum_{l=1,3,\ldots,}^{t} w_i(l) \geq (1-\alpha) \sum_{l=1,3,\ldots,}^{t} \sum_{j\in\mathscr{S}_u} \mu_{ij}(l), \forall i \in \mathscr{S}_a \tag{4.9d}$$

$$\alpha \sum_{l=1,3,\ldots,}^{t} u_i(l) - (1-\alpha) \sum_{l=1,3,\ldots,}^{t} \sum_{j\in\mathscr{S}_u} \nu_{ij}(l) \leq C_{\max}/TW, \forall i \in \mathscr{S}_a \tag{4.9e}$$

$$(4.7\text{d}) - (4.7\text{g}). \tag{4.9f}$$

The equivalence between (2.7) and (4.9) is given in Appendix 6. Next, by further investigating (4.9b), we can decouple these two constraints into four equivalent inequalities as

$$f(\mathbf{v}_i(t), \pi_i(t)) = \log\left(1 + \left|\mathbf{v}_i(t)^H \mathbf{h}_i(t)\right|^2 / \pi_i(t)\right) \geq w_i(t), \tag{4.10a}$$

$$\sum_{k\in\mathscr{F}\setminus i}\left|\mathbf{v}_k(t)^H\mathbf{h}_i(t)\right|^2+N_{01}\leq\pi_i(t), \tag{4.10b}$$

$$1+\frac{\left|\mathbf{v}_i(t)^H\mathbf{h}_i(t)\right|^2}{\phi_i(t)}\leq\exp(u_i(t)), \tag{4.10c}$$

$$g(\mathscr{V}(t)_{\sim i})=\sum_{k\in\mathscr{F}\setminus i}|\mathbf{v}_k(t)\mathbf{h}_i(t)|^2+N_{01}\geq\phi_i(t), \tag{4.10d}$$

where $\pi_i(t)\geq 0$ and $\phi_i(t)\geq 0$ are the newly introduced variables. Here, we denote $\Pi=\{\pi_i(t)\geq 0,\forall i\in\mathscr{S}_a,t=1,3,5\ldots,t_e\}$ and $\Phi=\{\phi_i(t)\geq 0,\forall i\in\mathscr{S}_a,t=1,3,5\ldots,t_e\}$. The equivalence between (4.9b) and (4.10a)-(4.10d) can be explained similar to the proof in Appendix 6. Clearly, (4.10a), (4.10c), and (4.10d) are non-convex constraints because the convex functions $f(\mathbf{v}_i(t),\pi_i(t))$, $\exp(u_i(t))$, and $g(\mathscr{V}(t)_{\sim i})$ lie on the greater sides of the inequalities. Thus, we approximate the convex functions in (4.10a), (4.10c), and (4.10d) by their lower bound concave approximation around the points $\mathbf{v}_i(t)^{(n)}$, $\pi_i(t)^{(n)}$, $u_i(t)^{(n)}$, and $\mathscr{V}(t)^{(n)}_{\sim i},\forall i\in\mathscr{S}_a$ as

$$\begin{aligned}f(\mathbf{v}_i(t),\pi_i(t))\geq F(\mathbf{v}_i(t),\pi_i(t);\mathbf{v}_i(t)^{(n)},\pi_i(t)^{(n)})=\log\left[1+\frac{\left|\mathbf{v}_i(t)^{(n)H}\mathbf{h}_i(t)\right|^2}{\pi_i(t)^{(n)}}\right]-\frac{\left|\mathbf{v}_i(t)^{(n)H}\mathbf{h}_i(t)\right|^2}{\pi_i(t)^{(n)}}\\+\frac{2\mathscr{R}\left(\mathbf{v}_i(t)^{(n)H}\mathbf{H}_i(t)\mathbf{v}_i(t)\right)}{\pi_i(t)^{(n)}}-\frac{\left|\mathbf{v}_i(t)^{(n)H}\mathbf{h}_i(t)\right|^2\left[\left|\mathbf{v}_i(t)^{(n)H}\mathbf{h}_i(t)\right|^2+\pi_i(t)\right]}{\pi_i(t)^{(n)}\left[\left|\mathbf{v}_i(t)^{(n)H}\mathbf{h}_i(t)\right|^2+\pi_i(t)^{(n)}\right]},\end{aligned} \tag{4.11}$$

$$\exp(u_i(t))\geq\exp(u_i(t)^{(n)})+\exp(u_i(t)^{(n)})(u_i(t)-u_i(t)^{(n)}), \tag{4.12}$$

$$\begin{aligned}g(\mathscr{V}(t)_{\sim i})\geq G\left(\mathscr{V}(t)_{\sim i};\mathscr{V}(t)^{(n)}_{\sim i}\right)=2\mathscr{R}\left(\sum_{k\in\mathscr{F}\setminus i}\mathbf{v}_k(t)^{(n)H}\mathbf{H}_i(t)\mathbf{v}_k(t)\right)\\-\sum_{k\in\mathscr{F}\setminus i}\mathbf{v}_i(t)^{(n)H}\mathbf{H}_i(t)\mathbf{v}_k(t)^{(n)}+N_{01},\end{aligned} \tag{4.13}$$

Here, we briefly discuss the properties of $F\left(\mathbf{v}_i(t),\pi_i(t);\mathbf{v}_i(t)^{(n)},\pi_i(t)^{(n)}\right)$. First, we note that $f(\mathbf{v}_i(t),\pi_i(t))=F\left(\mathbf{v}_i(t),\pi_i(t);\mathbf{v}_i(t)^{(n)},\pi_i(t)^{(n)}\right)$ when $\mathbf{v}_i(t)=\mathbf{v}_i(t)^{(n)}$ and $\pi_i(t)=\pi_i(t)^{(n)}$. Moreover, with this selection of $\mathbf{v}_i(t)$ and $\pi_i(t)$, we can easily check that the first derivative of $F\left(\mathbf{v}_i(t),\pi_i(t);\mathbf{v}_i(t)^{(n)},\pi_i(t)^{(n)}\right)$ with respect to either $\mathbf{v}_i(t)$ or $\pi_i(t)$ is equal to that of $f(\mathbf{v}_i(t),\pi_i(t))$. The same argument can be made for $G\left(\mathscr{V}(t)_{\sim i};\mathscr{V}(t)^{(n)}_{\sim i}\right)$ and its counterpart

$g(\mathscr{V}(t)_{\sim i})$. These properties are important to conclude the convergence of the iterative algorithm developed later. Alternatively, we can rewrite (4.9c) into two equivalent inequalities as

$$\log\left(p_{ij}(t)\left|h_{iij}(t)\right|^2+I_{ij}(\mathscr{P}(t))\right)\geq \nu_{ij}(t)+\log\left(I_{ij}(\mathscr{P}(t))\right) \tag{4.14a}$$

$$\log\left(p_{ij}(t)\left|h_{iij}(t)\right|^2+I_{ij}(\mathscr{P}(t))\right)\leq \mu_{ij}(t)+\log\left(I_{ij}(\mathscr{P}(t))\right) \tag{4.14b}$$

Both (4.14a) and (4.14b) are non-convex constraints because the concave functions lie in the lesser side of the inequalities. By denoting $g_{ij}^{I}(\mathscr{P}(t))=I_{ij}(\mathscr{P}(t))$ and $g_{ij}^{SI}(\mathscr{P}(t))=p_{ij}(t)\left|h_{iij}(t)\right|^2+I_{ij}(\mathscr{P}(t))$, we employ an approximate to function $\log\left(g_{ij}^{I}(\mathscr{P}(t))\right)$ in (4.14a) and $\log\left(g_{ij}^{SI}(\mathscr{P}(t))\right)$ in (4.14b) around the point $\mathscr{P}(t)^{(n)}$ by their upper bound convex approximation as follow

$$\begin{cases}\log\left(g_{ij}^{I}(\mathscr{P}(t))\right)\leq G_{ij}^{I}\left(\mathscr{P}(t);\mathscr{P}(t)^{(n)}\right)= & \log\left(g^{I}\left(\mathscr{P}(t)^{(n)}\right)\right)+\frac{g_{ij}^{I}(\mathscr{P}(t))-g_{ij}^{I}\left(\mathscr{P}(t)^{(n)}\right)}{g_{ij}^{I}\left(\mathscr{P}(t)^{(n)}\right)}\\ \log\left(g_{ij}^{SI}(\mathscr{P}(t))\right)\leq G_{ij}^{SI}\left(\mathscr{P}(t);\mathscr{P}(t)^{(n)}\right)= & \log\left(g^{SI}\left(\mathscr{P}(t)^{(n)}\right)\right)+\frac{g_{ij}^{SI}(\mathscr{P}(t))-g_{ij}^{SI}\left(\mathscr{P}(t)^{(n)}\right)}{g_{ij}^{SI}\left(\mathscr{P}(t)^{(n)}\right)}\end{cases} \tag{4.15}$$

By applying the approximation in (4.11)–(4.13) to the non-convex constraints (4.10a) and (4.10d), (4.12)to constraint (4.10c), and (4.15) to the constraints (4.14a)–(4.14b), problem (4.9) can be solved by iteratively solving the following approximated convex problem formulated at the $n+1$ iteration index as

$$\max_{\substack{\mathscr{V},\mathscr{P},\mathscr{U},\mathscr{W}\\ \mathscr{X},\mathscr{Y},\Pi,\Phi}} \quad \sum_{t=1,3,\ldots}^{t_e}\sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u}\nu_{ij}(t) \tag{4.16a}$$

$$\text{s.t.} \quad F\left(\mathbf{v}_i(t),\pi_i(t);\mathbf{v}_i(t)^{(n)},\pi_i(t)^{(n)}\right)\geq w_i(t),\forall i\in\mathscr{S}_a, \tag{4.16b}$$

$$\sum_{k\in\mathscr{F}\setminus i}\left|\mathbf{v}_k(t)^H\mathbf{h}_i(t)\right|^2+N_0\leq\pi_i(t),\forall i\in\mathscr{S}_a, \tag{4.16c}$$

$$\left|\mathbf{v}_i(t)^H\mathbf{h}_i(t)\right|^2/\phi_i(t)+1\leq e^{u_i(t)^{(n)}}+e^{u_i(t)^{(n)}}(u_i(t)-u_i(t)^{(n)}),\forall i\in\mathscr{S}_a, \tag{4.16d}$$

$$G\left(\mathscr{V}_{\sim i}(t);\mathscr{V}_{\sim i}(t)^{(n)}\right)\geq\phi_i(t),\forall i\in\mathscr{S}_a, \tag{4.16e}$$

$$\log\left(p_{ij}(t)\left|h_{iij}(t)\right|^2+I_{ij}(\mathscr{P}(t))\right)\geq\nu_{ij}(t)+$$

$$G_{ij}^{I}\left(\mathscr{P}(t);\mathscr{P}(t)^{(n)}\right),\forall(i,j)\in(\mathscr{S}_a,\mathscr{S}_u), \tag{4.16f}$$

$$\mu_{ij}(t)+\log\left(I_{ij}(\mathscr{P}(t))\right)\geq G_{ij}^{SI}\left(\mathscr{P}(t);\mathscr{P}(t)^{(n)}\right)\forall i\in\mathscr{S}_a,\forall(i,j)\in(\mathscr{S}_a,\mathscr{S}_u), \tag{4.16g}$$

$$(4.9d),(4.9e),(4.7d)-(4.7g). \tag{4.16h}$$

Note that parameters $\mathscr{P}^{(n)},\mathscr{V}^{(n)},\Pi^{(n)},\Phi^{(n)},\mathscr{U}^{(n)}$ are iteratively updated by the optimal solution achieved by solving (4.16) at each iteration. The pseudo code that presents the algorithm to solve problem (4.9) is summarized in Algorithm 4.

Algorithm 4: Offline Joint Beamforming and Power Allocation Optimization (JBPAO) Algorithm

1: Set $n:=0$;
2: Initialize starting point of $\mathscr{P}^{(n)},\mathscr{V}^{(n)},\Pi^{(n)},\Phi^{(n)},\mathscr{U}^{(n)}$;
3: **repeat**
4: Solve the convex problem in (4.16) to achieve the optimal solution $\mathscr{V}^{\star},\mathscr{P}^{\star},\mathscr{U}^{\star},\mathscr{W}^{\star},\mathscr{X}^{\star},\mathscr{Y}^{\star},\Pi^{\star},\Phi^{\star}$;
5: Set $n:=n+1$;
6: Update $\mathscr{P}^{(n)}=\mathscr{P}^{\star},\mathscr{V}^{(n)}=\mathscr{V}^{\star},\Pi^{(n)}=\Pi^{\star},\Phi^{(n)}=\Phi^{\star},\mathscr{U}^{(n)}=\mathscr{U}^{\star}$;
7: **until** Convergence of (4.16a);

Convergence Analysis: Let $f^{(n)}$ denote the optimal objective value and $\Omega^{(n)}$ denote the optimal solution set at the n^{th} iteration of Algorithm 4. Due to the convex approximation in (4.11)–(4.13) and (4.15), the updating rules in Algorithm 4 ensure that the solution set $\Omega^{(n)}$ is a feasible solution to problem (4.16) at step $n+1$. This subsequently leads to the results of $f^{(n+1)}\geq f^{(n)}$, which means that Algorithm 4 generates a non-decreasing sequence of objective function values. Due to the limited power constraints, the sequence of $f^{(n)},n=1,2,\ldots$ is bounded above and therefore, Algorithm 4 guarantees that the objective converges.

Generation of starting point: A starting point of Algorithm 4 is the value of the set of variables $\mathscr{P},\mathscr{V},\Pi,\Phi,\mathscr{U}$ which are feasible for problem (4.9). Although (4.9) is non-convex, we can

choose its feasible solution by solving the convex problem

$$\min_{\mathscr{V},\mathscr{P}} \quad \sum_{t=1}^{t_e}\sum_{i\in\mathscr{F}} \|\mathbf{v}_i(t)\|_2^2 + \sum_{t=1}^{t_e}\sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} p_{ij}(t) \tag{4.17a}$$

$$\text{s.t.} \quad p_{i_1j_1}(t)\left|h_{i_1i_1j_1}(t)\right|^2 = (e^{r^{\min}}-1)I_{i_1j_1}(\mathscr{P}(t)), \tag{4.17b}$$

$$\mathrm{Re}(\mathbf{v}_{i_1}(t)^H\mathbf{h}_{i_1}(t)) \geq A^{0.5}\left(\sum_{l\in\mathscr{F}\setminus i}\left|\mathbf{v}_l(t)^H\mathbf{h}_{i_1}(t)\right|^2 + N_{01}\right)^{0.5} \tag{4.17c}$$

$$\mathrm{Re}(\mathbf{v}_k(t)^H\mathbf{h}_k(t)) \geq (e^{R_{\min}}-1)^{0.5}\left(\sum_{l\in\mathscr{F}\setminus k}\left|\mathbf{v}_l(t)^H\mathbf{h}_k(t)\right|^2 + N_{01}\right)^{0.5} \tag{4.17d}$$

$$(4.7\text{d}) - (4.7\text{g}). \tag{4.17e}$$

for $i_1 \in \mathscr{S}_a^{\text{sub}}, j_1 \in \mathscr{S}_u^{\text{sub}}, k \in \mathscr{M}$, where $A = (1-\alpha)\left(e^{\sum_{j\in\mathscr{S}_u} r^{\min}} - 1\right)/\alpha$. Note that problem (4.17) is convex since the objective function and all constraints are convex with respect to variables $\mathscr{V}, \mathscr{P}$. By denoting $\mathscr{V}^{(0)}, \mathscr{P}^{(0)}$ as the optimal solution of (4.17), we can compute each i^{th} element in $\Pi^{(0)}, \Phi^{(0)}, \mathscr{U}^{(0)} \forall i \in \mathscr{S}_a$ as follow

$$\pi_i(t)^{(0)} = \phi_i(t)^{(0)} = \sum_{k\in\mathscr{F}\setminus i}\left|\mathbf{v}_k(t)^{(0)H}\mathbf{h}_i(t)\right|^2 + N_{01} \tag{4.18}$$

$$u_i(t)^{(0)} = \log(1 + \left|\mathbf{v}_i(t)^{(0)H}\mathbf{h}_i(t)\right|^2/\phi_i(t)^{(0)}) \tag{4.19}$$

4.5 Proposed Online Problem Formulation

In the previous section, a low complexity offline JBPAO algorithm is proposed to determine the maximum sum WA data at the SUEs within t_e time slots. Since assuming the availability of the current and future CSI at the transmitters is infeasible in practice, we tackle the design problem by assuming that the transmitters only know the CSI at the current time slot. We denote this scheme as Proposed Online Scheme. Similar to Section 4.3.1, we assume that the CSI is the same in time slot t and varies independently each consecutive time slot. Recall that in the offline setting, jointly solving the beamforming and power allocation allows us to allocate the best tuple of WB and WA rates at all time slots, in which the total WA rate performance has been

shown to be more superior than the Conventional Scheme. On the contrary, this cannot be done in the same manner in the online setting due to the lackness of future CSI. Therefore, in order to improve the WA sum achieved data by employing the buffering strategy in Section 4.2.4, the MBS and SAP must set their transmit beamforming and power to maximize both WA and WB rates subject to the new online backhaul-access constraint. Similar to (4.8b), this online constraint requires that the data received at the SUE at the current time slot should be less than or equal to the data left in the buffer from the previous time slots combined with the received data from the MBS at the SAP. Thus, the online problem of interest at time slot t can be written as two separate problems and is solved in two stages. The first stage problem is given by

$$\left(\mathscr{P}^1_{\text{on}}\right): \max_{\mathscr{V}(t),\mathscr{P}(t)} \quad \sum_{(i,j)\in(\mathscr{S}_a,\mathscr{S}_u)} r_{ij}(\mathscr{P}(t)) \tag{4.20a}$$

$$\text{s.t.} \quad D_i(\mathscr{V}(t)) + Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2)) \geq \sum_{j\in\mathscr{S}_u} d_{ij}(\mathscr{P}(t)), \forall i \in \mathscr{S}_a \tag{4.20b}$$

$$D_i(\mathscr{V}(t)) + Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2)) \leq C_{\max}, \forall i \in \mathscr{S}_a \tag{4.20c}$$

$$(4.7\text{d}) - (4.7\text{g}), \tag{4.20d}$$

where $Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2)) = \sum_{l=1,3,\ldots,}^{t-2} D_i(\bar{\mathscr{V}}(l)) - \sum_{l=1,3,\ldots,}^{t-2} \sum_{j\in\mathscr{S}_u} d_{ij}(\bar{\mathscr{P}}(l))$. The second stage problem is given by

$$\left(\mathscr{P}^2_{\text{on}}\right): \max_{\mathscr{V}(t)} \quad U\left(\Gamma(\mathscr{V}(t))\right) \tag{4.21a}$$

$$\text{s.t.} \quad D_i(\mathscr{V}(t)) + Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2)) \geq \sum_{j\in\mathscr{S}_u} d_{ij}(\bar{\mathscr{P}}(t)), \forall i \in \mathscr{S}_a \tag{4.21b}$$

$$D_i(\mathscr{V}(t)) + Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2)) \leq C_{\max}, \forall i \in \mathscr{S}_a \tag{4.21c}$$

$$(4.7\text{d}) - (4.7\text{g}) \tag{4.21d}$$

where $U\left(\Gamma(\mathscr{V}(t))\right)$ is the utility function of received SINR at each SAP, which is often chosen as a concave and increasing function with respect to $\Gamma(\mathscr{V}(t))$, e.g.,

$$U\left(\Gamma(\mathscr{V}(t))\right) = \begin{cases} \sum_{i\in\mathscr{S}_a} \log\left(1+\Gamma_i(\mathscr{V}(t))\right) & \text{, sum rate} \\ \min_{i\in\mathscr{S}_a} \Gamma_i(\mathscr{V}(t)) & \text{, max-min-fairness} \end{cases} \tag{4.22}$$

Here, we note that the difference between the problem formulation (4.20)-(4.21) of the Proposed Online Scheme and (2.7) of the Proposed Offline Scheme lies in the attempt to maximize both WB and WA rates. Since future CSI is unavailable in the online setting, to take advantage of the data stored in the buffer queue (as discussed in Section 4.2.4), an intuition is to additionally maximize the WB rate beside WA sum rate. By doing that, when the WB channel is good, we can achieve WB rate strictly higher than the corresponding WA rate. Then, the difference between WB and WA rates can be stored in the buffer. When there is a bad WB channel in future time slots and the WA communication can potentially achieve a larger rate than the corresponding WB communication rate, the stored data can be added to the WB data to improve the upper bound of the WA rate, as reflected in (4.20b). This observation can be exploited by first attempting to find the solution that maximizes the sum WA rates at all the SAPs in (4.20) first while still ensuring the backhaul-access relationship as in constraint (4.20b). Once we obtain the desired solution for (4.20), we can evaluate the corresponding $Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2))$ and $d_{ij}(\bar{\mathscr{P}}(t))$ and fix these values to solve the problem of maximizing utility function of received SINR at each SAP as in (4.21). Again, we also note that unlike (4.7b), constraints (4.20b) and (4.21b) in the Proposed Online Scheme show that the data received at each SUE in the i^{th} small cell via WA at the t^{th} time slot should be upper bounded by the remaining data in the buffer at the previous $(t-2)^{\text{th}}$ time slot together with the data received at the i^{th} SAP via WB communication in the t^{th} time slot. In addition, to integrate the limited buffer constraint, the amount of data stored in the i^{th} buffer should be upper bounded by $C_{\max}$ as in (4.8c). The following proposition characterizes the superiority of the Proposed Online Scheme compared to the Conventional Scheme.

Proposition 4. *The sum of optimal objective in* (4.7) *(Conventional Scheme) over t_e time slots is always less than or equal to the sum of optimal objective achieved in* (4.20) *(Proposed Online Scheme) over t_e time slots.*

Proof. The proof is similar to that of Proposition 3 in Appendix 7. □

4.5.1 Proposed Online Algorithm

By taking a similar analysis to problem (4.8) in Section 4.3.2, we also observe that (4.20) and (4.21) are non-convex optimization problems. Thus, by applying similar techniques of equivalent transformation and invoking the convex approximation in Section 4.4, the approximated online optimization problem of (4.20) can be presented as

$$\max_{\substack{\mathscr{V}(t),\bar{\mathscr{P}}(t),\mathscr{U}(t),\mathscr{W}(t),\\ \mathscr{X}(t),\mathscr{Y}(t),\Pi(t),\Phi(t)}} \quad \sum_{i\in\mathscr{S}_a}\sum_{j\in\mathscr{S}_u} v_{ij}(t) \tag{4.23a}$$

$$\text{s.t.} \quad \alpha w_i(t) + \frac{Q_i(\bar{\mathscr{V}}(t-2),\bar{\mathscr{P}}(t-2))}{TW} \geq (1-\alpha)\sum_{j\in\mathscr{S}_u}\mu_{ij}(t), \forall i\in\mathscr{S}_a \tag{4.23b}$$

$$\alpha u_i(t) - (1-\alpha)\sum_{j\in\mathscr{S}_u} v_{ij}(t) + \frac{Q_i(\bar{\mathscr{V}}(t-2),\bar{\mathscr{P}}(t-2))}{TW} \leq \frac{C_{\max}}{TW}, \forall i\in\mathscr{S}_a \tag{4.23c}$$

$$(4.16b)-(4.16g),(4.7d)-(4.7g). \tag{4.23d}$$

and the approximated online optimization problem of (4.21) can be presented as

$$\max_{\substack{\mathscr{V}(t),\mathscr{U}(t),\mathscr{W}(t),\\ \Pi(t),\Phi(t)}} \quad \sum_{i\in\mathscr{S}_a} U\left(w_i(t)\right) \tag{4.24a}$$

$$\text{s.t.} \quad \alpha w_i(t) + \frac{Q_i(\bar{\mathscr{V}}(t-2),\bar{\mathscr{P}}(t-2))}{TW} \geq \frac{\sum_{j\in\mathscr{S}_u} d_{ij}(\bar{\mathscr{P}}(t))}{TW}, \forall i\in\mathscr{S}_a \tag{4.24b}$$

$$\alpha u_i(t) - \frac{\sum_{j\in\mathscr{S}_u} d_{ij}(\bar{\mathscr{P}}(t)) - Q_i(\bar{\mathscr{V}}(t-2),\bar{\mathscr{P}}(t-2))}{TW} \leq \frac{C_{\max}}{TW}, \forall i\in\mathscr{S}_a \tag{4.24c}$$

$$(4.16b)-(4.16e),(4.7d)-(4.7g). \tag{4.24d}$$

where we denote $\mathscr{U}(t)=\{u_i(t)\geq 0,\forall i\in\mathscr{S}_a\}$, $\mathscr{W}(t)=\{w_i(t)\geq 0,\forall i\in\mathscr{S}_a\}$, $\mathscr{X}(t)=\{\mu_{ij}(t)\geq 0,\forall i\in\mathscr{S}_a,\forall j\in\mathscr{S}_u\}$, $\mathscr{Y}(t)=\{\nu_{ij}(t)\geq 0,\forall i\in\mathscr{S}_a,\forall j\in\mathscr{S}_u\}$, $\Pi(t)=\{\pi_i(t)\geq 0,\forall i\in\mathscr{S}_a\}$ and $\Phi(t)=\{\phi_i(t)\geq 0,\forall i\in\mathscr{S}_a\}$. Note that $\mathscr{P}(t)^{(n)},\mathscr{V}(t)^{(n)},\Pi(t)^{(n)},\Phi(t)^{(n)},\mathscr{U}(t)^{(n)}$ are the parameters of problem (4.23) that are iteratively updated by the optimal solution achieved by solving (4.23) for current time slot t at each iteration. Once solving (4.23) for current time slot t, the solution is used to evaluate $Q_i(\bar{\mathscr{V}}(t-2),\bar{\mathscr{P}}(t-2))$ and $d_{ij}(\bar{\mathscr{P}}(t))$ and then fed into problem (4.21). Similarly, we can solve (4.21) for time slot t by iteratively solving the approximated convex problem (4.24) and update parameters $\mathscr{V}(t)^{(n)},\Pi(t)^{(n)},\Phi(t)^{(n)},\mathscr{U}(t)^{(n)}$ until convergence to attain the solution. Then, the solution of (4.21) can be used to evaluate $Q_i(\bar{\mathscr{V}}(t),\bar{\mathscr{P}}(t))$ and fed into problem (4.20) to solve at time slot $t+2$. Algorithm 5 presents the pseudo code that solves the online problems (4.20) and (4.21). By following the same argument for the convergence of the offline Algorithm 4, we can prove the convergence of Algorithm 5 from Steps 5–9 and Steps 13–17. Since Algorithm 5 operates in a finite number of time slots t_e, its convergence is guaranteed.

4.5.2 Delay-based Online Algorithms

Our proposed approach for Algorithm 5 contains a step of maximizing the WB rate at the small cell, so that at each time slot, the buffer of some small cells may be non-empty. The buffered data can be queued for a long period since the future wireless channel may not be sufficiently good for transmission. Consequently, delay at these small cell links may occur. To avoid such scenario, we propose an online delay-based resource allocation design by formulating a two-stage optimization problem as in Section 4.5.1. The first stage problem, denoted by $\left(\mathscr{P}^1_{\text{delay}}\right)$, is similar to problem (4.20), which aims at maximizing the small cell sum rate at time slot t. On the contrary, at the second stage, we aim at minimizing the cost while guaranteeing that all the backhaul rate must be greater than or equal to the access sum rate. The second stage

Algorithm 5: Online JBPAO Algorithm.

1: $t = 1$;
2: **repeat**
3: Set $n := 0$;
4: Initialize starting points of $\mathscr{P}(t)^{(n)}, \mathscr{V}(t)^{(n)}, \Pi(t)^{(n)}, \Phi(t)^{(n)}, \mathscr{U}(t)^{(n)}$;
5: **repeat**
6: Solve the convex problem in (4.23) to achieve the optimal solution $\mathscr{V}(t)^{\star}, \mathscr{P}(t)^{\star}, \mathscr{U}(t)^{\star}, \mathscr{W}(t)^{\star}, \mathscr{X}(t)^{\star}, \mathscr{Y}(t)^{\star}, \Pi(t)^{\star}, \Phi(t)^{\star}$;
7: Set $n := n+1$;
8: Update $\mathscr{P}(t)^{(n)} = \mathscr{P}(t)^{\star}, \mathscr{V}(t)^{(n)} = \mathscr{V}(t)^{\star}, \Pi(t)^{(n)} = \Pi(t)^{\star}, \Phi(t)^{(n)} = \Phi(t)^{\star}, \mathscr{U}(t)^{(n)} = \mathscr{U}(t)^{\star}$;
9: **until** Convergence of (4.23);
10: Compute $Q_i(\bar{\mathscr{V}}(t-2), \bar{\mathscr{P}}(t-2))$ and $d_{ij}(\bar{\mathscr{P}}(t)), \forall (i,j) \in (\mathscr{S}_a, \mathscr{S}_u)$ with the obtained $\mathscr{V}(t)^{\star}, \mathscr{P}(t)^{\star}$ and feed into (4.21);
11: Reuse the points $\mathscr{V}(t)^{(n)}, \Pi(t)^{(n)}, \Phi(t)^{(n)}, \mathscr{U}(t)^{(n)}$ from Step 9;
12: Set $n := 0$;
13: **repeat**
14: Solve the convex problem in (4.24) to achieve the optimal solution $\mathscr{V}(t)^{\star}, \mathscr{U}(t)^{\star}, \mathscr{W}(t)^{\star}, \Pi(t)^{\star}, \Phi(t)^{\star}$;;
15: Set $n := n+1$;
16: Update $\mathscr{V}(t)^{(n)} = \mathscr{V}(t)^{\star}, \Pi(t)^{(n)} = \Pi(t)^{\star}, \Phi(t)^{(n)} = \Phi(t)^{\star}, \mathscr{U}(t)^{(n)} = \mathscr{U}(t)^{\star}$;
17: **until** Convergence of (4.24a);
18: Compute $Q_i(\bar{\mathscr{V}}(t), \bar{\mathscr{P}}(t))$, $\forall i \in \mathscr{S}_a$ with the obtained $\mathscr{V}(t)^{\star}$ and feed into (4.20);
19: $t := t+2$;
20: **until** $t = t_e$;

problem is formulated as

$$\left(\mathscr{P}^2_{\text{delay}}\right) : \min_{\mathscr{V}(t)} \left\{ \sum_{k \in \mathscr{F}} \|\mathbf{v}_k(t)\|^2 \,|(4.21\text{b}), (4.21\text{c}), (4.7\text{d}) - (4.7\text{g}) \right\} \tag{4.25}$$

The intuition of problem $\left(\mathscr{P}^2_{\text{delay}}\right)$ is that we only allocate sufficient beamforming to the WB transmissions to ensure that the WB rate can support their corresponding WA rates, so that constraint (4.21b) occurs with equality at optimality (the proof can be sketched similar to Appendix 6). Since the backhaul rate is equal to the access rate, the SAP can use all the buffered data in the queue to transmit to its SUE. This means that at the end of each time slot, no data are stored in the queue, so that there will be no delay at all small cell links. We note that the result of Delay-based online problem is different from the Conventional problem in (4.7) in terms of

the achieved backhaul transmitted data. In Conventional scheme, the achieved backhaul transmitted data at each time slot might possibly be greater than the access transmitted data at each SAP, while the second stage of Delay-based online problem ensures that the resulted backhaul and access transmitted data to be equal. Thus, we can follow similar steps of transformation and approximation in Section 4.5.1 to solve problems $(\mathscr{P}^1_{\text{delay}})$ and $(\mathscr{P}^2_{\text{delay}})$.

In addition, we propose another design of resource allocation which stabilizes the small cell buffer in the wireless backhaul two-tier HetNets based on the Dynamic Back-Pressure (DBP) algorithm. In fact, when $C_{\max}$ is sufficiently large, maximizing the small cell data objective (like in (4.20)) without stabilizing the queue can cause excessive delay and instability at some SAPs. The work on buffer stability has been widely investigated in the literature and summarized in (Georgiadis *et al.*, 2006), where the effective DBP algorithm was developed to achieve a close-to-optimal solution of buffer usage. In our system, we can follow the same steps as in (Georgiadis *et al.*, 2006, Chapter 4) to formulate the DBP problem formulation and the subsequent DBP algorithm as follows

$$(\mathscr{P}_{\text{DBP}}): \max_{\mathscr{V}(t),\mathscr{P}(t)} \left\{ \sum_{i\in\mathscr{S}_a} Q_i(\bar{\mathscr{V}}(t-2),\bar{\mathscr{P}}(t-2)) \times \left(\sum_{j\in\mathscr{S}_u} r_{ij}(\mathscr{P}(t)) \right) \middle| (4.20\text{b}) - (4.20\text{d}) \right\} \quad (4.26)$$

where the objective function is the weighted small cell sum rate. We note that the weight at each link is proportional to the amount of buffered data in the queue from previous time slot. The intuition of this formulation is that we prioritize the maximization of the links which have longer queues so that more data rate of that link can be achieved and released from the buffer. Thus, DBP scheme can avoid the buffer overflow, maintain queue stability, and subsequently reduce the delay on each link.

Complexity analysis: We discuss the worst-case per-iteration computational complexity of offline algorithm. Note that online algorithms can be done similarly. We observe that problem (4.16) is a generalized convex programming due to the existence of the exponential cone in

constraints (4.14a) and (4.14b). For tractability of complexity analysis, we can employ the second order cone (SOC) approximation in (Nguyen *et al.*, 2016b, Section V, (27)) on each constraint (4.14a) and (4.14b) with an additional $m+4$ slack variables and $m+4$ SOC constraints to equivalently transform (4.16) into a SOC programming. Under this approximation, there are approximately $t_e(N(M+S_a)+4S_a+3S_aS_u+2S_aS_u(m+4))$ continuous variables in (4.16) with $(M+S_a+1)t_e$ constraints of dimension S_a+M+1, $t_eS_a(1+2S_u(m+4))$ SOC constraints of dimension 2 and $S_at_e(5+2S_u)$ SOC constraints of dimension 3. After omitting the small order, the worst-case per-iteration complexity of Algorithm 4 is $\mathcal{O}(t_e^4(N^2(M^4+S_a^4)+S_a^2S_u^2(m+4)^2(M^2+S_a^2))+t_e^3S_a^3S_u^3(m+4)^3)$.

4.5.3 Discussions

As discussed by the end of Section 4.3.1, the proposed Offline and Online solution in Section 4.3.2 can avoid the inherent issues from the conventional design, (c.f. optimization problem (4.7)) such as packet drop and retransmissions of backhaul data. Without retransmissions, the overall backhaul data is transmitted with more efficient power consumption in each time slot and can achieve a potential higher small cell rate than the conventional scheme. This implies an improvement on network energy efficiency, computed as the ratio of sum spectral efficiency over total network power consumption. Although research on energy efficiency under the usage of small cell buffer in wireless backhaul networks is an interesting topic, it is currently out of the scope of this paper and is left for future work.

4.6 Numerical Results

In this section, we evaluate the performance of the proposed algorithms. We assume that the CSI is the same in one time slot and varies independently in the next time slot. In addition, we consider Rayleigh fading channels and the pathloss component is calculated as $\left(d_{ij}/d_0\right)^{-3.8}$, where d_{ij} is the distance between the i^{th} transmitter and the j^{th} receiver, and $d_0 = 10$ m is the reference distance. We apply a circular coverage of macrocell with radius $20d_0$. The MBS is positioned at the center, where S_a small cells uniformly positioned in the macrocell coverage

(Nguyen *et al.*, 2016b). For ease of illustration, we choose $S_a = 4$ unless otherwise being mentioned. We assume that each small cell has a circular coverage of radius d_0 with its SAP at the center and one SUE is randomly located at the circumference of each small cell coverage. In addition, we assume $M = 2$ MUEs randomly scattered across the macrocell coverage. Unless being mentioned elsewhere, we choose this scenario as the standard model for the numerical simulation. The setting of other parameters is given in Table 4.1.

Table 4.1 Parameters setting

Description	**Value**
MBS antenna	$N = 4$
Maximum power at the SAP	$p_{\max} = 23$ dBm
Bandwidth	$W = 10$ MHz
Number of time slots	$t_e = 39$
MUE QoS requirement	$R_{\min} = 10^5$ bits/s
Maximum power at the MBS	$P_{\max} = 43$ dBm
Time slot period	$T = 20$ ms (Xiang *et al.*, 2017)
Spectrum division factor	$\alpha = 0.5$
Buffer storage size	$C_{\max} = 1$ Mbits
Noise power	$W\sigma_0 = -120$ dB

Fig. 4.3 and 4.4 show the convergence of the total amount of data received at all the SUEs by applying Algorithm 4 in the Offline Scheme and Algorithm 5 in the Online Sum Scheme, where $U(\Gamma(\mathscr{V}(t)) = \sum_{i \in \mathscr{S}_a} \log(1 + \Gamma_i(\mathscr{V}(t)))$. In addition, the performance of the proposed algorithms are also benchmarked by the global optimal solution achieved by applying the Branch and Bound (BnB) algorithm developed in (Tervo *et al.*, 2015b) on the offline problem (2.7). In Fig. 4.3 and 4.4, we choose $t_e = 2, 4$. Based on (Tervo *et al.*, 2015b), the BnB algorithm recursively branches the considered potential set which contains the global optimal solution into smaller sets, checking the feasibility of the newly branched sets and updating the stopping criteria of the BnB algorithm until convergence. The stopping criteria used in the BnB algorithm is the difference of upper and lower bound functions, which are computed from the sets containing the potential global optimal solution. From Fig. 4.3, we observe that Algorithm 4 in the Offline Scheme and Algorithm 5 in the Online Sum Scheme keep increasing and converge

after a few iterations, which validates the convergence analysis. Moreover, the obtained results from low-complexity algorithms are very close to the baseline optimal value obtained from the BnB algorithm. This again shows the efficiency of our low-complexity offline and online algorithms in solving the non-convex NP-hard problems with much less computational cost, while still maintaining a close-to-optimal performance.

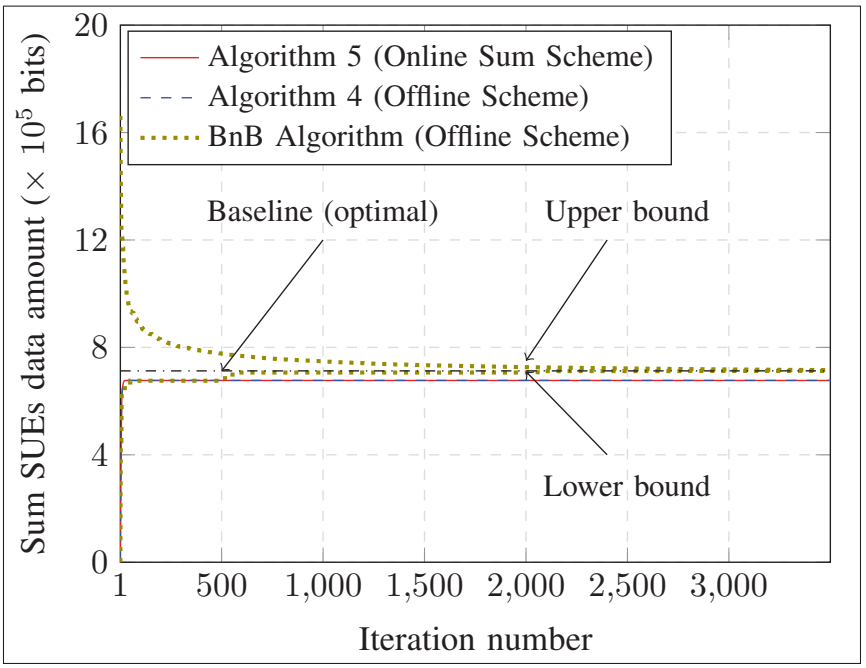

Figure 4.3 Convergence between the optimal BnB algorithm, Offline Scheme, and Online Sum Scheme when $t_e = 2$ and $S_a = 4$.

Fig. 4.5 compares the total amount of data received at the SUEs between the Offline, Online, DBP (c.f. (4.26)), Conventional Scheme ((Wang *et al.*, 2016; Zhao *et al.*, 2015; Nguyen *et al.*, 2016c)), Schemes in (Peng *et al.*, 2016) and (Guo *et al.*, 2017) with respect to the SAP power budget p_m. The online approaches consider the Online Sum Scheme and the Online Max-Min Fairness Scheme where $U(\Gamma(\mathscr{V}(t)) = \min_{i\in\mathscr{S}_a}\Gamma_i(\mathscr{V}(t))$. We observe from Fig. 4.5 that the Offline and Online Schemes always outperform the Conventional Scheme. At high p_m, the Offline, Online Sum and Max-Min Fairness Scheme achieve the gains of 93.67% , 93.42% and 66.86% compared to the Conventional Scheme, respectively, which again validates the superi-

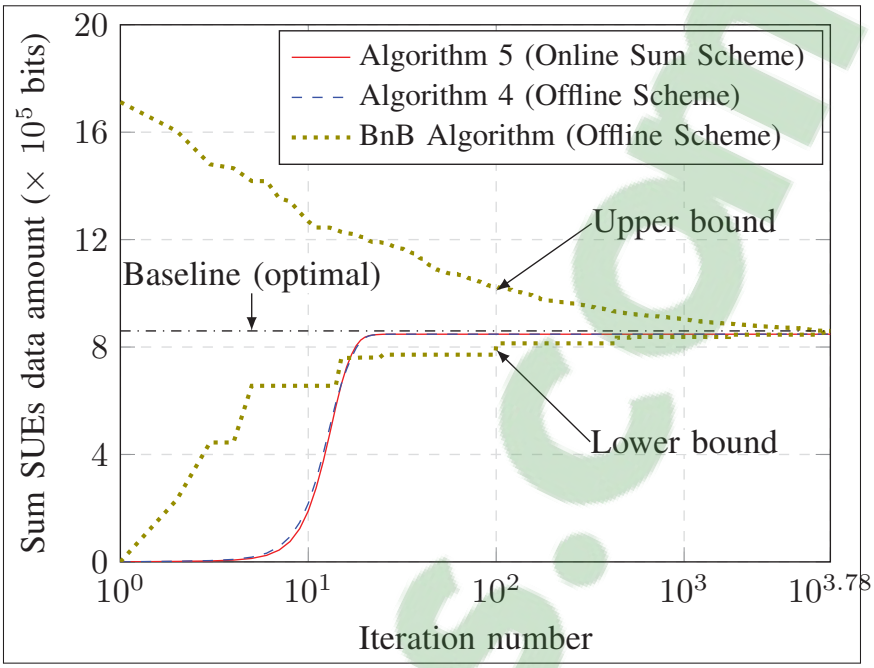

Figure 4.4 Convergence between the optimal BnB algorithm, Offline Scheme, and Online Sum Scheme when $t_e = 2$ and $S_a = 4$.

ority of our proposed online approach compared to the conventional design in the literature as stated in Propositions 3 and 4. The DBP Scheme and Schemes in (Peng *et al.*, 2016) and (Guo *et al.*, 2017) perform lower than the Offline and Online Sum Scheme. This can be explained since in DBP Scheme, we only aim at maximizing the WA weighted sum rate rather than WA sum rate. Alternatively, the schemes in (Peng *et al.*, 2016) and (Guo *et al.*, 2017) respectively aim at maximizing the energy efficiency and the total admission rate (similar to the total wireless backhaul rate in our case) subject to the stability of the buffer queue rather than the WA sum rate maximization. Therefore, the performance of DBP Scheme and the Scheme in (Peng *et al.*, 2016) and (Guo *et al.*, 2017) are less than the performance of Offline and Online Sum Schemes. Here, we note that the gap between the Offline and the Online Sum Scheme is close when p_m grows. In addition, when p_m increases, the total amount of data of all the schemes increases and saturates at higher value of p_m. This is because at high p_m, each SAP stops increasing its power to avoid creating more interference to its neighboring small cell, which leads to the saturation in the amount of data transmitted to the SUEs in all the considered schemes.

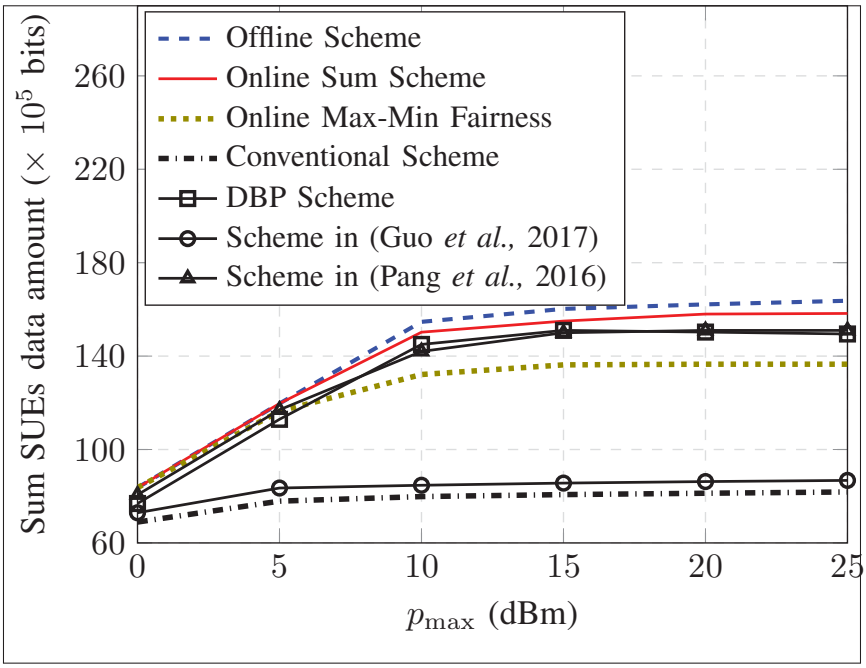

Figure 4.5 Comparison of the achieved amount of data transmitted to the SUEs between offline and online algorithms.

Fig. 4.6 shows the total amount of data that is transmitted to the SUEs with respect to the buffer size $C_{\max}$ at two values of $P_m = 33, 43$ dBm. We observe that the amount of access transmitted data increases with $C_{\max}$. To explain this, note that at low $C_{\max}$, the SAP cannot attain a WB rate higher than the buffer capacity to buffer the received backhaul data and the amount of data transmitted to the SUEs is limited. When $C_{\max}$ grows, Online Sum and Online Max-Min Fairness Schemes outperform Conventional Scheme since more data can be buffered for later transmissions, as being analyzed in Section 4.2.4. When $C_{\max}$ is sufficiently large, this amount of data converges at a saturation value since the SAPs cannot use all the buffer storage to store their data. Note that the performance of all the schemes is higher at higher values of P_m. This is obvious since with higher power budget P_m, the MBS can transmit higher WB rate to the SAPs and the SAPs can subsequently improve their WA rates. Fig. 4.7 compares the total delay of all the small cell buffer with respect to $C_{\max}$ between the Online Sum, DBP, and Delay-based schemes (c.f. (4.25)). Here, we are inspired by the formula of time-average delay over infinite time domain (Guo *et al.*, 2017) to compute the delay at each small cell buffer link i over t_e time

slots as

$$\text{Delay}_i = \frac{1/t_e \sum_{t=1}^{t_e} Q_i(\bar{\mathscr{V}}(t), \bar{\mathscr{P}}(t))}{1/t_e \sum_{t=1}^{t_e} \alpha W R_i(\bar{\mathscr{V}}(t))}$$

When $C_{\max}$ increases, we observe that the Delay-based scheme always achieve the zero delay. This is because it allocates the backhaul rate that is equal to the access rate at each small cell link, so that no data is left in the buffer at the end of each time slot. On the other hand, the proposed Online Sum scheme always achieves non-zero delay below 2.5×10^{-2} seconds. This is because it stores data in the buffer of some the small cell, but in the next time slots, these data still remain inside the buffer. When $C_{\max}$ increases, due to the limited power budget and the interference from other links, the backhaul rate is saturated (as shown in Fig. 4.6) so that the data in the buffer do not increase much. Therefore, the computed delay under the Online Sum scheme does not vary much at high $C_{\max}$. We also observe that the DBP scheme outperforms the Online Sum scheme, but still achieves non-zero delay. This is because the DBP algorithm prioritizes the maximization of the small cell access rates with the longer queues, so that it reduces the amount of data and subsequently the delay in those buffers.

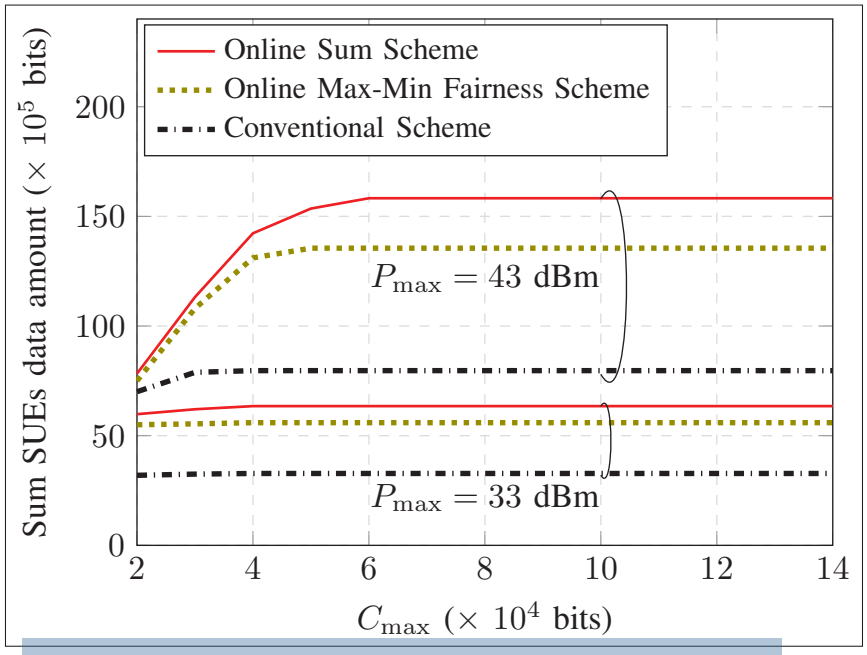

Figure 4.6 Achieved amount of data transmitted to the SUEs with respect to $C_{\max}$.

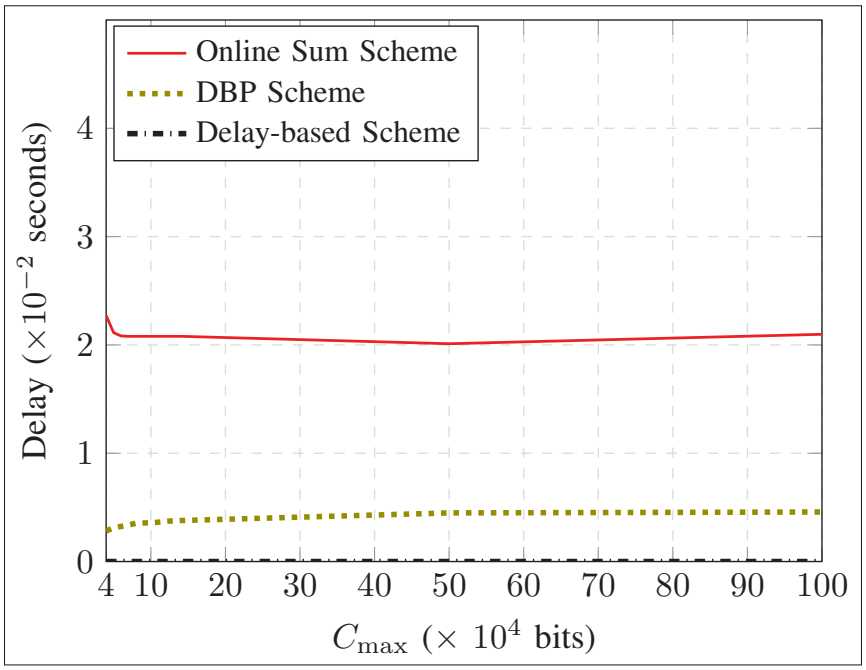

Figure 4.7 Delay with respect to $C_{\max}$.

In Fig. 4.8, we show the performance of the Online Sum Scheme with respect to the spectrum splitting factor α at $P_m = 33, 43$ dBm. We observe that when α varies, the total small cell data first increases and then decreases. There exists a value of α which maximizes the total access transmitted data, where this maximal value is different for each P_m. This is because when α is too low, more bandwidth partitioning is dedicated for the transmissions from MBS to SAPs and MUEs. This leads to a low performance of small cell sum rate since less bandwidth is provided for the transmissions from the SAPs. When α grows, more bandwidth is drawn for the transmissions from SAPs. However, since less bandwidth is available for the backhaul transmissions, this subsequently limits the upper bound performance of each SUE rate, and restricts the total small cell transmitted data.

In Fig. 4.9a and Fig. 4.9b, we show the evolution in time of the amount of buffered data. In particular, Fig. 4.9a shows the data stored in the buffer right after the MBS transmits data to the SAP 1 in the Online Sum and Conventional Scheme. As shown in the figure, the data stored in the buffer at SAP 1 in the Online Sum Scheme is always greater than in the Conventional

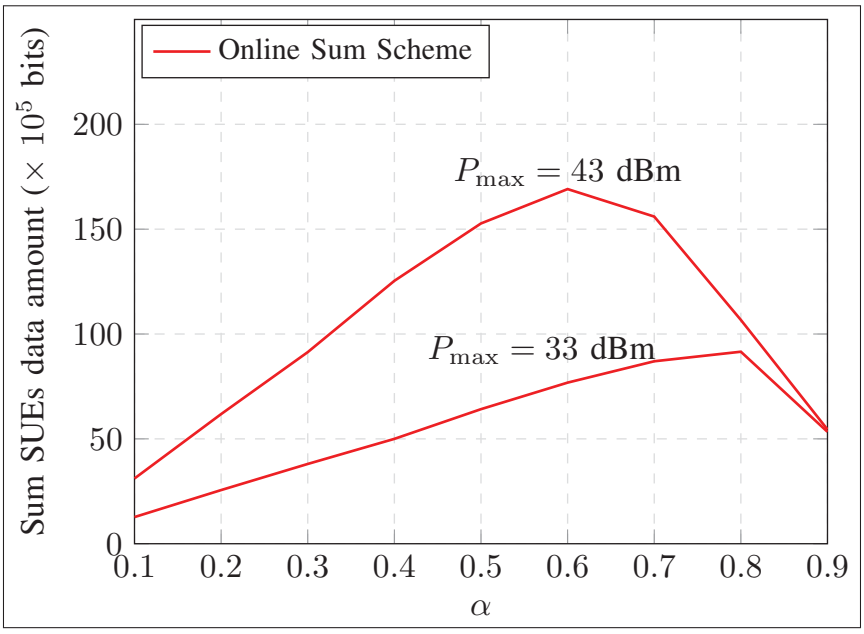

Figure 4.8 Achieved amount of data transmitted to the SUEs versus α.

Scheme. This can be explained as in the Conventional Scheme, the MBS only transmits at a sufficient rate to support the rate requested by the SAP. However, the Online Sum Scheme aims at maximizing the WB rate beside the WA rate, and thus more WB data can be achieved and stored in the buffer compared to the Conventional Scheme. In addition, Fig. 4.9b examines the behavior of the data in the buffer for the Online Sum Scheme at SAP 1 and SAP 2 after the small cells transmit to their intended users. Again, we observe that some data remain in the buffer of each small cell in some time slots. This data can be transmitted in the next time slots to improve the upper-bounded backhaul rate corresponding to each small cell and thus improve the small cell access sum rate.

Fig. 4.10 shows the amount of data that is transmitted to the SUE with respect to the number of small cells at $P_m = 33$ or 43 dBm. In addition, we compare the performance between the Online Sum, Online Max-Min Fairness and Conventional Scheme. We observe that the Online Sum and Online Max-Min Fairness Scheme always perform better than the Conventional Scheme at all values of number of small cell, which is consistent with the observations from the previous

figures. When P_m is small, the achieved amount of data transmitted to the SUE is low. This can be explained as when the MBS has a smaller power budget to transmit wireless backhaul data to the SAP, the WA rate at each SAP will be upper bounded by a smaller valued WB rate. Thus, the optimal sum WA rate only achieves a low value at low $P_{\max}$.

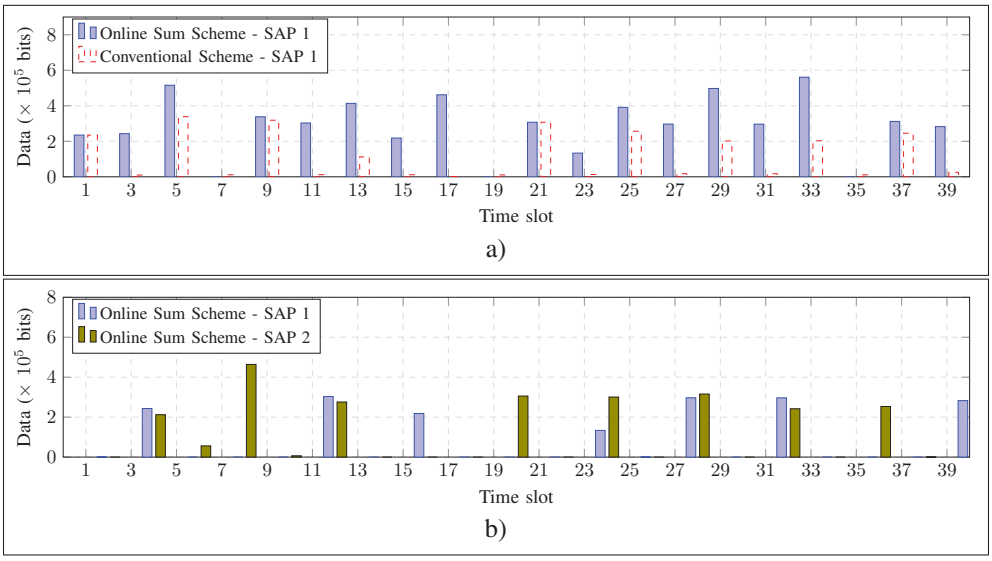

Figure 4.9 Buffered data in the queue of: (a) SAP 1 after the MBS transmits; (b) of SAP 1 and 2 after the SAPs transmit.

4.7 Concluding Remarks

This paper studied a novel model of two-tier wireless backhaul small cell HetNets where each SAP is equipped with a finite buffer. By proposing the time–spectrum downlink transmission allocation to manage interference, we developed a joint optimized transmit beamforming and power allocation algorithm. By exploiting the advantage of the buffering capability at the SAPs, we investigate the performance gain of the proposed network via solving a constrained optimization problem of maximizing small cell sum rate. To solve the formulated problem, we proposed an ideal offline and a practical online algorithm based on different assumption of CSI knowledge. Both algorithms rely on the framework of convex approximation to be developed to attain the solution within polynomial time, whereas the offline algorithm only serves for

benchmarking purpose to evaluate the performance of the online algorithms. Our theoretical results prove that our proposed wireless backhaul with buffer-aided small cells always performs better than conventional work. Numerical results validate our derived theory and confirms that our proposed model with advanced buffering strategy outperforms the conventional design in terms of small cell access rate for both the offline and online algorithms.

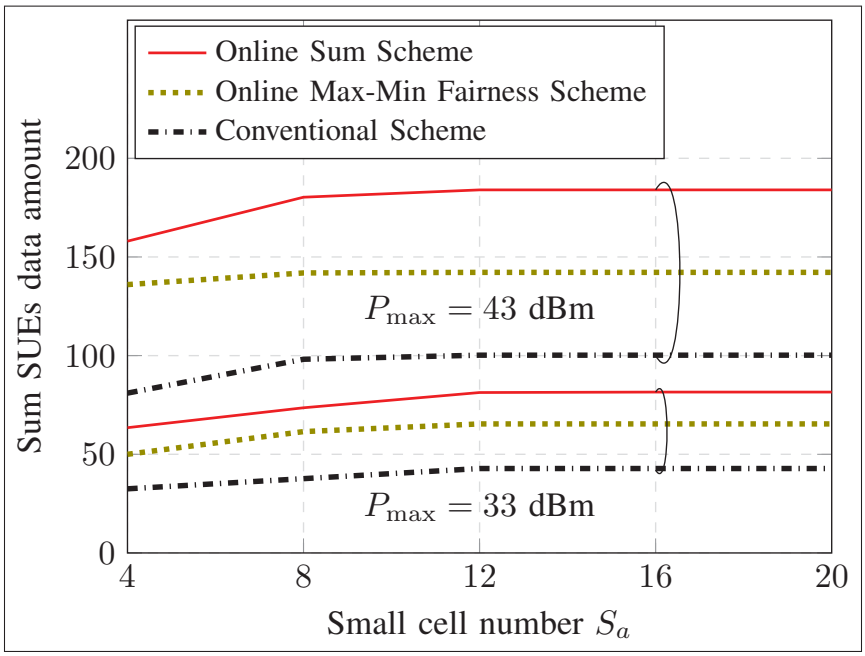

Figure 4.10 Achieved amount of data transmitted to the SUEs with respect to number of small cell.

CHAPTER 5

A NOVEL COOPERATIVE NON-ORTHOGONAL MULTIPLE ACCESS (NOMA) IN WIRELESS BACKHAUL TWO-TIER HETNETS

Tri Minh Nguyen[1], Wessam Ajib[2], Chadi Assi[3]

[1] Département de Génie Électrique, École de Technologie Supérieure,
[2] Département d'Informatique, Université de Québec à Montréal
[3] Concordia Institute for Information Systems Engineering, Concordia University

This article was published in *IEEE Transaction on Wireless Communications* in May 2018 (Nguyen *et al.*, 2018b).

5.1 Introduction

In the last recent years, mobile wireless networks have been witnessing a major upgrade towards their Fifth Generation (5G). The anticipated strategy which fulfills 5G goals by enhancing system spectral and energy efficiency to cope with the thousand-fold increase in data traffic is unprecedentedly expected (Andrews *et al.*, 2014; Kela *et al.*). Many key technologies are proposed to reach this expectation (Wong *et al.*, 2017). Among these, dense (or ultra-dense) heterogeneous networks (HetNets) solution is shown to suit the 5G perspective in many studies (Bhushan *et al.*, 2014). Although they are famous for self-deploying small cells, dense HetNets must inevitably deal with the indispensable installation problem by connecting excessive fiber backhaul links from the small cells (SCs) to the core network (Ge *et al.*, 2014). Thus, a more appealing solution, instead of wiring each single SC access point (SAP)–core network connection, is to transmit the backhaul data via wireless channels. Therefore, it can remarkably reduce the cost of network deployment and maintenance (Siddique *et al.*, 2015a).

As proposed in (Ghosh *et al.*, 2012), wireless backhaul (WB) in HetNets enables simultaneous wireless transmissions of backhaul signals between the SAPs and macro base station (MBS). Although WB offsets the economical and deployment issues, it in turn creates two fundamental bottle-necks which restrain the overall achievable rate performance (Zhao *et al.*, 2015; Wang

et al., 2016; Nguyen *et al.*, 2016b,a, 2017b). First, deploying WB communications concurrently with the existing wireless access (WA) communications renders the interference more severe. Second, it is apparent that each SAP should rely on the backhaul data to transmit to its intended users (Siddique *et al.*, 2015a). Therefore, it is essential to impose that the WB rate achieved at a SAP exceeds the WA rate. Hence, designing the WB architecture in HetNets to harness all of its potential advantages must explicitly consider these two drawbacks to operate a sustainable system. However, this introduces more challenges for the resource allocation and efficient methods to serve the above design criteria are imperative to yield the optimal solution.

5.1.1 Related Work

Employment and standardization of WB were initially proposed for IEEE 802.16 mesh networks (Viswanathan & Mukherjee, 2006). By enabling WB transmission, the authors in (Viswanathan & Mukherjee, 2006) developed an algorithm based on linear program to find the optimal routing and scheduling strategy so as to attain the maximal network throughput. In the context of 5G, WB was more likely readdressed in different network scenarios and, interestingly, looked at a wider variety of spectrum bands (Siddique *et al.*, 2015a). Particularly, Hur *et al.* (2013b) presented a novel idea for SCs equipped with mm-wave transmitters to effectively align their transmit beam under wind induced impairments. In the sub–6 GHz band, some works proposed to utilize the available hardware and current implementation to accommodate WB communications concurrently with the WA ones. By employing the reverse time division duplex model, the authors in (Sanguinetti *et al.*, 2015; Wang *et al.*, 2016) considered the joint bandwidth allocation and user association which maximizes the achievable downlink sum log rate of SCs under the massive multiple input multiple output (MIMO) MBS. In (Zhao *et al.*, 2015), the authors aimed at finding the maximum admitted WB-enabled SAP to serve as many SAPs as possible while guaranteeing quality-of-service (QoS) rates. To better exploit the available spectrum, Siddique *et al.* (2017) introduced a hybrid strategy of out-of-band and in-band full-duplex and formulated a problem which optimizes the spectrum allocation under the proposed scheme. On the other hand, the authors of (Niu *et al.*, 2017) presented an

optimization problem of scheduling and power control which maximizes the mm-wave WB energy efficiency. Unlike the mentioned works, the authors of (Yang *et al.*, 2016a) studied the WB networks and analyzed the energy efficiency of the SCs using stochastic geometry.

In the aforementioned work, in-band and out-of-band WB within sub–6 GHz band are more frequently studied due to their technological maturities. To facilitate sophisticated task in managing WB with other coexisting communications, accommodation of WB and WA communications on orthogonal frequency is more likely considered. However, this not only reduces the spectral reuse capability but also irradically handles severe interference under ultra-dense networks. Hence, more advanced and pragmatic technologies without overusing such orthogonal spectrum partition manner are always vital. In such scenario, non-orthogonal multiple access (NOMA) (Saito *et al.*, 2013) emerges as a promising candidate to improve both WB rate and its respective WA rate. Unlike conventional orthogonal multiple access strategies, NOMA allows many simultaneous user's transmissions and receptions to access the same sub-carrier. To handle the inter-user interference, multi-user detection technique, such as successive interference cancellation (SIC), is applied at each receiver. A general survey of NOMA techniques was given in (Islam *et al.*, 2017) while the particular application on downlink NOMA for 5G networks was surveyed in (Wei *et al.*, 2016). In general, advanced by the characteristics such as multiplexing transmitting technique, SIC receiving technique, and efficient network resource allocation, NOMA has been able to show its prominent superiority through many research work in the related field. Ding *et al.* (2014) investigated the network performance of NOMA in a random user deployment scenario. In (Shi *et al.*, 2016), the authors jointly solved the power allocation and decoding order for an outage balancing downlink NOMA system. Cui *et al.* (2018) studied the NOMA design in the MISO mmwave system. However, the authors employed a fixed precoding beamformer prior to the design of NOMA so that the problem of interest is reduced to the joint design of power allocation together decoding order. Unlike these work, (Al-Imari *et al.*, 2014) applied an uplink NOMA scheme and formulated a problem of joint power and sub-carrier allocation optimization which maximizes the total network achievable rate. A cooperative NOMA scheme is investigated in (Ding *et al.*, 2015). Inspired by (Ding

et al., 2015), Wei *et al.* (2017a) proposed a more advanced hybrid downlink-uplink cooperative NOMA scheme to achieve a better trade-off between spectral efficiency and signal reception. Interestingly, the concept of NOMA was recently applied to visible light communication systems in (Marshoud *et al.*, 2016). When multiple antenna are available at the transmitter or/and receiver, studies on multiple input single output (MISO) or MIMO NOMA are necessary so as to better harness all the potential degrees of freedom for maximizing the performance enhancement. In (Choi, 2015) and (Hanif *et al.*, 2016), the authors formulated a problem of designing beamforming in the MISO NOMA to minimize the total consumed power or maximize the sum spectral efficiency, respectively. In a particular case of two-antenna-user, Sun *et al.* (2015a) formulated a throughput maximization problem and developed two algorithms to find solutions. In (Ding *et al.*, 2016), a novel framework of downlink and uplink MIMO NOMA was inspired by the application of signal alignment. The authors invoked the stochastic geometry tool to derive a closed-form evaluation formula, in which two power allocation schemes were studied to reflect the impact of the proposed scheme. Furthermore, Wei *et al.* (2017b) and Sun *et al.* (2017a) respectively studied the power-efficient and system throughput resource allocation optimization for multicarrier (MC)-NOMA system and full duplex MC-NOMA. In (Sun *et al.*, 2017a), the authors developed a polyblock optimal algorithm based on the monotonic branch-and-bound algorithm and low complexity successive convex approximation algorithm to solve the formulated problem optimally and sub-optimally. Similarly, Wei *et al.* (2017b) also adopted the monotonic branch-and-bound algorithm and developed a low complexity algorithm based on difference of convex (DC) programming to solve for global optimal and sub-optimal solutions.

5.1.2 Motivation and Contribution

According to the above discussed works, improving the macrocell WB rate while simultaneously increasing the small cell WA rate is imperative. To the best of our knowledge, the existing literature lacks of work which jointly enhances the WB and WA downlink transmissions prior to the maximization of network performance. Being motivated by these observations, this pa-

per proposes a novel cooperative transmission scheme for the WB and WA communications based on NOMA strategy. In particular, we study the downlink two-tier WB small cell HetNets which comprises of one macrocell coexisting with multiple small cells. To efficiently and tractably manage cross-tier and self-interference from the new WB transmission type, an interference management scheme based on cochannel time division duplexing (CoTDD) (Hoydis *et al.*, 2013) is considered. Under this CoTDD setting, we contribute the following points in our work:

- Our first important contribution lies in the advanced usage of the MISO NOMA on the macrocell downlink. Rather than fixing the decoding order in the MISO NOMA process prior to the optimization of the resource allocation as in (Hanif *et al.*, 2016) or solving for a decoding order in the SISO NOMA homogeneous networks as in (Shi *et al.*, 2016), our work proposes to present the decoding order between the SAPs as a set of binary variables and propose to jointly determine these binary variables together with the radio resource allocation which optimizes the objective of interest.
- In our second and most significant contribution, we observe that some small cells can opportunistically decode the backhaul data which is intended for other small cells thanks to the proposed NOMA SIC process. Based on this observation, we present a novel cooperative transmission scheme for the small cell downlink which allows different SAPs to cooperatively transmit their decoded messages to the small cell users (SUEs). This provides additional degrees of freedom for small cell access transmissions, which can potentially improve their achievable rates. However, appropriately controlling the SAPs' cooperation must comply with the decoding rule from the NOMA process, which is undetermined initially. Therefore, our alternative contribution is to formulate the proper WA rate formula taking into account the proposed cooperation. Note that this formula must cope with the the decoding order rule represented by the binary variables.
- Under the novel scheme, we formulate two problems of jointly determining the decoding order together with allocating the transmit beamforming and power at the MBS and SAPs which optimizes the problems of interest. The first problem aims at maximizing the total

macrocell user (MUE) and SUE sum rate. The second problem aims at maximizing the number of MUEs and SUEs which can satisfy the minimum rate requirements.

- We observe that both formulated problems are mixed integer non-convex and are generally NP-hard. Finding a global optimal solution is very difficult and cannot be done using any existing off-the-shelf polynomial-time algorithm. Thus, we propose to approach the solution via a more efficient and pragmatic method. In particular, we first employ the DC functions to present the formulated binary variables and then equivalently transform the optimization problems into more tractable forms. Finally, we present an iterative low-complexity algorithm based on successive convex approximation (SCA) technique and majorization minimization method (MMM) to attain the solution, which is provable to eventually converge at a sub-optimal solution.

The rest of this chapter is organized as follows. Section 5.2 introduces the system model and proposes the novel transmission scheme. In Section 5.3, we formulate two optimization problems and analyze their characteristics. In Section 5.4, we develop a unified algorithm to solve both formulated problems. Section 5.5 presents and discusses our numerical results. Finally, the concluding remarks of this chapter are given in Section 5.6.

5.2 System Model

5.2.1 Spatial Model and Interference Management

We consider the downlink of a two-tier WB HetNet consisting of one MBS in the macrocell tier and S SAPs in the small cell tier as shown in Fig. 4.1. The MBS is equipped with $N > 1$ antennas to communicate with the SAPs and its MUEs. In this work, we assume the WB transmission between the MBS and SAP operates concurrently with other WB transmissions and with the WA transmissions from the MBS to its MUEs on the same spectrum. For tractability, we consider that each SAP serves one SUE within its coverage, where all SAPs, SUEs, and MUEs are equipped with a single antenna each. We also consider the design of this system under perfect and imperfect channel state information in this work. In a more complicated

case, the SAPs can be equipped with multiple antenna to enable multiple-stream transmissions from the MBS and allow each SAP to simultaneously serve many SUEs within its coverage. The resource allocation must then deal with the design of transmit and receive beamforming for the MBS–SAPs transmission to suit with the MIMO configuration and satisfy the inherent characteristics of the WB HetNet. Such scenario is a non-trivial extension of our current work, which is especially complicated when considering the NOMA concept in the subsequent sections, and thus deserves further study in future works.

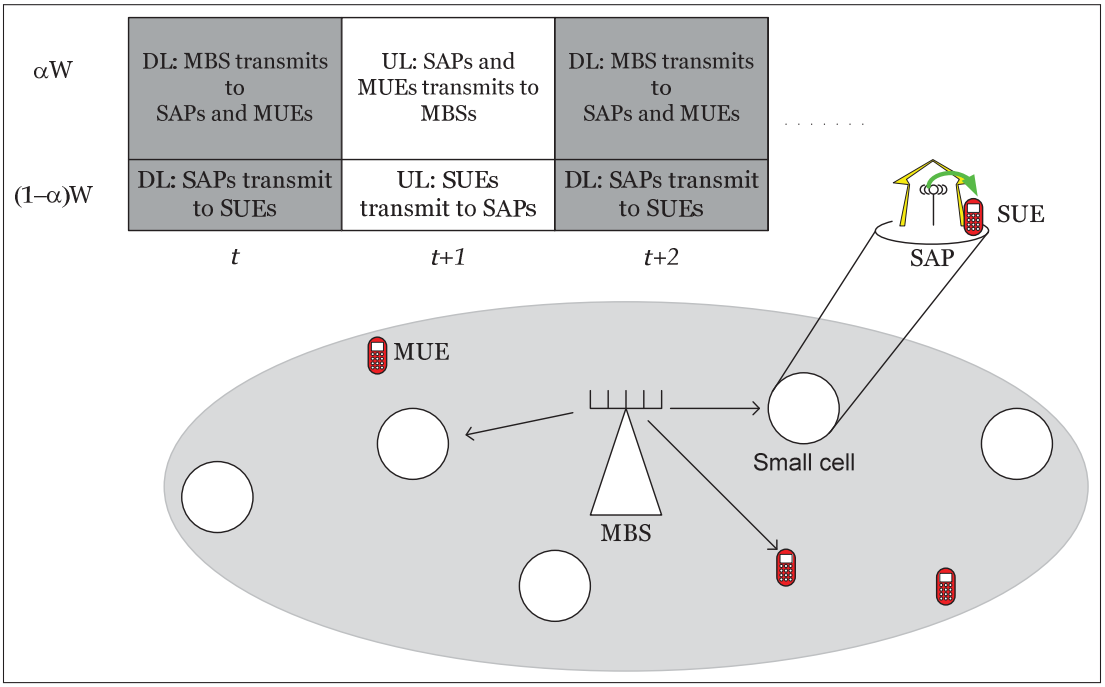

Figure 5.1 Two-tier WB HetNets with CoTDD scheme.

To facilitate the radio resource allocation design presented later for the half-duplex transmitters and receivers, we employ the CoTDD scheme as in (Sanguinetti *et al.*, 2015; Hoydis *et al.*, 2013) to accommodate the WB and WA communications on the available channel resources. More specifically, the time domain is divided into multiple time slots to separate the downlink and uplink transmissions. The spectrum band W is split by a factor of α. In each time slot dedicated for the downlink transmission, αW Hertz is reserved to accommodate the backhaul downlink transmissions while $(1-\alpha)W$ Hertz is reserved for the access downlink transmis-

sions. In another time slot dedicated for the uplink transmission, αW Hertz is reserved to accommodate the backhaul uplink transmissions while $(1-\alpha)W$ Hertz is reserved for the access uplink transmissions. The proposed time–spectrum allocation is illustrated in Fig. 5.1.

5.2.2 Transmission Model

5.2.2.1 Macrocell Downlink Transmissions

In this work, we assume that the channel is time invariant for the considered time slots. We first consider the resource blocks dedicated for the macrocell downlink transmissions. We assume that this resource block has T seconds and αW Hertz. We denote the SAP and MUE indices set are $\mathscr{S}$ and $\mathscr{M}$, where $\mathscr{F} = \{\mathscr{S}, \mathscr{M}\}$. Let $\mathbf{w}_j \in \mathbb{C}^{N\times 1}$ be the beamformer from the MBS to the j^{th} receiver (including the SAPs and MUEs), s_j be the message intended for the j^{th} receiver with unit average power, e.g., $\mathbb{E}\left\{s_j s_j^*\right\} = 1$. Then, the received signal at the j^{th} receiver is

$$y_j = \mathbf{h}_j^H \mathbf{w}_j s_j + \sum_{k\in\mathscr{F}\setminus j} \mathbf{h}_j^H \mathbf{w}_k s_k + n_j, \tag{5.1}$$

where $\mathbf{h}_j \in \mathbb{C}^{N\times 1}$ is the channel state vector from the MBS to the j^{th} receiver, which includes fading gain and path-loss components. Note that the path-loss is calculated according to the distance and carrier frequency dependent model (Goldsmith, 2005) as

$$\text{PL}_j\,(\text{dB}) = -10\log_{10}\frac{G_l(c/f_c)^2}{(4\pi d_j/d_0)^2}, \tag{5.2}$$

where d_j is the distance from the MBS to the j^{th} receiver, and d_0 is the reference distance. In addition, G_l is the non-directional antenna gain, c is the speed of light, f_c is the carrier frequency. In addition, n_j is an additive white Gaussian noise (AWGN) at the j^{th} receiver, which is distributed according to a normal distribution $\mathscr{CN}(0, N_{01})$, where N_{01} is the noise power over the allocated spectrum.

At the receiver side, each SAP applies the SIC technique used for the NOMA downlink (Islam *et al.*, 2017) prior to decoding its intended message. Here, we do not consider the SIC protocol among the MUEs. Depending on the determined decoding order, the j^{th} SAP might decode some messages intended for the other i^{th} SAP, for some $i \in \mathscr{S}$. It is worth to mention that in single-input single-output (SISO) downlink NOMA system (Saito *et al.*, 2013), where the receivers indices follow the ascending order of the channel quality, the j^{th} receiver can decode all the messages of the i^{th} receiver, for $i \leq j$. However, such ordering mechanism cannot be straightforwardly applied in MISO systems since computing channel quality in this case differs from SISO environment. This is because with multiple transmit antenna, we must deal with channel and beamforming vectors rather than scalar channels and powers in single antenna case. To convey the determination of an appropriate decoding order, let us introduce binary variable $a_{\ell,j} \in \{0,1\}$, where $a_{\ell,j} = 0$ implies that the j^{th} SAP can decode message s_ℓ intended for the ℓ^{th} SAP and subtract the interfered signal caused by message s_ℓ out of the total received signal; and $a_{\ell,j} = 1$ otherwise. Therefore, the achievable rates, computed in bps/Hz, after applying SIC–NOMA for the downlink transmission at the j^{th} receiver to detect the ℓ^{th} SAP's message and its own j^{th} messages ar

$$R_{j,\ell}(\mathbf{w},\mathbf{a}) = \alpha \log\left(1 + \frac{\left|\mathbf{w}_\ell^H \mathbf{h}_j\right|^2}{\sum\limits_{l \in \mathscr{S}\setminus j} a_{l,j}\left|\mathbf{w}_l^H \mathbf{h}_j\right|^2 + \sum\limits_{m \in \mathscr{M}}\left|\mathbf{w}_m^H \mathbf{h}_j\right|^2 + N_{01}}\right) \tag{5.3}$$

$$R_{j,j}(\mathbf{w},\mathbf{a}) = \alpha \log\left(1 + \frac{\left|\mathbf{w}_j^H \mathbf{h}_j\right|^2}{\sum\limits_{l \in \mathscr{S}\setminus j} a_{l,j}\left|\mathbf{w}_l^H \mathbf{h}_j\right|^2 + \sum\limits_{m \in \mathscr{M}}\left|\mathbf{w}_m^H \mathbf{h}_j\right|^2 + N_{01}}\right) \tag{5.4}$$

$$R_{k,k}(\mathbf{w}) = \alpha \log\left(1 + \frac{\left|\mathbf{w}_k^H \mathbf{h}_k\right|^2}{\sum\limits_{l \in \mathscr{F}\setminus k}\left|\mathbf{w}_l^H \mathbf{h}_k\right|^2 + N_{01}}\right), \tag{5.5}$$

where $j, \ell \in \mathscr{S}, \ell \neq j$ and $k \in \mathscr{M}$. In addition, we denote $\mathbf{w} = \left\{\mathbf{w}_j \in \mathbb{C}^N, \forall j \in \mathscr{F}\right\}$ and $\mathbf{a} = \left\{a_{\ell,j}, \forall \ell, j \in \mathscr{S}\right\}$. In addition, $N_{01} = \alpha W \sigma_0$ is the noise power over the αW band and σ_0 is the noise power spectral density. The characteristic of $\mathbf{a}$ deserves more elaboration. First, since

each j^{th} SAP can decode its own message s_j, it holds that $a_{j,j} = 0, \forall j \in \mathscr{S}$. Second, supposed that a decoding order is determined so that the j^{th} SAP is ranked before the ℓ^{th} SAP. According to the SIC principle, if the j^{th} SAP decodes the ℓ^{th} SAP's message s_ℓ, the ℓ^{th} SAP cannot decode the j^{th} SAP's message s_j or reversely. This implies that either $a_{\ell,j} = 0$ and $a_{j,\ell} = 1$ or $a_{\ell,j} = 1$ and $a_{j,\ell} = 0$. In general, we have $a_{j,\ell} + a_{\ell,j} = 1, \forall j, \ell \in \mathscr{S}$. Third, if the j^{th} SAP cannot decode the i^{th} SAP's message s_i, and the k^{th} SAP cannot decode the j^{th} SAP's message s_j, then the k^{th} SAP cannot decode the i^{th} SAP's message s_i. Mathematically, we can write the third relationship as $a_{i,k} \geq a_{i,j} a_{j,k} \forall i, j, k \in \mathscr{S}$. Finally, supposed that the j^{th} SAP can decode message s_ℓ intended for the ℓ^{th} SAP, $\forall \ell \neq j$, the following constraints, which are also reflected in (Sun *et al.*, 2017b, Eq. (4)), must hold

$$R_{j,\ell}(\mathbf{w}, \mathbf{a}) \geq \left(1 - a_{\ell,j}\right) R_{\ell,\ell}(\mathbf{w}, \mathbf{a}), \forall \ell, j \in \mathscr{S}, \ell \neq j. \tag{5.6}$$

5.2.2.2 Novel Cooperative Small Cell Downlink Transmissions

In the other resource block of T seconds and $(1-\alpha)W$ Hertz where the SAPs transmit to their SUEs, we propose a novel cooperative transmission scheme which directly inherits from the SIC detailed above. The details of this transmission can be elaborated by analyzing the following example with $S = 3$ as illustrated in Fig. 5.2.

An example case for $S = 3$: In this example, we assume that SAP 1 is ranked after SAP 2, and SAP 2 is ranked after SAP 3. As a consequence of the reception in the macrocell downlink transmission, SAP 1 can only decode data for its SUE in SC 1. SAP 2 decodes data not only for its SUE in SC 2 but also for the SUE in SC 1. Finally, SAP 3 decodes all the data for the SUEs in SCs 1,2, and 3. Since SAP 2 and 3 have the data of the SUE in SC 1, they can cooperate with SAP 1 to send data s_1 to the SUE of SC 1. Similarly, since SAP 3 also has the data of the SUE in SC 2, it can cooperate with SAP 2 to send data s_2 to the SUE in SC 2. Here, the cooperation between SAP 1, 2, and 3 is that SAP 2 and 3 can simultaneously transmit

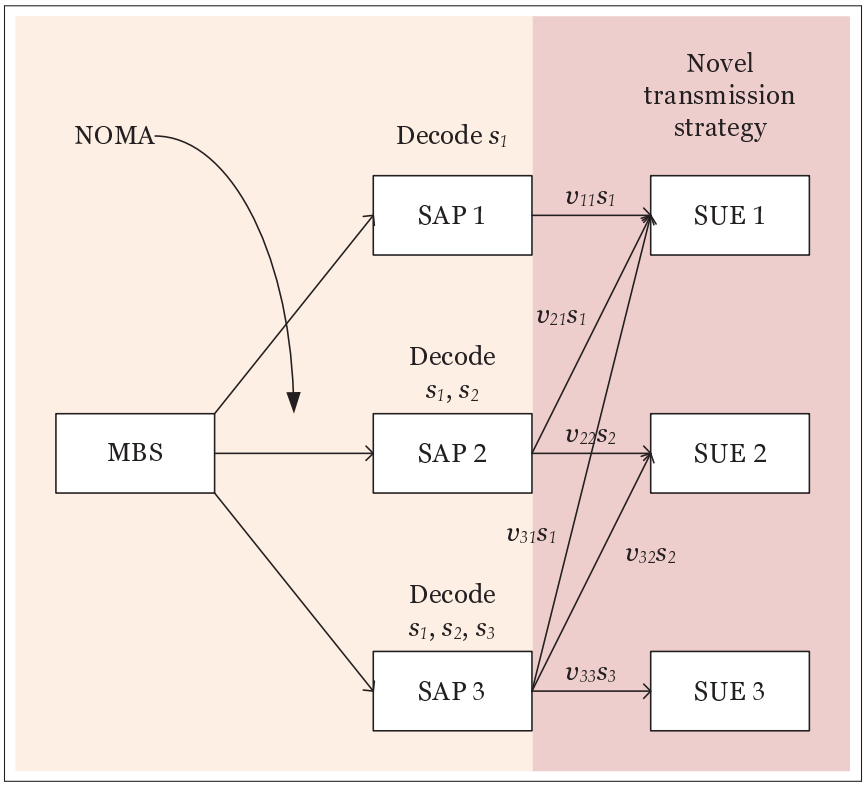

Figure 5.2 Details of macrocell and SC transmissions under the proposed cooperative NOMA scheme.

message s_1, in addition to their own messages s_2 and s_3, respectively, to help SAP 1 improve the signal reception at SUE 1. Without this cooperation, each SAP $\ell > 1$ will only transmit s_ℓ without helping to transmit s_1. By assuming that $\mathbf{v}_1 = [v_{11}, v_{21}, v_{31}], \mathbf{v}_2 = [v_{12}, v_{22}, v_{32}], \mathbf{v}_3 = [v_{13}, v_{23}, v_{33}]$ are the transmit beamformers from SAP 1, 2, and 3, respectively, the received signal at each SUE is

$$y_i = \mathbf{v}_i^H \mathbf{g}_i s_i + \sum_{k \neq i} \mathbf{v}_k^H \mathbf{g}_i s_k + z_i, \tag{5.7}$$

where $\mathbf{g}_i = [g_{1i}, g_{2i}, g_{3i}]^T$ is the vector which concatenates the channel from all the SAPs to the SUE in the i^{th} small cell. Note that we associate the cooperation capability for each SC by

coupling the binary variable $a_{j,i}$ with the weight v_{ij} using the following inequalities

$$\begin{cases} \left|v_{ij}\right|^2 \leq (1-a_{j,i})\lambda_{ij} & ,\forall i,j=1,2,3, \\ \sum_{j=1}^{3}\lambda_{ij} \leq p_{\max} & ,\forall i=1,2,3, \end{cases} \tag{5.8}$$

where $\lambda_{ij} \in \mathbb{R}^+$ represents the soft power level that the SAP i applies at the j^{th} SUE. In our example, $a_{1,2} = a_{1,3} = a_{2,3} = 1$ where the remaining elements of $\mathbf{a}$ are 0. This means that v_{12}, v_{13}, v_{23} are enforced to 0, where the remaining elements of each vector $\mathbf{v}_i, \forall i = 1,2,3$ can take any complex value whose norm is from 0 to λ_{ij}. In general case of S SCs, to mathematically derive the above cooperative transmission and reception, let us denote $\mathbf{v}_i = [v_{1i}, \ldots, v_{Si}]^T$ and $\mathbf{g}_i = [g_{1i}, \ldots, g_{Si}]^T$ as the array of weights and channel vector from all the SAPs to the SUE of the i^{th} small cell. Then, the received signal at the SUE of the i^{th} small cell is

$$y_i = \mathbf{v}_i^H \mathbf{g}_i s_i + \sum_{k \in \mathscr{S} \setminus i} \mathbf{v}_k^H \mathbf{g}_i s_k + z_i, \tag{5.9}$$

where we also have

$$\left|v_{ij}\right|^2 \leq (1-a_{j,i})\lambda_{ij}, \forall i,j \in \mathscr{S} \tag{5.10}$$

$$\sum_{j \in \mathscr{S}} \lambda_{ij} \leq p_{\max}, \forall i \in \mathscr{S}, \tag{5.11}$$

and s_i is the message intended for the SUE in the i^{th} small cell with unit average power, e.g., $\mathbb{E}\{s_i s_i^*\} = 1$, z_j is a AWGN at the SUE in the i^{th} small cell, distributed according to a normal distribution $\mathscr{CN}(0, N_{02})$. Treating interference as noise, the achievable rate $r_j(\mathbf{v})$ of the j^{th} SAP, computed in bps/Hz, is

$$r_i(\mathbf{v}) = (1-\alpha)\log\left(1 + \frac{\left|\mathbf{v}_i^H \mathbf{g}_i\right|^2}{\sum_{k \in \mathscr{S} \setminus i}\left|\mathbf{v}_k^H \mathbf{g}_i\right|^2 + N_{02}}\right), \tag{5.12}$$

where we denote $N_{02} = (1-\alpha)W\sigma_0$ as the noise power over the $(1-\alpha)W$ band.

5.3 Problem Formulations

In this section, we aim at jointly optimizing the decoding order together with the radio resource allocation to solve two optimization problems. The first problem is to maximize the sum rate of the SUEs and MUEs, which can be formulated as

$$(\mathscr{P}_1): \max_{\mathbf{w},\mathbf{v},\mathbf{a},\lambda} \quad \sum_{j\in\mathscr{S}} r_j(\mathbf{v}) + \sum_{k\in\mathscr{M}} R_{k,k}(\mathbf{w}) \tag{5.13a}$$

$$\text{s.t.} \quad R_{j,j}(\mathbf{w},\mathbf{a}) \geq r_j(\mathbf{v}), \forall j \in \mathscr{S} \tag{5.13b}$$

$$R_{j,\ell}(\mathbf{w},\mathbf{a}) \geq (1-a_{\ell.j}) R_{\ell,\ell}(\mathbf{w},\mathbf{a}), \forall \ell, j \in \mathscr{S} \tag{5.13c}$$

$$\left|v_{ij}\right|^2 \leq (1-a_{j,i})\lambda_{ij}, \forall i, j \in \mathscr{S} \tag{5.13d}$$

$$\sum_{j\in\mathscr{S}} \lambda_{ij} \leq p_{\max}, \forall j \in \mathscr{S}; \sum_{i\in\mathscr{F}} \|\mathbf{w}_i\|^2 \leq P_{\max} \tag{5.13e}$$

$$a_{j,\ell} + a_{\ell.j} = 1, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.13f}$$

$$a_{j,j} = 0, a_{i,\ell} \geq a_{i,j} a_{j,k}, \forall i, j, k \in \mathscr{S} \tag{5.13g}$$

$$a_{j,\ell} \in \{0,1\}, \forall j, \ell \in \mathscr{S}, \tag{5.13h}$$

The second problem is to maximize the number of satisfied SUEs and MUEs whose achievable rates exceed the minimum rate requirement $\gamma_{\min}$ and $\Gamma_{\min}$, which can be formulated as

$$(\mathscr{P}_2): \max_{\mathbf{w},\mathbf{v},\mathbf{a},\lambda,\mathbf{x},\mathbf{y},\omega} \quad \sum_{j\in\mathscr{S}} x_j + \sum_{k\in\mathscr{M}} y_k \tag{5.14a}$$

$$\text{s.t.} \quad r_j(\mathbf{v}) \geq x_j \gamma_{\min}, \forall j \in \mathscr{S}; R_{k,k}(\mathbf{w}) \geq y_k \Gamma_{\min}, \forall k \in \mathscr{M} \tag{5.14b}$$

$$\left\|\mathbf{w}_j\right\|^2 \leq x_j \omega_j, \forall j \in \mathscr{S}; \|\mathbf{w}_k\|^2 \leq y_k \omega_k, \forall k \in \mathscr{M} \tag{5.14c}$$

$$\sum_{i\in\mathscr{F}} \omega_i \leq P_{\max}; \sum_{j\in\mathscr{S}} \lambda_{ij} \leq x_i p_{\max}, \forall j \in \mathscr{S} \tag{5.14d}$$

$$x_j = \{0,1\}, \forall j \in \mathscr{S}, y_k = \{0,1\}, \forall k \in \mathscr{M} \tag{5.14e}$$

$$(5.13b) - (5.13d), (5.13f) - (5.13h) \tag{5.14f}$$

The characteristics of these introduced problems deserve more elaboration. In $(\mathscr{P}_1)$, we aim at maximizing the continuous value of the total achievable rate at the SUEs and MUEs. This means that in some scenario, there might exist some admitted users with low rates while the other user rates are much higher to arrive at the optimal sum rate objective (5.13a). On the contrary, problem $(\mathscr{P}_2)$ guarantees that each admitted user must exceed a minimum rate requirement, e.g., $\gamma_{\min}$ or $\Gamma_{\min}$, via constraint (5.14b) when the maximum number of admitted users is achieved.

To convey the property of the WB networks into these optimization problems, constraints (5.13b) is imposed to restrict that the WB downlink rate of each SAP at the t^{th} time slot should always be larger than or equal to the corresponding WA downlink rate at each SAP at the $(t+2)^{\text{th}}$ time slot. Since we consider time invariant channel within the considered time slots, we omit the notation of time index for ease of presentation. Note that the backhaul rate formula, e.g., $R_{j,j}(\mathbf{w},\mathbf{a}), \forall j \in \mathscr{S}$, under the proposed NOMA technique in (5.4) is more complicated than the existing expression in (Nguyen *et al.*, 2016b; Wang *et al.*, 2016) due to the NOMA-related decoding capability constraints in (5.13c). (5.13b) also implies a complicated coupling between variables $\mathbf{w}, \mathbf{v}$, and $\mathbf{a}$ by non-concave non-convex functions, i.e., $R_{j,j}(\mathbf{w},\mathbf{a}), r_j(\mathbf{v}), \forall j \in \mathscr{S}$. The first and second terms in (5.13e) are the maximum power budget constraint at the SAP and MBS, respectively. To reflect the impact of decoding capability of the proposed NOMA in Section 5.2.2.1, we also present the binary constraints (5.13f)–(5.13h). Note that the relationship of the decoding capability at each small cell also affects the setting of each weight's value at each SAP, as shown in (5.13d). In addition, in $(\mathscr{P}_2)$, we present the maximum power budget constraints (5.14c)–(5.14d) slightly differently by adopting the concept of perspective reformulation in (Günlük & Linderoth, 2010) and the slack variables $\omega = \{\omega_i \geq 0, \forall i \in \mathscr{F}\}$. By this, when the binary variable $x_j, \forall j \in \mathscr{S}$ (or $y_k, \forall k \in \mathscr{M}$) take value 0, the respective beamforming vector $\mathbf{w}_j$ and $\mathbf{v}_j$ for $j \in \mathscr{S}$ (or $\mathbf{w}_k$ for $k \in \mathscr{M}$) will be enforced to 0. Towards this end, we provide the following remark to characterize the performance of problem $(\mathscr{P}_1)$ over the conventional works (Wang *et al.*, 2016; Zhao *et al.*, 2015; Nguyen *et al.*, 2016b) in certain channel conditions.

Proposition 5. *Without applying NOMA in Section 5.2.2, the problem* $(\mathscr{P}_1)$ *is rewritten as*

$$(\mathscr{P}_{non}): \max_{\mathbf{W},\mathbf{V}} \quad \sum_{j\in\mathscr{S}} \tilde{r}_j(\mathbf{v}) + \sum_{k\in\mathscr{M}} \tilde{R}_{k,k}(\mathbf{w}) \tag{5.15a}$$

$$s.t. \quad \tilde{R}_{j,j}(\mathbf{w}) \geq \tilde{r}_j(\mathbf{v}), \forall j \in \mathscr{S} \tag{5.15b}$$

$$\sum_{i\in\mathscr{F}} \|\mathbf{w}_i\|_2^2 \leq P_{\max}, \left|v_{jj}\right|^2 \leq p_{\max}, \forall j \in \mathscr{S}, \tag{5.15c}$$

where

$$\tilde{R}_{j,j}(\mathbf{w}) = \alpha \log\left(1 + \frac{\left|\mathbf{w}_j^H \mathbf{h}_j\right|^2}{\sum\limits_{k\in\mathscr{F}\setminus j} \left|\mathbf{w}_k^H \mathbf{h}_j\right|^2 + N_{01}}\right) \tag{5.16}$$

$$\tilde{r}_j(\mathbf{v}) = (1-\alpha) \log\left(1 + \frac{\left|v_{jj}^* g_{jj}\right|^2}{\sum_{k\in\mathscr{S}\setminus j} \left|v_{kk}^* g_{kj}\right|^2 + N_{02}}\right), \tag{5.17}$$

which is similar to the formulation in (Nguyen et al.*, 2016c). We observe that in some specific scenarios of* $(\mathscr{P}_1)$ *where* $R_{j,j}(\mathbf{w},\mathbf{a})$ *from* (5.4) *satisfies constraints* (5.13c) *for all feasible* $\mathbf{w}$*, the optimal solution of* $(\mathscr{P}_1)$ *provides a backhaul rate tuple higher than the optimal solution of* $(\mathscr{P}_{non})$*. In such cases, the optimal objective of* $(\mathscr{P}_1)$ *is provable to be greater than or equal to that of* $(\mathscr{P}_{non})$*. The detailed proof for this is given in the Appendix 8.*

In order to solve $(\mathscr{P}_1)$ and $(\mathscr{P}_2)$, we first assume that global channel state information (CSI) is available at all the transmitters (Wang *et al.*, 2016; Niu *et al.*, 2017; Zhao *et al.*, 2015). The assumption of global CSI to embrace the centralized computation of problem $(\mathscr{P}_1)$ and $(\mathscr{P}_2)$ requires that all the necessary CSI are gathered at one central computational node. This is necessary to execute the later developed algorithm to attain the solution. Thus, a cost of exchanging of $|\mathbf{h}|$ and $|\mathbf{g}|$ CSI elements, e.g. $C(|\mathbf{h}|+|\mathbf{g}|)$, towards to the computational node, where $\mathbf{h} = \{\mathbf{h}_i, \forall i \in \mathscr{F}\}$ and $\mathbf{g} = \{g_{j\ell}, \forall j,\ell \in \mathscr{S}\}$, is required.

Toward this end, we observe that both problems $(\mathscr{P}_1)$ and $(\mathscr{P}_2)$ are mixed integer non-convex optimization problems, which are generally NP-hard and difficult to solve. The non-convexity of $(\mathscr{P}_1)$ and $(\mathscr{P}_2)$ is due to the existence of the non-convex rate functions which lie in the objective functions and the constraints (5.13b)–(5.13c). Finding a global solution for each problem $(\mathscr{P}_1)$ or $(\mathscr{P}_2)$ often requires a high–complexity exhaustive search algorithm, which is impractical and cumbersome. This high computational complexity is illustrated in the numerical results section, where we employ an exhaustive search based on branch and bound (BnB) algorithm to solve for the global optimal solution. Motivated by these, we only aim at developing a more appealing low-complexity algorithm to attain the solution within a polynomial time. In the next sections, we mainly concentrate to develop our framework to solve $(\mathscr{P}_1)$, where solving $(\mathscr{P}_2)$ can follow the similar steps.

5.4 Proposed Efficient Solution Methodology

5.4.1 Equivalent Problem Transformation

Let us first equivalently rewrite the binary constraint in (5.13h) in a D.C form as

$$a_{j,\ell} - a_{j,\ell}^2 \leq 0 \tag{5.18a}$$

$$0 \leq a_{j,\ell} \leq 1. \tag{5.18b}$$

Next, we can rewrite (5.13g) into a system of linear binary constraints as

$$\begin{cases} a_{i,k} \geq b_{i,j,k} \\ b_{i,j,k} \geq a_{i,j}, b_{i,j,k} \geq a_{j,k} \quad ,\forall i,j,k \in \mathscr{S} \\ b_{i,j,k} \geq (a_{i,j} + a_{j,k}) - 1 \end{cases} \tag{5.19}$$

where $b_{i,j,k}$ is another slack variable. For convenience, we denote $\mathbf{b} = \{b_{i,j,k}, \forall i,j,k \in \mathscr{S}\}$. Next, we also introduce some other slack variables $\boldsymbol{\chi} = \{\chi_j \geq 0, \forall j \in \mathscr{S}\}$, and $\boldsymbol{\gamma} = \{\gamma_k \geq 0, \forall k \in \mathscr{M}\}$, $\boldsymbol{\xi} = \{\xi_j \geq 0, \forall j \in \mathscr{S}\}$, and $\boldsymbol{\theta} = \{\theta_j \geq 0, \forall j \in \mathscr{S}\}$ and equivalently represent

constraints (5.13b) and (5.13c) as

$$\xi_j \geq r_j(\mathbf{v}) \geq \chi_j, \forall j \in \mathscr{S}, \tag{5.20}$$

$$R_{k,k}(\mathbf{w}) \geq \gamma_k, \forall k \in \mathscr{M}, \tag{5.21}$$

$$1/\theta_j \geq R_{j,j}(\mathbf{w},\mathbf{a}) \geq \xi_j, \forall j \in \mathscr{S}, \tag{5.22}$$

$$R_{j,\ell}(\mathbf{w},\mathbf{a}) \geq \frac{(1-a_{\ell.j})^2}{\theta_j}, \forall \ell, j \in \mathscr{S}, \tag{5.23}$$

Similarly, we introduce slacking variables $\mathbf{t} = \left\{t_j \geq 0, \forall j \in \mathscr{S}\right\}$, $\mathbf{q} = \left\{q_j \geq 0, \forall j \in \mathscr{S}\right\}$ to rewrite the right side of (5.20) as (5.24b), (5.24d), (5.24e) and $\mathbf{u} = \left\{u_j \geq 0, \forall j \in \mathscr{S}\right\}$, to rewrite the left side of (5.20) as (5.24c) and (5.24f). Then, we also introduce $\tau^m = \{\tau_{kk} \geq 0, \forall k \in \mathscr{M}\}$, $\tau^s = \left\{\tau_{j\ell} \geq 0, \forall \ell, j \in \mathscr{S}\right\}$, $\mathbf{s} = \left\{s_{\ell j} \geq 0, \forall \ell, j \in \mathscr{S}\right\}$, $\zeta = \left\{\zeta_{j\ell} \geq 0, \forall \ell, j \in \mathscr{S}\right\}$ to rewrite constraint (5.21) and the right side of (5.22) as constraint (5.24g)–(5.24l). Finally, we also introduce $\mu = \left\{\mu_j \geq 0, \forall j \in \mathscr{S}\right\}$, $\pi = \left\{\pi_{\ell j} \geq, \forall l, j \in \mathscr{S}\right\}$ to rewrite the left side of (5.22) as (5.24m)–(5.24o). The resulted equivalent transformation of problem (5.13) is recasted as

$$\max_{\substack{\mathbf{w},\mathbf{v},\mathbf{a},\mathbf{b},\lambda,\\ \chi,\gamma,\xi,\theta,\pi,\mathbf{q},\\ \mathbf{s},\mathbf{u},\mu,\mathbf{t},\tau,\zeta}} \quad \sum_{j\in\mathscr{S}} \chi_j + \sum_{k\in\mathscr{M}} \gamma_k \tag{5.24a}$$

$$\text{s.t.} \quad (1-\alpha)\log\left(1+t_j\right) \geq \chi_j, \forall j \in \mathscr{S} \tag{5.24b}$$

$$(1-\alpha)\log\left(1+u_j\right) \leq \xi_j, \forall j \in \mathscr{S} \tag{5.24c}$$

$$\left|\mathbf{v}_j^H \mathbf{g}_j\right|^2 / q_j \geq t_j, \forall j \in \mathscr{S} \tag{5.24d}$$

$$\sum_{k\in\mathscr{S}\setminus j} \left|\mathbf{v}_k^H \mathbf{g}_j\right|^2 + N_{02} \leq q_j, \forall j \in \mathscr{S} \tag{5.24e}$$

$$\left|\mathbf{v}_j^H \mathbf{g}_j\right|^2 / u_j \leq \sum_{k\in\mathscr{S}\setminus j} \left|\mathbf{v}_k^H \mathbf{g}_j\right|^2 + N_{02}, \forall j \in \mathscr{S} \tag{5.24f}$$

$$\alpha\log\left(1+\tau_{kk}\right) \geq \gamma_k, \forall k \in \mathscr{M} \tag{5.24g}$$

$$\alpha\log\left(1+\tau_{jj}\right) \geq \xi_j, \forall j \in \mathscr{S} \tag{5.24h}$$

$$\frac{\left|\mathbf{w}_k^H \mathbf{h}_k\right|^2}{\tau_{kk}} \geq \sum_{l\in\mathscr{F}\setminus k} \left|\mathbf{w}_l^H \mathbf{h}_k\right|^2 + N_{01}, \forall k \in \mathscr{M} \tag{5.24i}$$

$$\frac{\left|\mathbf{w}_j^H \mathbf{h}_j\right|^2}{s_{jj}+N_{01}} \geq \tau_{jj}, \forall j \in \mathscr{S} \tag{5.24j}$$

$$\sum_{\ell \in \mathscr{S}\setminus j} \frac{a_{\ell,j}^2}{\zeta_{\ell,j}} + \sum_{k \in \mathscr{M}} \left|\mathbf{w}_k^H \mathbf{h}_j\right|^2 \leq s_{jj}, \forall j \in \mathscr{S} \tag{5.24k}$$

$$\left|\mathbf{w}_\ell^H \mathbf{h}_j\right|^2 \leq \frac{1}{\zeta_{\ell,j}}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.24l}$$

$$\alpha \log\left(1+\mu_j\right) \leq \frac{1}{\theta_j}, \forall j \in \mathscr{S} \tag{5.24m}$$

$$\frac{\left|\mathbf{w}_j^H \mathbf{h}_j\right|^2}{\mu_j} \leq \sum_{\ell \in \mathscr{F}\setminus j} \pi_{\ell j}^2 + N_{01}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.24n}$$

$$a_{\ell,j} + \left|\mathbf{w}_\ell^H \mathbf{h}_j\right|^2 \geq \sqrt{\pi_{\ell j}^2 + \left(a_{\ell,j} - \left|\mathbf{w}_\ell^H \mathbf{h}_j\right|^2\right)^2}, \quad \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.24o}$$

$$\alpha \log\left(1+\tau_{\ell j}\right) \geq \frac{a_{\ell,j}^2}{\theta_j}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.24p}$$

$$\frac{\left|\mathbf{w}_j^H \mathbf{h}_\ell\right|^2}{s_{\ell j}+N_{01}} \geq \tau_{\ell j}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.24q}$$

$$\sum_{\ell \in \mathscr{S}\setminus j} \frac{a_{\ell,m}^2}{\zeta_{\ell,m}} + \sum_{k \in \mathscr{M}} \left|\mathbf{w}_k^H \mathbf{h}_\ell\right|^2 \leq s_{\ell j}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.24r}$$

$$\left|\mathbf{w}_m^H \mathbf{h}_\ell\right|^2 \leq \frac{1}{\zeta_{\ell,m}}, \forall \ell, m \in \mathscr{S}, \ell \neq m \tag{5.24s}$$

$$(5.13\text{d}) - (5.13\text{f}), (5.18), (5.19). \tag{5.24t}$$

5.4.2 Successive Convex Approximation (SCA) Method

In this section, owing to the fact that (5.24) is still difficult to solve, we will present the SCA method to approximate the non-convex problem (5.24) and develop an iterative low-complexity algorithm to attain the solution. The difficulties of problem (5.24) are recognized through some existing constraint types. The first type of constraint involves in the binary variables $\mathbf{a}$, e.g., (5.13f), (5.18), and (5.19). If we directly employ the SCA method under the consideration

of constraint (5.18), there will be some steps where solving the approximated problem result is infeasible because of (5.18), which leads to a failed convergence of the mentioned iterative algorithm. Inspired by the approach in (Vu *et al.*, 2016a), we overcome this obstacle by considering the relaxed version of problem (5.24) as

$$\max_{\substack{\mathbf{w},\mathbf{v},\mathbf{a},\mathbf{b},\lambda,\phi,\\ \chi,\gamma,\xi,\theta,\pi,\mathbf{q},\\ \mathbf{s},\mathbf{u},\mu,\mathbf{t},\tau,\zeta}} \quad \sum_{j\in\mathscr{S}} \chi_j + \sum_{k\in\mathscr{M}} \gamma_k - A \sum_{j,\ell\in\mathscr{S}} \phi_{j,\ell} \tag{5.25a}$$

$$\text{s.t.} \quad (5.24\text{b}) - (5.24\text{s}), (5.13\text{d}) - (5.13\text{f}), (5.18\text{b}), (5.19), \tag{5.25b}$$

$$a_{j,\ell} - a_{j,\ell}^2 \leq \phi_{j,\ell} \forall \ell, j \in \mathscr{S}, \tag{5.25c}$$

where $\phi = \left\{\phi_{j,\ell} \geq 0 \forall \ell, j \in \mathscr{S}\right\}$ is a new slack variable and $A > 0$ is the penalty parameter. It is obvious to remark that (5.24) and (5.25) are equivalent when $\phi_{j,\ell} = 0$. Under the relaxed problem (5.25), we also remark that the second constraint type includes all the convex constraints (5.24b), (5.24e), (5.24g), (5.24h), (5.24k), (5.24n), (5.24p), (5.24r), and (5.13d)–(5.13f). The third type contains the non-convex constraints (5.24c), (5.24d), (5.24f), (5.24i), (5.24l), (5.24m), (5.24o), (5.24q), and (5.24s), (5.25c). The non-convexity of the constraints (5.24d), (5.24f), (5.24i), (5.24j), (5.24l), (5.24m), (5.24o), (5.24q), and (5.24s) are due to the existence of the convex functions on the greater side of the inequalities. It is worth to mention that the greater sides of (5.24d), (5.24i), (5.24l), (5.24m), (5.24q), (5.24s) have a generic form as $f(x,y) = \frac{|y|^2}{x}, \forall y \in \mathbb{C}, \forall x \in \mathbb{R}^+$. On the other hand, the greater sides of (5.24o), (5.24f) has the quadratic form with respect to variable $\mathbf{w}$ or $\mathbf{v}$. Beside, the non-convexity of constraints (5.24c), (5.24m), and (5.25c) are due to the concave functions which lie on the lesser side of the inequality. Toward this end, we first proceed to approximate function $f(x,y)$ by its lower–bounded concave approximation $F(x,y;x^{(n)},y^{(n)})$ as (Nguyen *et al.*, 2016b):

$$f(x,y) \geq F(x,y;x^{(n)},y^{(n)}) = \frac{2\text{Re}\left\{y^{(n)*}y\right\}}{x^{(n)}} - \frac{\left|y^{(n)}\right|^2}{\left(x^{(n)}\right)^2}x \tag{5.26}$$

Similarly, we can also approximate the function $\hat{h}_j(\mathbf{g},\mathbf{v}) = \sum_{k\in\mathscr{S}\setminus j}\left|\mathbf{v}_k^H\mathbf{g}_j^k\right|^2 + N_{02}$ respectively appeared in constraint (5.24f) and functions $g(u_j) = \log\left(1+u_j\right), g(\mu_i) = \log\left(1+\mu_i\right)$ respectively appeared in constraints (5.24c) and (5.24m) by their lower-bounded concave and upper-bounded convex approximation $\hat{H}_j(\mathbf{g}.\mathbf{v};\mathbf{v}^{(n)})$, and $G(u_j;u_j^{(n)})$, respectively, as:

$$\hat{h}_j(\mathbf{g},\mathbf{v}) \geq \hat{H}_j(\mathbf{g},\mathbf{v};\mathbf{v}^{(n)}) = \sum_{k\in\mathscr{S}\setminus j} 2\mathrm{Re}\left(\mathbf{v}_k^{(n)H}\mathbf{G}_j\mathbf{v}_k\right) - \sum_{k\in\mathscr{S}\setminus j}\left|\mathbf{v}_k^{(n)H}\mathbf{g}_j\right|^2 \tag{5.27}$$

$$g(u_j) \leq G(u_j;u_j^{(n)}) = \log\left(1+u_j^{(n)}\right) + \frac{u_j - u_j^{(n)}}{1+u_j^{(n)}}. \tag{5.28}$$

where we denote $\mathbf{G}_j^k = \mathbf{g}_j^k\left(\mathbf{g}_j^k\right)^H, \forall j \in \mathscr{S}$. The characteristics of the above approximations deserve some elaboration. First, it is obvious that (5.26) achieves its equality when $x = x^{(n)}$ and $y = y^{(n)}$. Second, $\nabla f(x,y) = \nabla(x,y;x^{(n)},y^{(n)})$ when $x = x^{(n)}$ and $y = y^{(n)}$. Third, $f(x,y)$ is lower bounded by a convex approximated function $F(x,y;x^{(n)},y^{(n)})$. Similar analyses can be employed for (5.27) and (5.28). According to the theory (Marks & Wright), employing these approximation creates a sequence of convex approximate of the continuous non-convex problem $(\mathscr{P}_1)$, where sequentially optimizing these convex approximated problems produces a non-decreasing sequence of objective values and eventually leads to convergence. Motivated by this, we replace the non-convex constraints (5.24d), (5.24i), (5.24l), (5.24m), (5.24q), (5.24s) by their approximation using the proper lower–bounded concave approximation in (5.26), the non-convex constraint (5.24f) by the approximation in (5.27), and non-convex constraints (5.24c) and (5.24m) by the approximation in (5.28). Then, the solution of (5.24) can be attained by solving the following approximated problem at iteration $(n+1)$ as

$$\max_{\substack{\mathbf{w},\mathbf{v},\mathbf{a},\mathbf{b},\lambda,\phi,\\ \chi,\gamma,\xi,\theta,\pi,\mathbf{q},\\ \mathbf{s},\mathbf{u},\mu,\mathbf{t},\tau,\zeta}} \quad \sum_{j\in\mathscr{S}}\chi_j + \sum_{k\in\mathscr{M}}\gamma_k - A\sum_{j,\ell\in\mathscr{S}}\phi_{j,\ell} \tag{5.29a}$$

$$\text{s.t.} \quad (1-\alpha)\log\left(1+t_j\right) \geq \chi_j, \forall j \in \mathscr{S} \tag{5.29b}$$

$$(1-\alpha)G(u_j;u_j^{(n)}) \leq \xi_j, \forall j \in \mathscr{S} \tag{5.29c}$$

$$F(q_j,\mathbf{v}_j^H\mathbf{g}_j;q_j^{(n)},\mathbf{v}_j^{(n)H}\mathbf{g}_j) \geq t_j, \forall j \in \mathscr{S} \tag{5.29d}$$

$$\frac{\left|\mathbf{v}_j^H \mathbf{g}_j\right|^2}{u_j} \leq \hat{H}_j(\mathbf{g}.\mathbf{v};\mathbf{v}^{(n)}), \forall j \in \mathscr{S} \tag{5.29e}$$

$$\alpha \log(1+\tau_{kk}) \geq \gamma_k, \forall k \in \mathscr{M}; \alpha \log\left(1+\tau_{jj}\right) \geq \xi_j, \forall j \in \mathscr{S} \tag{5.29f}$$

$$F(\tau_{kk}, \mathbf{w}_k^H \mathbf{h}_k; \tau_{kk}^{(n)}, \mathbf{w}_k^{(n)H} \mathbf{h}_k) \geq \sum_{l \in \mathscr{F} \setminus k} \left|\mathbf{w}_l^H \mathbf{h}_k\right|^2 + N_{01}, \forall k \in \mathscr{M} \tag{5.29g}$$

$$F(s_{jj}+N_{01}, \mathbf{w}_j^H \mathbf{h}_j; s_{jj}^{(n)}+N_{01}, \mathbf{w}_j^{(n)H} \mathbf{h}_k) \geq \tau_{jj}, \forall j \in \mathscr{S} \tag{5.29h}$$

$$\left|\mathbf{w}_\ell^H \mathbf{h}_j\right|^2 \leq F(\zeta_{\ell,j}, 1; \zeta_{\ell,j}^{(n)}, 1), \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.29i}$$

$$(1-\alpha) G(\mu_j; \mu_j^{(n)}) \leq F(\theta_j, 1; \theta_j^{(n)}, 1), \forall j \in \mathscr{S} \tag{5.29j}$$

$$\frac{\left|\mathbf{w}_j^H \mathbf{h}_j\right|^2}{\mu_j} \leq \hat{H}_j(1, \pi; 1, \pi^{(n)}) + N_{01}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.29k}$$

$$a_{\ell,j} + 2\mathrm{Re}\left(\mathbf{w}_\ell^{(n)H} \mathbf{H}_j \mathbf{w}_\ell\right) - \left|\mathbf{w}_\ell^{(n)H} \mathbf{h}_j\right|^2 \geq \sqrt{\pi_{\ell j}^2 + \left(a_{\ell,j} - \left|\mathbf{w}_\ell^H \mathbf{h}_j\right|^2\right)^2}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.29l}$$

$$\alpha \log\left(1+\tau_{\ell j}\right) \geq a_{\ell,j}^2 / \theta_j, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.29m}$$

$$F(s_{\ell j}+N_{01}, \mathbf{w}_j^H \mathbf{h}_\ell; s_{\ell j}^{(n)}+N_{01}, \mathbf{w}_j^{(n)H} \mathbf{h}_\ell) \geq \tau_{\ell j}, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.29n}$$

$$\left|\mathbf{w}_m^H \mathbf{h}_\ell\right|^2 \leq F(\zeta_{\ell,m}, 1; \zeta_{\ell,m}^{(n)}, 1), \forall \ell, m \in \mathscr{S}, \ell \neq m \tag{5.29o}$$

$$a_{j,\ell} - \left(2a_{j,\ell}^{(n)} a_{j,\ell} - \left(a_{j,\ell}^{(n)}\right)^2\right) \leq \phi_{j,\ell}, \forall \ell, j \in \mathscr{S} \tag{5.29p}$$

$$(5.24e), (5.13d)-(5.13f), (5.18b), (5.19), (5.24k), (5.24n), (5.24r). \tag{5.29q}$$

5.4.3 Majorization-Minimization Method (MMM) based Approximation and Proposed Algorithm

In the previous subsection, problem (5.24) has been convexified into a sequence of convex problems (5.29), where obtaining the solution of (5.24) involves sequentially solving a sequence of convex problems (5.29). We observe that the approximated problem (5.29) at the $(n+1)^{\text{th}}$ iteration is categorized as a generalized nonlinear convex programming (GNCP), which in general can be solved within a polynomial time. When a GNCP contains an exponential cone constraint, we can employ a universal convex solver MATLAB's FMINCON

to seek for the optimal solution. However, this often results in high computational time compared to other standard convex programs dedicated for quadratic problems (Nguyen *et al.*, 2016b). Therefore, this motivates us to approximate (5.29b), (5.29f), and (5.29m) into a convex quadratic constraint. In the existing work (Nguyen *et al.*, 2016b), to approximate each exponential cone constraint by a system of second order cone programming (SOCP), extra $m+4$ slack variables and $m+4$ new constraints must be additionally introduced. This approaches might become excessive when we meet a large-sized problem. Hence, in this work, we propose a novel approximation based on Majorization-Minimization method (Parikh & Boyd, 2014b) which does not require any addition of slack variables and constraints. We provide the following proposition to approximate the generic function $\log(1+x)$.

Proposition 6. *The function $\ell(x) = \log(1+x), \forall x \in R$ can be approximated by its lower-bounded quadratic function $L(x;x^{(n)})$, which is given by*

$$\ell(x) \geq L(x;x^{(n)}) = \log\left(1+x^{(n)}\right) + \frac{x-x^{(n)}}{1+x^{(n)}} - \frac{L}{2}\left(x-x^{(n)}\right)^2 \tag{5.30}$$

where $L=1$ is the Lipschitz constant of $\partial\ell(x)/\partial x$.

Proof. The detailed proof of deriving (5.30) and explaining the chosen value $L=1$ is given in Appendix 9. □

By applying the result of Proposition 6 on the logarithm function in constraints (5.29b), (5.29f), and (5.29m), we can rewrite the GNCP problem (5.29) into a quadratic optimization problem at the $(n+1)$th iteration as

$$\max_{\substack{\mathbf{w},\mathbf{v},\mathbf{a},\mathbf{b},\lambda,\phi,\\ \chi,\gamma,\xi,\theta,\pi,\mathbf{q},\\ \mathbf{s},\mathbf{u},\mu,\mathbf{t},\tau,\zeta}} \quad \sum_{j\in\mathscr{S}}\chi_j + \sum_{k\in\mathscr{M}}\gamma_k - A\sum_{j,\ell\in\mathscr{S}}\phi_{j,\ell} \tag{5.31a}$$

$$\text{s.t.} \quad (1-\alpha)L(t_j;t_j^{(n)}) \geq \chi_j, \forall j \in \mathscr{S} \tag{5.31b}$$

$$\alpha L\left(\tau_{kk};\tau_{kk}^{(n)}\right) \geq \gamma_k, \forall k \in \mathscr{M} \tag{5.31c}$$

$$\alpha L\left(\tau_{jj};\tau_{jj}^{(n)}\right) \geq \xi_j, \forall j \in \mathscr{S} \tag{5.31d}$$

$$\alpha L\left(\tau_{\ell j};\tau_{\ell j}^{(n)}\right) \geq a_{\ell,j}^2/\theta_j, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.31e}$$

$$(5.29c)-(5.29e),(5.29g)-(5.29l),(5.29n)-(5.29q). \tag{5.31f}$$

where $\mathbf{u}^{(n)},\mathbf{q}^{(n)},\mathbf{v}^{(n)},\mathbf{t}^{(n)},\tau^{(n)},\mathbf{w}^{(n)},\mathbf{s}^{(n)},\zeta^{(n)},\mu^{(n)},\theta^{(n)},\pi^{(n)},\mathbf{a}^{(n)}$ are not the optimization variables but the problem parameters that are iteratively updated by the obtained optimal solution of (5.31) at each iteration. Algorithm 6, which is based on successive convex approximation and majorization-minimization method (SCAMMM), presents the pseudo code to solve problem (5.24). First, we initially set A at small value. Then after each iteration n, we increase A using a constant $c>1$ to ensure that $\sum_{j,\ell}\phi_{j,\ell}^{\star}=0$ when $n\to\infty$. The proof that there exist a finite value of $A_{\max}$ and n_1 such that $\sum_{j,\ell}\phi_{j,\ell}^{\star}=0$ and $\left|\sum_{j\in\mathscr{S}}\chi_j^{(n_2)}+\sum_{k\in\mathscr{M}}\gamma_k^{(n_2)}-\sum_{j\in\mathscr{S}}\chi_j^{(n_2-1)}+\sum_{k\in\mathscr{M}}\gamma_k^{(n_2-1)}\right|=0$ at iteration $n_2,\forall n_2>n_1$ is similar to (Vu *et al.*, 2016a). We omit the proof here due to space limitation.

Algorithm 6: Iterative SCAMMM Algorithm

1: Initialize starting points of $\mathbf{u}^{(n)},\mathbf{q}^{(n)},\mathbf{v}^{(n)},\mathbf{t}^{(n)},\tau^{(n)},\mathbf{w}^{(n)},\mathbf{s}^{(n)},\zeta^{(n)},\mu^{(n)},\theta^{(n)},\pi^{(n)},\mathbf{a}^{(n)}$ and $A^{(0)}$;
2: Set $n:=0$;
3: **repeat**
4: Solve the convex problem in (5.31) to achieve the optimal solution $\mathbf{w}^{\star},\mathbf{v}^{\star},\mathbf{a}^{\star},\mathbf{b}^{\star},\lambda^{\star},\phi^{\star},\chi^{\star},\gamma^{\star},\xi^{\star},\theta^{\star},\pi^{\star},\mathbf{u}^{\star},\mu^{\star},\mathbf{t}^{\star},\tau^{\star},\zeta^{\star}$;
5: Set $n:=n+1$;
6: Update $\mathbf{u}^{(n)}=\mathbf{u}^{\star},\mathbf{q}^{(n)}=\mathbf{q}^{\star},\mathbf{v}^{(n)}=\mathbf{v}^{\star},\mathbf{t}^{(n)}=\mathbf{t}^{\star},\tau^{(n)}=\tau^{\star},\mathbf{w}^{(n)}=\mathbf{w}^{\star},\mathbf{s}^{(n)}=\mathbf{s}^{\star},\zeta^{(n)}=\zeta^{\star},\mu^{(n)},=\mu^{\star},\theta^{(n)}=\theta^{\star},\pi^{(n)}=\pi^{\star},\mathbf{a}^{(n)}=\mathbf{a}^{\star}$;
7: Update $A^{(n)}=\min\left\{cA^{(n-1)},A_{\max}\right\}$;
8: **until** Convergence of the objective (5.31a);

By applying similar steps of transformation and approximations, $(\mathscr{P}_2)$ can also be transformed and approximated by its quadratic approximation at iteration $(n+1)$ as

$$\max_{\substack{\mathbf{w},\mathbf{v},\omega,\mathbf{a},\mathbf{b},\lambda,\phi,\\ \mathbf{x},\mathbf{y},\xi,\theta,\pi,\mathbf{q},\Phi,\\ \mathbf{s},\mathbf{u},\mu,\mathbf{t},\tau,\zeta}} \sum_{j\in\mathscr{S}}x_j+\sum_{k\in\mathscr{M}}y_k-A\sum_{j,\ell\in\mathscr{S}}\phi_{j,\ell}-B\left(\sum_{j\in\mathscr{S}}\Phi_j^x+\sum_{k\in\mathscr{M}}\Phi_k^y\right) \tag{5.32a}$$

$$\text{s.t.} \quad (1-\alpha)L(t_j;t_j^{(n)}) \geq x_j\gamma_{\min}, \forall j \in \mathscr{S} \tag{5.32b}$$

$$\alpha L(\tau_{kk};\tau_{kk}^{(n)}) \geq y_k\Gamma_{\min}, \forall k \in \mathscr{M} \tag{5.32c}$$

$$\alpha L\left(\tau_{jj};\tau_{jj}^{(n)}\right) \geq \xi_j, \forall j \in \mathscr{S} \tag{5.32d}$$

$$\alpha L\left(\tau_{\ell j};\tau_{\ell j}^{(n)}\right) \geq a_{\ell,j}^2/\theta_j, \forall \ell, j \in \mathscr{S}, \ell \neq j \tag{5.32e}$$

$$x_j - \left(2x_j^{(n)}x_j - \left(x_j^{(n)}\right)^2\right) \leq \Phi_j^x, \forall j \in \mathscr{S} \tag{5.32f}$$

$$y_k - \left(2y_k^{(n)}y_k - \left(y_k^{(n)}\right)^2\right) \leq \Phi_k^y, \forall k \in \mathscr{M}, \tag{5.32g}$$

$$0 \leq x_j \leq 1, \forall j \in \mathscr{S}; 0 \leq y_k \leq 1, \forall k \in \mathscr{M}, \tag{5.32h}$$

$$(5.29c)-(5.29e),(5.29g)-(5.29l),(5.29n)-(5.29p),(5.24e),$$

$$(5.14c)-(5.14e),(5.13f),(5.18b),(5.19),(5.24k),(5.24n),(5.24r). \tag{5.32i}$$

where $\Phi = \left\{\Phi_j^x \geq 0, \Phi_k^y \geq 0, \forall j \in \mathscr{S}, \forall k \in \mathscr{M}\right\}$. Note that $\mathbf{u}^{(n)}, \mathbf{q}^{(n)}, \mathbf{v}^{(n)}, \mathbf{t}^{(n)}, \tau^{(n)}, \mathbf{w}^{(n)}, \mathbf{s}^{(n)}$, $\zeta^{(n)}, \mu^{(n)}, \theta^{(n)}, \pi^{(n)}$, $\mathbf{a}^{(n)}, \mathbf{x}^{(n)}, \mathbf{y}^{(n)}$ are not the optimization variables but the problem parameters that are iteratively updated by the obtained optimal solution of (5.32) at each iteration. We can directly employ Algorithm 6 to solve (5.14) by simply replacing problem (5.31) by problem (5.32) in Step 4 of Algorithm 6.

5.4.4 Initial point setting

To start Algorithm 6, we must choose an initial point $\mathbf{u}^{(0)}, \mathbf{q}^{(0)}, \mathbf{v}^{(0)}, \mathbf{t}^{(0)}, \tau^{(0)}, \mathbf{w}^{(0)}, \mathbf{s}^{(0)}, \zeta^{(0)}$, $\mu^{(0)}, \theta^{(0)}, \pi^{(0)}, \mathbf{a}^{(0)}$ which is feasible to (5.24). Although (5.24) is non-convex, we can determine this feasible point by the following steps. First, a feasible point $\mathbf{w}^{(0)}$ can be chosen as

$$\begin{cases} \mathbf{w}_1^{(0)} = \cdots = \mathbf{w}_{S-1}^{(0)} = 0; \mathbf{w}_S^{(0)} = \frac{P_{\max}}{S+M}[1,0,\ldots,0]^T \\ \mathbf{w}_k^{(0)} = \frac{P_{\max}}{S+M}[1,0,\ldots,0]^T, \forall k \in \mathscr{M} \\ \zeta_{\ell,j}^{(0)} = \varepsilon, \forall \ell, j \in \mathscr{S}, \ell \neq j, \end{cases} \tag{5.33}$$

where ε can be arbitrary small ($\varepsilon = 10^{-3}$). Then, we can calculate $\mathbf{s}^{(0)}, \forall \ell, j \in \mathscr{S}$ by setting (5.24k) and (5.24r) with equality. In addition, we set $\mu_j^{(0)}$ equal to $\frac{\left|\mathbf{w}_j^{(n)H}\mathbf{h}_j\right|^2}{s_{jj}^{(n)}+N_{01}}$ and $\theta_j^{(0)}$ equal to $\frac{1}{\alpha \log\left(1+\mu_j^{(0)}\right)}, \forall j \in \mathscr{S}$. Alternatively, we can calculate $\boldsymbol{\pi}^{(0)}$ by setting (5.24s) with equality. Then, we choose $\mathbf{v}_1^{(0)} = \cdots = \mathbf{v}_{S-1}^{(0)} = 0$ and choose $\mathbf{v}_S^{(0)}$ by solving

$$\min_{\mathbf{v}_S} \quad \|\mathbf{v}_S\|^2 \tag{5.34a}$$

$$\text{s.t.} \quad \text{Re}\left\{\mathbf{v}_S^H \mathbf{g}_S\right\} \geq \sqrt{\frac{\left|\mathbf{w}_S^{(0)H}\mathbf{h}_S\right|^2}{N_{01}} N_{02}}, \forall j \in \mathscr{S} \tag{5.34b}$$

$$\|\mathbf{v}_S\|^2 \leq p_{\max} \tag{5.34c}$$

and calculate $\mathbf{q}^{(0)}$ at by solving (5.24e) at equality. We can set $u_1^{(0)} = \cdots = u_{S-1}^{(0)} = \varepsilon$, $t_1^{(0)} = \cdots = t_{S-1}^{(0)} = 0$ and choose $u_S^{(0)} = t_S^{(0)} = \frac{\left|\mathbf{v}_S^{(0)H}\mathbf{g}_S\right|^2}{N_{01}}$. Finally, we can choose $a_{1,S}^{(n)} = \cdots = a_{S-1,S}^{(n)} = 0$ and the other element of $\mathbf{a}^{(n)}$ is equal to 0.5. It is worth to note that we can find the initial point using another approach by simply setting an arbitrary point without checking its feasibility. Such approach, which was used in (Venkatraman *et al.*, 2017), employs a penalty method which allows our SCA-based algorithm, e.g., SCAMMM algorithm, to start from any initial point, even if it is infeasible to (5.24). In particular, some slack variables are introduced at each constraint as the violations and these violations are additionally considered to be minimized in the objective. In this way, some first iterations of Algorithm 6 may be infeasible to (5.32), but violations are minimized to be zero as the algorithm iterates. We refer to (Lipp & Boyd, 2016, Algorithm 3.1) for a complete description of this initialization method.

5.4.5 Computational complexity analysis

We first present the worst-case per-iteration computational complexity analysis of Algorithm 6. We note that (5.31) is in the standard quadratic programming, which can be easily rewritten into a second order programming (SOCP). By following the result in (Lobo *et al.*, 1998), we observe that there are approximately $S^3 + 8S^2 + 7S + 2M + N(S+M)$ continuous variables in

(5.31) with $6S+M+3S^2+2S^3$ SOC constraints of dimension 1, $7S^2-4S$ SOC constraints of dimension 2, S SOC constraints of dimension S, S^2+M+1 SOC constraints of dimension $S+M+1$. After omitting the small order, the worst-case per-iteration complexity of Algorithm 6 is approximately $\mathscr{O}(S^9+S^8M+\left(S^3+S^2M\right)N^2(S^2+M^2))$.

5.4.6 Imperfect channel state information (CSI)

In the previous sections, we assume perfect global CSI at the transmitters. In practice, CSI knowledge at the transmitter can be imperfect. Under imperfect CSI, solution from solving (5.13) using estimated CSI knowledge might not belong to the feasible solution domain of (5.13) with perfect CSI. This is because imperfect CSI may lead to the violation of constraint (5.13b) and (5.13c) (Nguyen *et al.*, 2016b). To evaluate the impact of imperfect CSI on the resource allocation design, we propose a robust design of radio resource allocation in the erroneous CSI case using the worst-case design as in (Iserte *et al.*, 2006; Jeong *et al.*, 2011; Zhang *et al.*, 2016). For simplicity, we consider that the erroneous CSI is only due to the channel estimation error. Considering the outdated CSI can be left for future work. Under this assumption, the estimated channel values are obtained with some estimation errors lying in some bounded sets of known size. For instance, the estimated channel model can be given as in (Jeong *et al.*, 2011; Zhang *et al.*, 2016): $\mathbf{h}_i=\hat{\mathbf{h}}_i+\boldsymbol{\varepsilon}_h, \forall i\in\mathscr{F}$, and $\mathbf{g}_j=\hat{\mathbf{g}}_j+\boldsymbol{\varepsilon}_g, \forall j\in\mathscr{S}$, where $\hat{\mathbf{h}}_i, \hat{\mathbf{g}}_j$ are the estimated CSI and $\|\boldsymbol{\varepsilon}_h\|^2\leq\delta_h, \left\|\boldsymbol{\varepsilon}_g\right\|^2\leq\delta_g$ reflect that the estimation errors lie in a bounded set of size δ_h and δ_g, respectively. The robust optimization problem is given by:

$$\left(\mathscr{P}_{\text{imp}}\right): \max_{\mathbf{w},\mathbf{v},\mathbf{a},\lambda} \quad \min_{\left\|\boldsymbol{\varepsilon}_g\right\|^2\leq\delta_g}\sum_{j\in\mathscr{S}} r_j(\mathbf{v}) + \min_{\|\boldsymbol{\varepsilon}_h\|^2\leq\delta_h}\sum_{k\in\mathscr{M}} R_{k,k}(\mathbf{w}) \tag{5.35a}$$

$$\text{s.t.} \quad \min_{\|\boldsymbol{\varepsilon}_h\|^2\leq\delta_h} R_{j,j}(\mathbf{w},\mathbf{a}) \geq \max_{\left\|\boldsymbol{\varepsilon}_g\right\|^2\leq\delta_g} r_j(\mathbf{v}), \forall j\in\mathscr{S} \tag{5.35b}$$

$$\min_{\|\boldsymbol{\varepsilon}_h\|^2\leq\delta_h} R_{j,\ell}(\mathbf{w},\mathbf{a}) \geq \max_{\|\boldsymbol{\varepsilon}_h\|^2\leq\delta_h} (1-a_{\ell,j})R_{\ell,\ell}(\mathbf{w},\mathbf{a}), \forall \ell, j\in\mathscr{S} \tag{5.35c}$$

$$(5.13\text{d})-(5.13\text{h}). \tag{5.35d}$$

where in this problem, transmitters only have the estimated CSI $\hat{\mathbf{h}}_i, \hat{\mathbf{g}}_j$. For simplicity, we follow the framework in (Jeong *et al.*, 2011) to derive an approximation of each term in (5.35b) and (5.35c) as in (5.36a)–(5.36d).

$$\min_{\|\varepsilon_g\|^2 \le \delta_g} r_j(\mathbf{v}) \approx (1-\alpha)\log\left[1+\frac{|\mathbf{v}_j^H \hat{\mathbf{g}}_j|^2 - \delta_g\|\mathbf{v}_j\|^2}{\sum_{k\in\mathscr{S}\setminus j}\left(|\mathbf{v}_k^H \hat{\mathbf{g}}_j|^2 + \delta_g\|\mathbf{v}_k\|^2\right)+N_{02}}\right] \tag{5.36a}$$

$$\max_{\|\varepsilon_g\|^2 \le \delta_g} r_j(\mathbf{v}) \approx (1-\alpha)\log\left[1+\frac{|\mathbf{v}_j^H \hat{\mathbf{g}}_j|^2 + \delta_g\|\mathbf{v}_j\|^2}{\sum_{k\in\mathscr{S}\setminus j}\left(|\mathbf{v}_k^H \hat{\mathbf{g}}_j|^2 - \delta_g\|\mathbf{v}_k\|^2\right)+N_{02}}\right] \tag{5.36b}$$

$$\min_{\|\varepsilon_h\|^2 \le \delta_h} R_{j,i}(\mathbf{w},\mathbf{a}) \approx$$
$$\alpha\log\left[1+\frac{|\mathbf{w}_i^H \hat{\mathbf{h}}_j|^2 - \delta_h\|\mathbf{w}_i\|^2}{\sum_{\ell\in\mathscr{S}\setminus j}\left(a_{\ell,j}|\mathbf{w}_\ell^H \hat{\mathbf{h}}_j|^2 + \delta_h a_{\ell,j}\|\mathbf{w}_\ell\|^2\right)+\sum_{m\in\mathscr{M}}\left(|\mathbf{w}_m^H \hat{\mathbf{h}}_j|^2 + \delta_h\|\mathbf{w}_m\|^2\right)+N_{01}}\right] \tag{5.36c}$$

$$\max_{\|\varepsilon_h\|^2 \le \delta_h} R_{j,j}(\mathbf{w},\mathbf{a}) \approx$$
$$\alpha\log\left[1+\frac{|\mathbf{w}_j^H \hat{\mathbf{h}}_j|^2 + \delta_h\|\mathbf{w}_j\|^2}{\sum_{\ell\in\mathscr{S}\setminus j}\left(a_{\ell,j}|\mathbf{w}_l^H \hat{\mathbf{h}}_j|^2 - \delta_h a_{\ell,j}\|\mathbf{w}_l\|^2\right)+\sum_{m\in\mathscr{M}}\left(|\mathbf{w}_m^H \hat{\mathbf{h}}_j|^2 - \delta_h\|\mathbf{w}_m\|^2\right)+N_{01}}\right]. \tag{5.36d}$$

Towards this end, we can follow the same steps developed in Sections 5.4.2 and 5.4.3 to transform, approximate and develop a low-complexity algorithm similar to Algorithm 6 to solve (5.35) for an efficient solution.

5.5 Numerical Results

In this section, we evaluate the network performance with the proposed algorithm. We assume time-invariant flat Rayleigh fading channels and the pathloss component is computed according to (5.2), wehre $d_0 = 10$ m, $G_l = 1$, $c = 3\times10^8$ m/s, $f_c = 2.36$ GHz. In most of the numerical instances, we apply the spatial scenario where we assume a circular coverage of the macrocell with radius $20d_0$. The MBS is positioned at the center and there are $S = 6$ small cells placed within the considered coverage. We assume that each small cell has a circular coverage of radius $2d_0$ with its SAP at the center and a SUE at the circumference of the coverage. In

addition, we assume $M = 2$ MUEs scattered across the macrocell coverage. Unless being mentioned elsewhere, we choose this scenario as the standard simulation mode, where the number of transmit antenna as $N = 4$ and the maximum transmit power as $P_{\max} = 43$ dBm. On the other hand, the maximum transmit power at the SAPs are set as $p_{\max} = 30$ dBm. We choose the minimum rate requirement $\Gamma_{\min} = 1.5$ bps/Hz and $\gamma_{\min} = 0.1$ bps/Hz. The noise power is $W\sigma_0 = -120$ dB and the bandwidth is $W = 10$ MHz. The convergence of Algorithm 6 is determined when the difference of objective value between two consecutive iteration is below 10^{-3}.

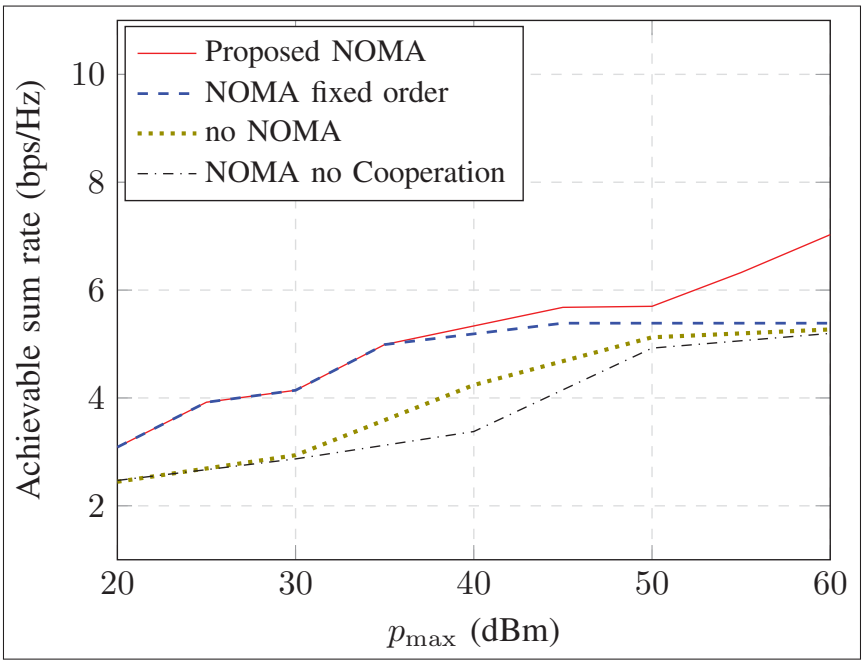

Figure 5.3 Total throughput comparison of the proposed NOMA, NOMA fixed order, no NOMA, and NOMA no cooperation schemes with respect to $p_{\max}$.

In Fig. 5.3, we show the performance of the proposed NOMA, the "NOMA fixed order" (c.f. (Hanif *et al.*, 2016)), the conventional "no NOMA", and the "NOMA no cooperation" schemes. According to (Hanif *et al.*, 2016), the "NOMA fixed order" scheme sorts the decoding order by ranking the norm of channel vectors from the MBS to the SAPs prior to optimizing the resource allocation. The conventional "no NOMA" scheme (c.f. (Wang *et al.*, 2016; Zhao *et al.*,

2015; Nguyen *et al.*, 2016b)) can be translated into the optimization problem in (5.15), and the "NOMA no cooperation" scheme can be interpreted similar to problem (5.13), where the backhaul and MUE rates satisfy the conditions in (5.6), but the SUE rates take the formulas in the right side of (5.16) as a result of no cooperation. We evaluate the sum throughput at the MUEs and SUEs by the solution obtained at the termination of Algorithm 6 with respect to $p_{\max}$. We first validate that our proposed NOMA scheme always performs better than the "NOMA fixed order" scheme. However, depending on the studied channel realization and system parameter setting, the gap between these two schemes may vary. In addition, our proposed NOMA scheme significantly outperforms the conventional "no NOMA" and the "NOMA no cooperation" schemes. Although "NOMA no cooperation" scheme performs almost similar to "no NOMA" scheme at some value of $p_{\max}$ in Fig. 5.3, their performance difference can be more significant under different set of channel realization input $\mathbf{h}_i, \forall i \in \mathscr{F}$ and $\mathbf{g}_j, \forall j \in \mathscr{S}$, as illustrated in Fig. 5.4. This is because equation (5.6) used for "NOMA no cooperation" constraints make the feasible solution domain of "NOMA no cooperation" problem generally different with "no NOMA" problem, so that these two problems are generally different.

Fig. 5.5 compares the achieved throughput at the MUEs and SUEs between the aforementioned schemes by applying Algorithm 6. The throughput performance is shown with respect to the MBS power budget $P_{\max}$. When $P_{\max}$ increases, the throughput of all the considered schemes increase. However, the proposed NOMA scheme always outperforms the other schemes. This can be explained as when $P_{\max}$ increases, more power can be supported to improve the MUE and the backhaul rate at each SAP. Note that improving the backhaul rate enhances the upper bound of the corresponding access rate at each SAP, which can potentially increase the achievable rate delivered to the SUEs. Again, we observe that our proposed NOMA scheme slightly outperforms the "NOMA fixed order" scheme, but is considerably better than the "no NOMA" and "NOMA no cooperation" schemes. We also observe that "no NOMA" performance slightly differs from "NOMA no cooperation" at small $P_{\max}$. At high $P_{\max}$, these two schemes performs similarly. This can be explained as in Fig. 5.3 where due to the value of channel realization input, "no NOMA" and "NOMA no cooperation" schemes might share the

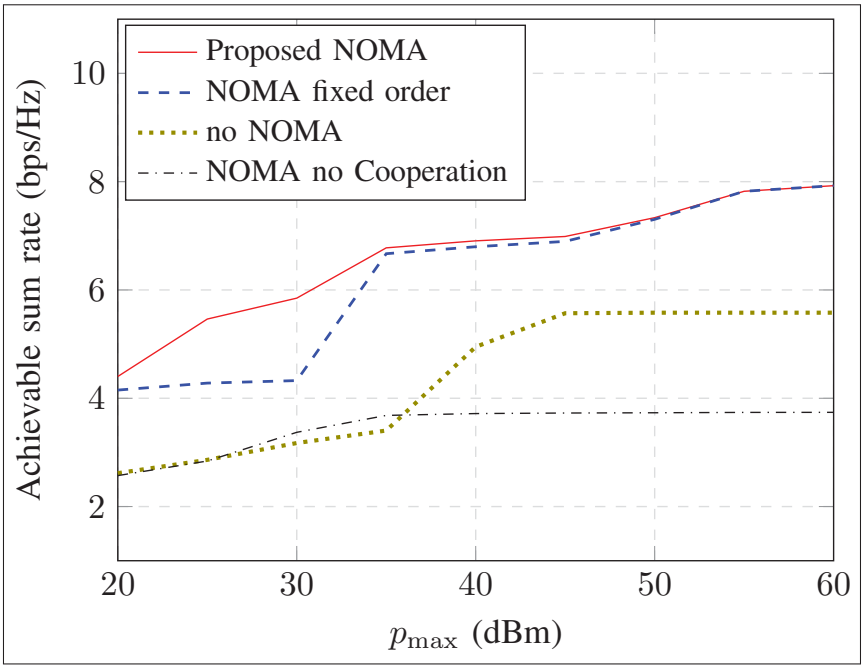

Figure 5.4 Total throughput comparison of the proposed NOMA, NOMA fixed order, no NOMA, and NOMA no cooperation schemes with respect to $p_{\max}$.

same optimal objective value at high $P_{\max}$, which results in the almost overlapped performance of two curves.

In Fig. 5.6, we examine the total achieved throughput of the proposed NOMA scheme by applying Algorithm 6 at different values of the spectrum partitioning factor α. We compare the achieved throughput performance between $P_{\max}$ =38, 43, and 45 dBm. We observe that when α increases from 0.1 to 0.98, the total achieved throughput at the MUEs and SUEs first increases and then decreases. There exists a value of α which maximizes the throughput, where this maximal α value varies for each $P_{\max}$ curve. This can be explained when α is too low, more bandwidth partition is dedicated for the macrocell downlink transmissions from the MBS to the SAPs and MUEs. This leads to a low performance of the total SUE throughput since less bandwidth is provided for the small cell downlink transmissions from the SAPs to the SUEs. When α is too high, more bandwidth is drawn for the small cell downlink transmissions from the SAPs to the SUEs. However, since less bandwidth is now available for the backhaul trans-

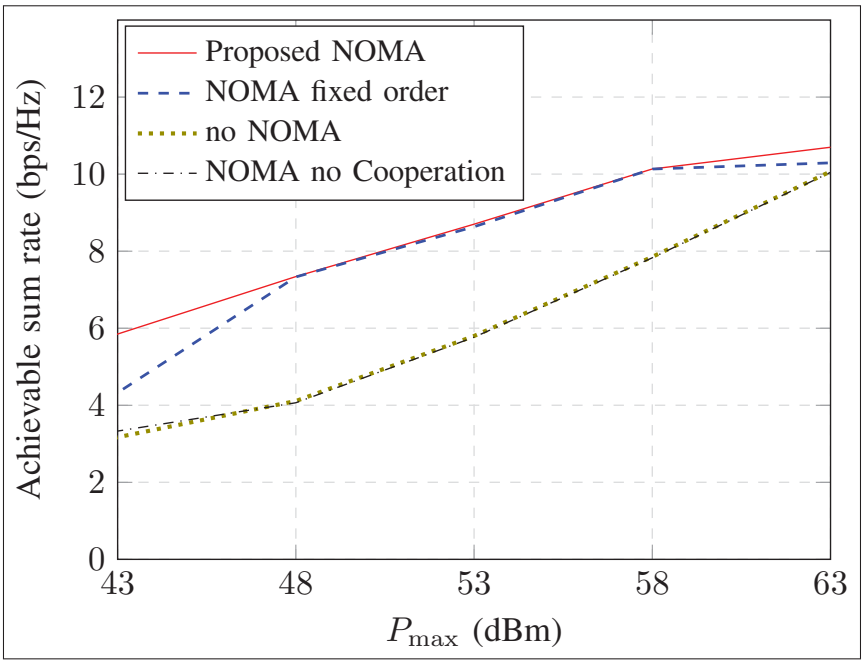

Figure 5.5 Total throughput comparison of the proposed NOMA, NOMA fixed order, no NOMA, and NOMA no cooperation schemes with respect to $P_{\max}$.

mission, this subsequently limits the upper bound performance of the SUE rate, and restricts the total achieved SUE throughput. It is worth to note that this value of α which maximizes the throughput is not a global optimal solution for (5.13). This is because (5.13) is non-convex if we consider a joint optimization of $\mathbf{w}, \mathbf{v}, \mathbf{a}, \lambda, \alpha$. In Fig. 5.6, we fix α and employ Algorithm 6 to seek for a joint sub-optimal solution, so that the achieved solution $\tilde{\alpha}$ together with the $\tilde{\mathbf{w}}$, $\tilde{\mathbf{v}}$, $\tilde{\mathbf{a}}$, and $\tilde{\lambda}$ is jointly sub-optimal. To find a jointly global optimal solution of $\mathbf{w}, \mathbf{v}, \mathbf{a}, \lambda, \alpha$, an exhaustive search algorithm like BnB in (Tervo *et al.*, 2015b) should be used. Since this is not the scope of our paper, we leave this study for future work.

In Fig. 5.7, we show the total achievable throughput versus the network size $S+M$ under the proposed NOMA, the "NOMA fixed order" and the conventional "no NOMA" schemes, where we choose $M = 2$. For each scheme, we compare the network performance between two values of $p_{\max} = 20$ and 30 dBm. We first observe that our proposed NOMA scheme always outperforms the "NOMA fixed order" and "no NOMA" schemes, which is consistent with the previous results showed in the above figures. Moreover, under the proposed NOMA

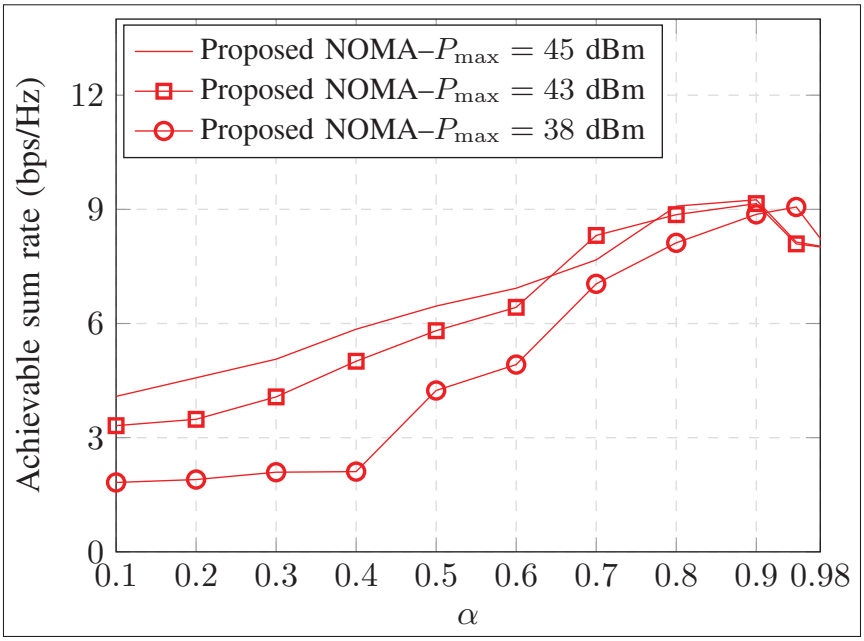

Figure 5.6 Throughput performance of the proposed NOMA scheme with respect to α.

scheme, when the network size increases, the total achieved throughput increases and saturates at high value of $S+M$. The proposed NOMA scheme of higher $p_{\max}$ outperforms the one with lower $p_{\max}$ in the regime of small network size. However, the performance gap between these curves diminishes at larger network size. This can be explained as when the network size grows, more number of small cell can cooperate as the consequence of the cooperative NOMA scheme to transmit to their SUEs. Under the scenario of large network size, the cooperation between these small cells result in more degrees of freedoms, which potentially leads to the maximum throughput solution at lower maximum power budget $p_{\max}$. On the contrary, in the conventional "no NOMA" scheme, the performance gap between the lower and higher $p_{\max}$ is more apparent. At large network size, each received signal must suffer more inter-cell interference in the conventional no NOMA scheme since no small cell cooperation is allowed. This leads to the rapid saturation of these curves when the network size grows.

Fig. 5.8 shows the achieved throughput in the imperfect CSI case versus $p_{\max}$ with different value of error bound $\delta_h = \delta_g = \delta = 0, 0.02, 0.04, 0.08$. When $p_{\max}$ increases, the performance of the proposed NOMA increases and saturates at high value of $p_{\max}$. This is obvious because

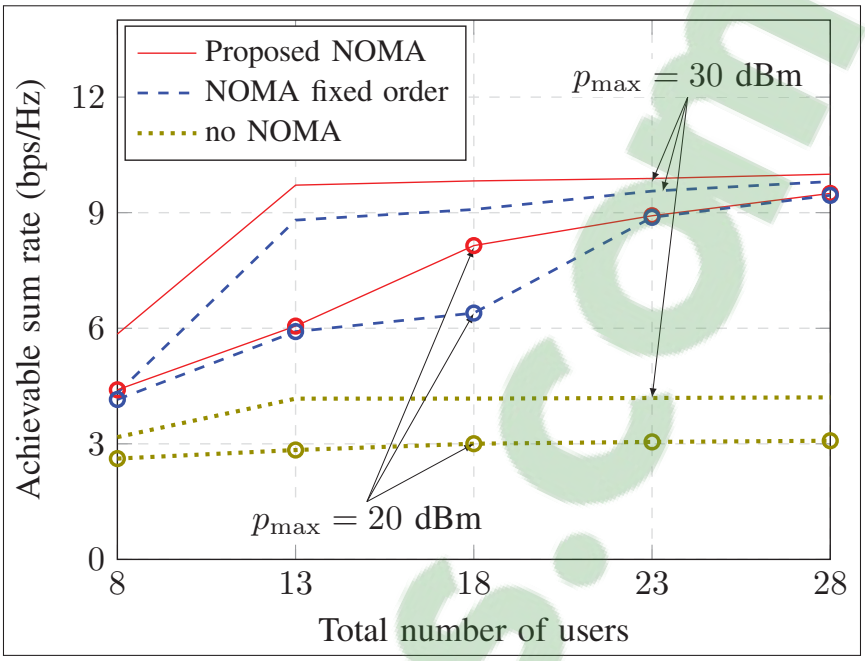

Figure 5.7 Throughput performance of the proposed NOMA scheme with respect to total user number.

when each SAP has higher power budget $p_{\max}$, it will transmit at higher power to achieve more throughput. However, due to the interference from concurrent SAP transmissions, these SAPs will not transmit at it maximum budget power $p_{\max}$ so as to maintain the maximal total throughput, which results in the performance saturation. We observe that the achieved throughput performance in the imperfect CSI case is less than the achieved throughput in the perfect CSI case. This is because the robust optimization problem in (5.35) is solved using the erroneous estimated CSI. Furthermore, when the CSI error grows, the achieved throughput reduces more.

In Fig. 5.9, we show the total number of satisfied MUEs and SUEs with respect to the minimum SUE rate requirement $\gamma_{\min}$, where we choose $\Gamma_{\min} = 1.5$ bps/Hz. In this figure, we compare the performance between the proposed NOMA and the conventional "no NOMA" schemes (c.f. (Wang *et al.*, 2016; Zhao *et al.*, 2015; Nguyen *et al.*, 2016b)). Note that the problem of maximizing the satisfied user number under the conventional "no NOMA" scheme is similar to problem $(\mathscr{P}_2)$, where where the backhaul, MUE, and SUE rates take the formulas in (5.16).

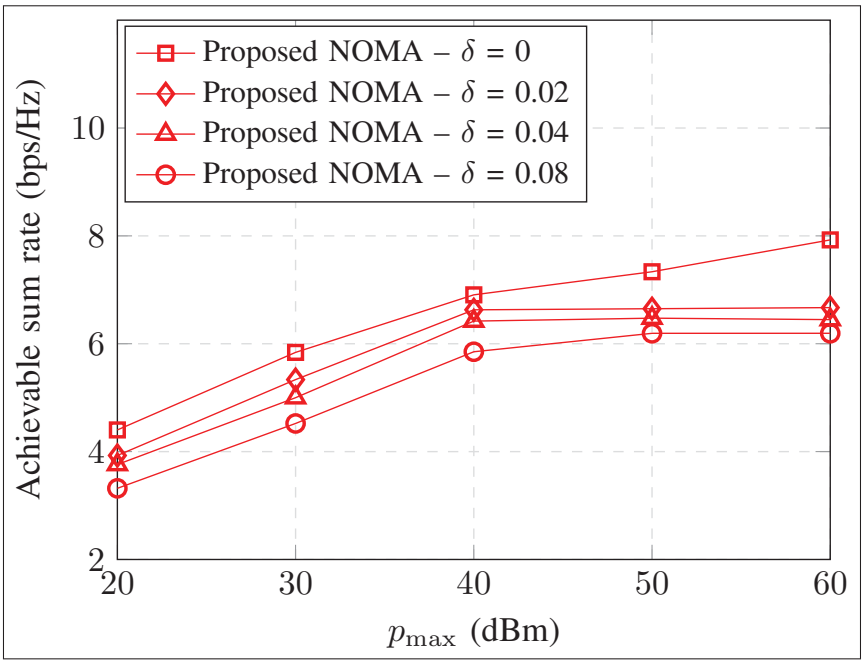

Figure 5.8 Comparison of the achieved throughput with respect to $p_{\max}$ when $\delta = 0, 0.02, 0.04, 0.08$.

The results in this figure are obtained by applying Algorithm 6 on problem $(\mathscr{P}_2)$ until convergence. We observe that the proposed NOMA scheme always provides more number of satisfied user compared to the conventional no NOMA scheme, which again justifies the usefulness of our NOMA strategy not only in boosting the throughput metrics but also in admitting more qualified served users. When $\gamma_{\min}$ increases, the number of satisfied users decreases. This is apparent because when the SUEs are obliged to achieve higher access rate, we must schedule to deny service to some insignificant SUEs in order to reduce the concurrent transmissions. This alleviates the interference introduced to the served SUEs possessing the better channel conditions, which in turn improves their achievable rates.

In Fig. 5.10, we show the number of satisfied MUEs and SUEs with respect to both the SUE minimum rate requirement $\gamma_{\min}$ and the maximum power budget $P_{\max}$, where we choose $\Gamma_{\min} =$ 1.5 bps/Hz. At high $P_{\max}$ and low $\gamma_{\min}$, we observe that all the MUEs and SUEs are active since the the available power budget is sufficient to guarantee the satisfaction of each user. However, when $\gamma_{\min}$ increases, the number of satisfied SUEs and MUEs decreases. This can be explained

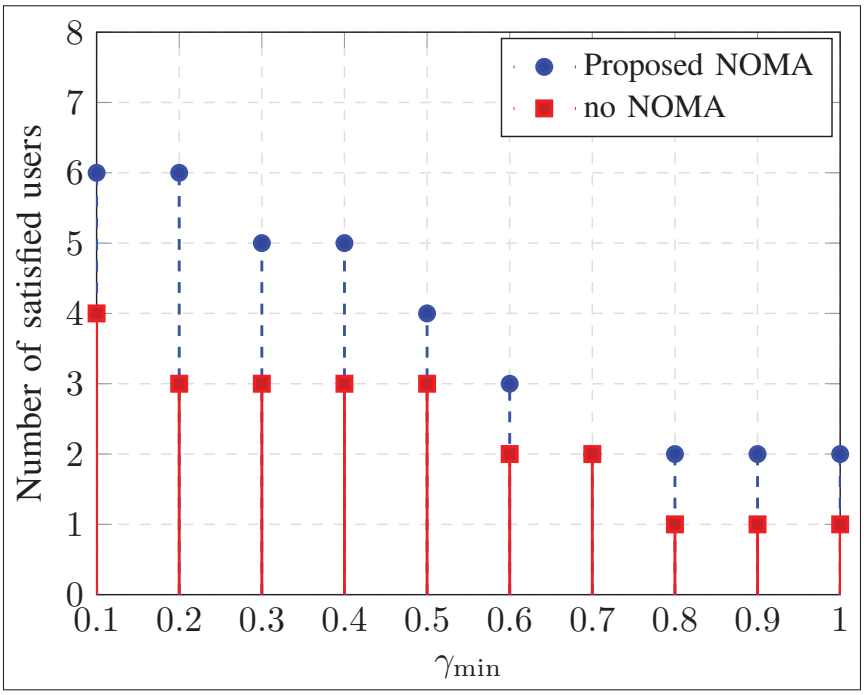

Figure 5.9 Number of satisfied users achieved at different schemes with respect to $\gamma_{\min}$.

similar to Fig. 5.9. Moreover, when $P_{\max}$ decreases, the number of satisfied SUEs and MUEs also reduces since with less power budget to support the WB communications, the achievable rates at some WB links are insufficient to satisfy the minimum rate requirement of the related small cell WA links, which results in the less number of active users.

In Fig. 5.11, we compare the performance of our proposed low-complexity Algorithm 6 with the global optimal branch-and-bound (BnB) algorithm developed in (Tervo *et al.*, 2015b) for Scenario 1 with $N = 4$, $S = 4$ and $M = 2$. According to the characteristic of BnB algorithm described in (Tervo *et al.*, 2015b), the BnB algorithm repeatedly operates the following steps: (i) partitioning the set which contains the global optimal solution into smaller sets, (ii) checking the feasibility of the newly branched sets, and (iii) updating the terminating criteria of the BnB algorithm until convergence. The BnB algorithm is terminated when the difference between upper and lower bound, as can be observed from Fig. 5.11, is below a threshold of 10^{-3}. Note that BnB algorithm is indeed an exhaustive search based algorithm, which has prohibitively

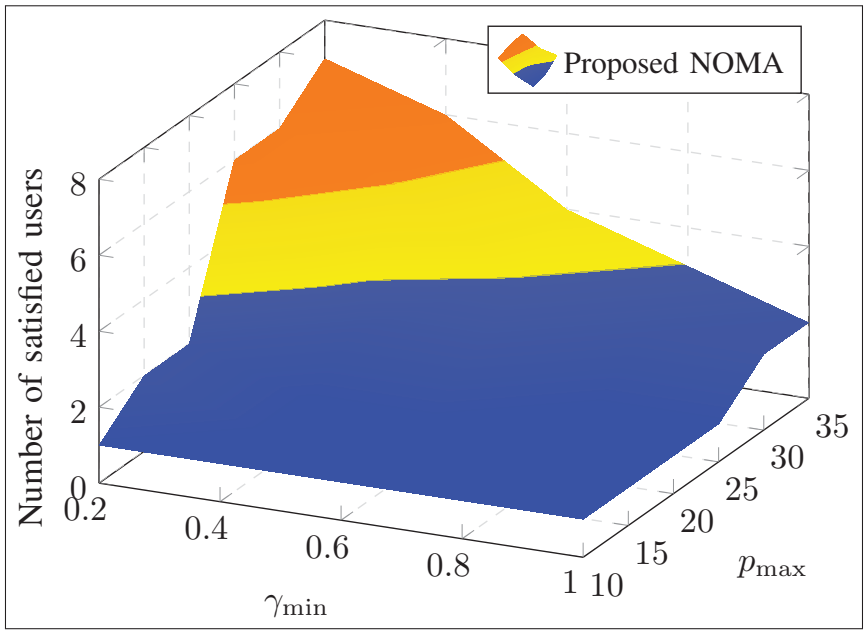

Figure 5.10 Number of satisfied users achieved at the proposed NOMA scheme with respect to $P_{\max}$ and $\gamma_{\min}$.

high computational complexity. Therefore, we choose to show the performance of this BnB algorithm under small network setup. For Algorithm 6, we examine its convergence behavior with two different set of initial points. We observe that these two curves of Algorithm 6 converge less than 20 iterations. However, the BnB algorithm keeps updating the upper and lower bound value and requires an excessive number of iteration to converge. We observe that the results from low-complexity algorithm are very close to the baseline optimal value obtained from the BnB algorithm. We also show the convergence of Algorithm 6 on Scenario 2 with $N = 4, S = 6, M = 2$. Due to the large network size, we skip the result of the BnB algorithm for Scenario 2. Again, we observe that the objective value (5.31a) resulted from Algorithm 6 finally converges, which again validates the convergence analysis in Section 5.4.3.

Fig. 5.12 shows the convergent behavior of Algorithm 6 applying to solve $(\mathscr{P}_2)$ on Scenario 2 with $N = 4, S = 6, M = 2$. We also show the convergence of BnB algorithm employed to solve $(\mathscr{P}_2)$ under Scenario 2. The BnB algorithm checks all binary combinations of $\{\mathbf{x}, \mathbf{y}\}$ for $(\mathscr{P}_2)$ and verifies the feasibility of each combination. As shown in Fig. 5.12, the result of 8

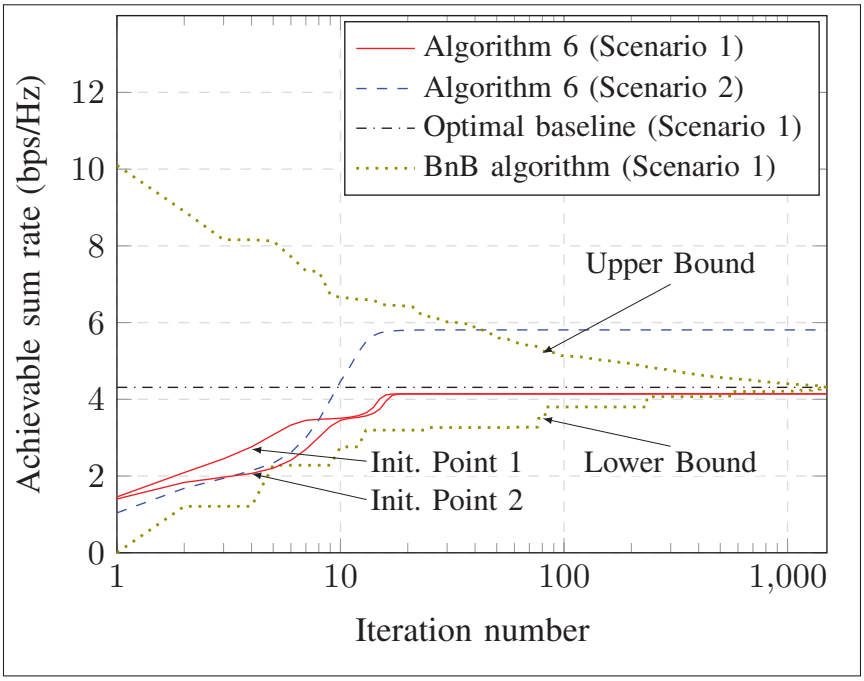

Figure 5.11 Convergence between the optimal branch-and-bound (BnB) algorithm and Algorithm 6 for (5.13).

active users (including SUEs and MUEs) is infeasible, which is shown by the cross marker. However, when examining the combinations of 7 active users, there are 1 feasible case, which are illustrated by the green circle markers. The BnB algorithm can exhaustively search to result in 7 maximum active users, while Algorithm 6 only achieve lower result with 6 maximum active users. However, BnB algorithm requires much more computation to solve the feasibility checking when each binary combination of $\{\mathbf{x}, \mathbf{y}\}$ is fixed, and this task is more excessive when the network size grows.

5.6 Concluding Remarks

We proposed a novel transmission scheme based on cooperative NOMA to aim at redesigning the WB two-tier HetNets system. By employing the CoTDD combined with spectrum partitioning scheme, we introduced two optimization problems which jointly design the NOMA decoding order together with the downlink transmit beamforming and power allocation at the

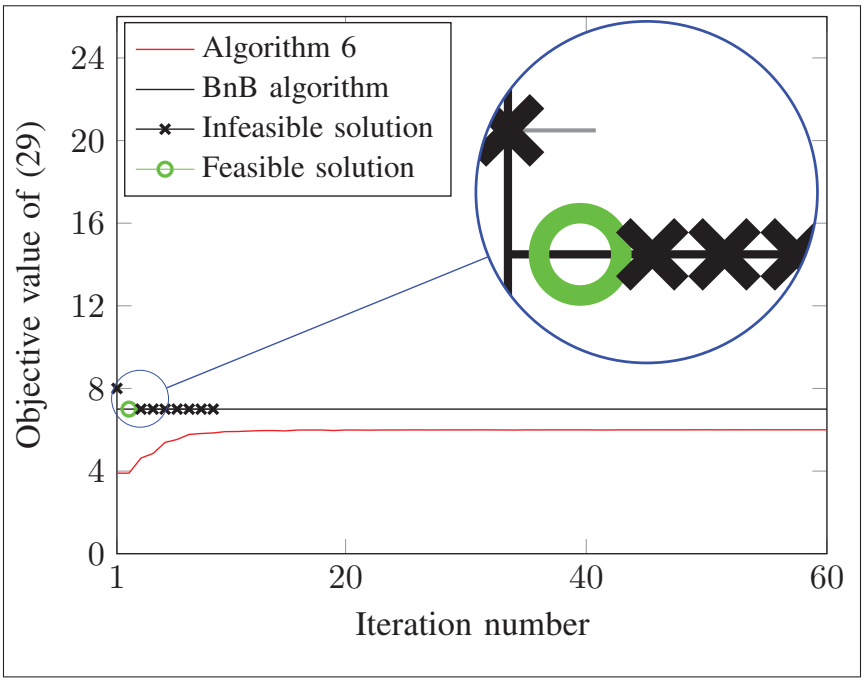

Figure 5.12 Convergence between the optimal branch-and-bound (BnB) algorithm and Algorithm 6 for (5.14).

MBS and SAP to maximize the two objectives. The formulated problems are generally NP-hard. To efficiently solve them, we first employed the DC functions to present the formulated binary variables and then equivalently transformed the optimization problems into more tractable forms. Finally, we developed an iterative low-complexity algorithm based on SCA technique and MMM to compute the sub-optimal solution, which is provable to eventually converge at a sub-optimal solution. Numerical results highlight that our proposed scheme when exploiting the novel cooperative transmissions as a result of the MISO-NOMA is more advanced and effective that the scheme which straightforwardly applies the MISO-NOMA from the literature. These achieved numerical results corroborate that our proposed scheme also outperforms the conventional designs in terms of total achievable rate.

CHAPTER 6

A NOVEL COOPERATIVE NOMA FOR DESIGNING UNMANNED AERIAL VEHICLE (UAV)–ASSISTED WIRELESS BACKHAUL NETWORKS

Tri Minh Nguyen[1], Wessam Ajib[2], Chadi Assi[3]

[1] Département de Génie Électrique, École de Technologie Supérieure,
[2] Département d'Informatique, Université de Québec à Montréal
[3] Concordia Institute for Information Systems Engineering, Concordia University

This article was published in *Journal on Selected Areas on Communications* on November 2018
(Nguyen *et al.*).

6.1 Introduction

In recent years, owing to their benefit, dense heterogeneous networks (HetNets) (Ghosh *et al.*, 2012) have constantly been promoted as a prime candidate technology for the evolution of wireless networks towards 5G (Andrews *et al.*, 2014). However, dense HetNets presented a tremendous economical drawback due to their expensive cost of installing and maintaining fiber backhaul infrastructure, especially at some hard-to-reach areas. To effectively exploit the benefits of HetNets within a reasonable budget, it is essential to replace the wired connections for backhauling data into the core with the wireless ones (Siddique *et al.*, 2015a).

Nonetheless, the proposed wireless backhaul (WB) solution could prove unable to bridge the connectivity to isolated users, especially, in disaster-prone situations. Therefore, rather than deploying small cell base stations, a more practical, cost-saving, and efficient solution is to dispatch unmanned aerial vehicles (UAVs) as flying small cell base stations (Zeng *et al.*, 2016a) which can communicate with ground users and backhaul the data to the core network (Zeng *et al.*, 2016b; Zeng & Zhang, 2017). While stimulating more opportunities, using UAVs to cope with the structure of point-to-multi-point (PtMP) WB must tackle the following inherent challenges. First, deploying WB links concurrently with the existing wireless access (WA) links renders the interference more severe (Nguyen *et al.*, 2016b). Second, each small cell

UAV cannot transmit more data than what it received from the core network (Siddique *et al.*, 2015a). Third, since UAV can ubiquitously move in space, a change of coordinates can lead to different channel gain realizations and modify the system performance. Thus, considering UAV-assisted WB networks must engineer a proper resource allocation design considering the above characteristics to fully reap their potential performance gains.

6.1.1 Related Work

The concept of WB for *IEEE* 802.16 mesh networks was investigated in (Viswanathan & Mukherjee, 2006). The authors used linear programming to maximize the network throughput by optimizing the routing and scheduling strategy. More generally, WB for 5G was extensively considered in different network scenarios and, interestingly, on different spectrum bands (Siddique *et al.*, 2015a). The authors in (Hur *et al.*, 2013b) presented a novel idea for mmWave small cells to efficiently align their transmit beamformer under wind induced impairments effect. WB can also be operated in the commercial sub–6 GHz band (Siddique *et al.*, 2015a) concurrently with the WA ones. In (Wang *et al.*, 2016), the authors employed the reverse time division duplex interference management and massive multiple input multiple output (MIMO) to jointly optimize the bandwidth allocation and user association. Siddique *et al.* (2017) introduced a hybrid strategy of out-of-band and in-band full-duplex for the considered small cells and formulated a problem which optimizes the spectrum allocation.

UAV (or drone) communications represent the data transmissions/receptions between the ground controller and the air-borne UAVs (Frew & Brown, 2008). UAVs used for communications are more advanced than conventional UAVs which are only commanded from ground controller for tasks such as transportation, infiltration, rescuing (Frew & Brown, 2008; Gupta *et al.*, 2016). They are expected to extend the coverage, guarantee more ubiquity, and support higher transmission rate to satisfy the user's experience. Related work on UAVs claimed that these gains can be obtained by optimizing the ubiquitous movement of UAVs jointly with the radio resources to reap the best network performance (Mozaffari *et al.*, 2017). Beside, investigating the UAV performance under various network settings such as device-to-device UAV (Mozaf-

fari *et al.*, 2016), UAV-assisted relay (Zeng *et al.*, 2016b), task offloading with UAV in mobile edge computing (Jeong *et al.*, 2018), coverage extension with UAV (Mozaffari *et al.*, 2017) were investigated. More particularly, in (Zeng *et al.*, 2016b), the authors optimized the transmit power along with the trajectory of the moving relay UAV to maximize the throughput and the energy-efficiency. In (Wu *et al.*, 2018), the authors modelled multi-UAV serving multi-user and formulated a problem of power allocation together with user association to maximize the minimum throughput. Despite its promising feature, it is still unclear how to interplay the role of UAV with existing WB infrastructure. When the question of poor scattering line-of-sight (LOS) channel condition between the macrocell base station (MBS) and UAVs is anticipated, the benefit of beamforming is now reduced to power allocation and is under-explored to retrieve the expected spectral efficiency. Without exploiting the multipath channel characteristic to beamform data to multiple receivers, applying a transmission scheme such as non-orthogonal multiple access (NOMA) is more effective to leverage the UAV's benefit.

Recently, NOMA (Saito *et al.*, 2013) was proposed as a pragmatic scheme to effectively handle simultaneous transmissions. Instead of separating transmissions in orthogonally split resources, NOMA consolidates the concurrent transmissions' benefit with improved spectral efficiencies by using a sophisticated successive interference cancellation (SIC) technique at the receiver side. A report on a downlink NOMA for 5G networks were surveyed in (Wei *et al.*, 2016). In general, by simultaneously exploiting the multiplexing transmission, SIC receptions, and efficient network resource allocation, many works have been able to highlight the NOMA's superiority compared to conventional designs. For instance, the network performance using NOMA under a random user deployment was studied in (Ding *et al.*, 2014). Power allocation and NOMA decoding order were jointly optimized in (Shi *et al.*, 2016) in an outage balancing downlink NOMA system. Tabassum *et al.* (2017) analyzed the impact of imperfect SIC on the network performance. Cui *et al.* (2018) studied the use of NOMA in multiple input single output (MISO) mmWave systems. On the other hand, a cooperative NOMA scheme was investigated in (Ding *et al.*, 2015). In (Hanif *et al.*, 2016), the authors formulated a problem of designing beamforming in the MISO NOMA to maximize the sum spectral efficiency.

6.1.2 Motivations and Contributions

In this paper, we study the downlink of UAV-assisted PtMP WB networks under a novel cooperative NOMA scheme. In one time slot, the MBS superimposes the WB messages to transmit to the UAVs, and the UAVs employ the SIC to decode the message depending on the determined decoding order. In the next time slot, we exploit the benefit that SIC allows some UAVs to decode the messages of other UAVs to enable the UAV's cooperation to transmit to the users. Based on this scheme, our contribution is as follows:

- Since CSI varies with the UAV's position, we jointly optimize the UAVs' positions, the transmit beamformer at the MBS and UAVs, and the decoding order of the NOMA-SIC so as to maximize the user sum rate.
- The formulated problem is a general mixed-integer non-linear programm, which is very difficult to solve optimally. Thus, we propose a framework based on difference of convex (DC) (Kha *et al.*, 2011) method characterized by the Lipschitz continuity (Parikh & Boyd, 2014a) to approximate the original problem into a series of convex approximate ones, and develop a low-complexity algorithm to sequentially solve for each approximate problem until convergence.

The rest of this chapter is as follows. Section 6.2 introduces the system model. In Section 6.3, we formulate mathematically our resource allocation as an optimization problem. The solution approach is given in Section 6.5. Section 6.6 presents and discusses our numerical results under different simulation setups. Finally, the concluding remarks are given in Section 6.7.

6.2 System Model

6.2.1 Spatial Model

We consider the downlink of a WB network comprising of one MBS, equipped with N sectorized antennas to split the coverage into 3 non-overlapped sub-sectors. The signals transmitted from each MBS's sub-sector are orthogonal with each other. Assuming that in each sector n,

where $n = 1,2,3$, there is a set $\mathscr{U}_n = \{1,\ldots,U\}$ of UAVs and clusters of users, where a UAV or a user has single antenna each. UAVs can wirelessly receive backhaul data from the MBS. Beside, each UAV is set to serve a distinct cluster of users based on a predetermined agreement between the operator and that cluster's representative. The MBS and each j^{th} user belonging to the i^{th} cluster in sector n are spatially placed at some fixed grounded positions, denoted by $\{\bar{x}_b, \bar{y}_b, H_b\}$ and $\{\bar{x}_{ni}^j, \bar{y}_{ni}^j, 0\}, \forall i \in \mathscr{U}_n$, respectively. On the other hand, each UAV can flexibly fly to any position in space. We later aim at determining each i^{th} UAV's position in each sector n, which is represented as $\mathbf{p}_{ni} = \{x_{ni}, y_{ni}, H\}$, where the height H is assumed constant (Wu *et al.*, 2018). We consider that user cannot be directly served by the MBS due to some geographical difficulties. For simplicity, we omit the service from MBS to macrocell users.

We assume that WB and WA communications are accommodated on a spectrum band of W Hertz within the sub–6 GHz band in consecutive time slots as depicted in Fig. 6.1. In the first time slot, the MBS transmits backhaul data to the UAVs. In the second time slot, all the UAVs simultaneously transmit the decoded backhaul data to their destinations. Here, we assume that each i^{th} UAV is scheduled to serve only one user belonging to the i^{th} cluster, where the remaining users are scheduled to be served later. For simplicity, we denote the user intended for the i^{th} UAV is the i^{th} user. We can also schedule each UAV to serve multi-user by orthogonally splitting the spectrum and accommodate users on each partition to avoid intra-cluster interference. In this way, determining UAV's positions should be optimized to provide service to multi-user at each time instant. For brevity, we assume that the scheduling is decided prior to our resource allocation problem and each i^{th} user's position in sector n is denoted as $\{\bar{x}_{ni}, \bar{y}_{ni}, 0\}, \forall i \in \mathscr{U}_n$.

6.2.2 Channel Model

We assume the fading channel remains unchanged within a coherence time. Following (Lyu *et al.*, 2018), we denote h_{nj} as the ground-to-air (GtA) channel, consisting of path-loss and small scale fading, from the MBS to the j^{th} UAV in sector n. Similarly, we denote g_{mnjk} as the air-to-ground (AtG) channel from the j^{th} UAV in sector m to the k^{th} user in sector n. h_{nj} and

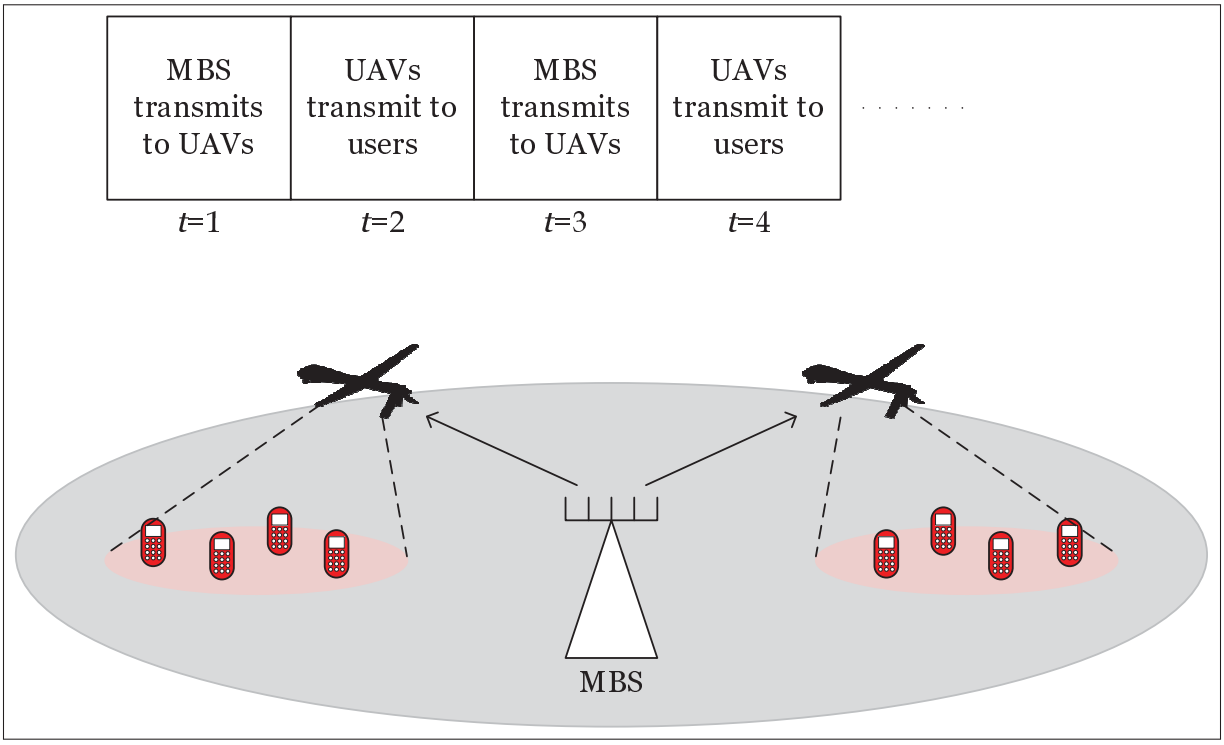

Figure 6.1 Spatial model of the UAV-assisted WB networks.

g_{mnjk} can be written as

$$h_{nj} = \delta_{nj}\tilde{h}_{nj} \tag{6.1}$$

$$g_{mnjk} = \Delta_{mnjk}\tilde{g}_{mnjk} \tag{6.2}$$

where each component $\tilde{h}_{nj}, \delta_{nj}$ respectively represents the small scale fading and path-loss coefficient between the MBS and the j^{th} UAV in sector n. Moreover, $\tilde{g}_{mnjk}$, Δ_{mnjk} respectively represent the small scale fading and path-loss coefficient between the j^{th} UAV in sector m and the k^{th} user in sector n. In particular, we can write

$$\delta_{nj} = [(\bar{x}_b - x_{nj})^2 + (\bar{y}_b - y_{nj})^2 + (H_b - H)^2]^{-1/2} \tag{6.3}$$

$$\Delta_{mnjk} = [(\bar{x}_{nk} - x_{mj})^2 + (\bar{y}_{nk} - y_{mj})^2 + H^2]^{-1/2} \tag{6.4}$$

Since the GtA and AtG channel are often governed by the LoS propagation, $\tilde{h}_{nj}$ (or $\tilde{g}_{mnjk}$) can follow a Rician distribution of factor K_b (or K_u), which consists of a deterministic LoS component $\bar{h}_{nj}$ (or $\bar{g}_{mnjk}$) with $|\bar{h}_{nj}| = 1$ (or $|\bar{g}_{mnjk}| = 1$) and a random scattered component $\hat{h}_{nj}$ (or $\hat{g}_{mnjk}$) as

$$\tilde{h}_{nj} = \sqrt{K_b/(1+K_b)}\bar{h}_{nj} + \sqrt{1/(1+K_b)}\hat{h}_{nj} \tag{6.5}$$

$$\tilde{g}_{mnjk} = \sqrt{K_u/(1+K_u)}\bar{g}_{mnjk} + \sqrt{1/(1+K_b)}\hat{g}_{mnjk} \tag{6.6}$$

where each of $\hat{h}_{nj}, \hat{g}_{mnjk}$ follows a complex Gaussian distribution $\mathscr{CN}(0,1)$.

6.2.3 Transmission Model

6.2.3.1 Transmissions from MBS to UAVs

In sector n, we employ NOMA technique (Sun *et al.*, 2017b) to let the MBS superimpose the messages of all UAVs and transmit. Under this strategy, the received signal at the j^{th} UAV in sector n is given by

$$y_{nj} = h_{nj}\sqrt{w_{nj}}s_{nj} + h_{nj}\sum_{l\in\mathscr{U}_n\setminus j}\sqrt{w_{nl}}s_{nl} + \tilde{n}_{nj}, \tag{6.7}$$

where $w_{nj} \in \mathbb{R}^+$ is the power from the MBS to the j^{th} UAV in sector n, s_{nj} is the message for the j^{th} UAV in sector n with unit average power, e.g., $\mathbb{E}\left\{s_{nj}s_{nj}^*\right\} = 1$, and $\tilde{n}_{nj}$ as the AWGN at the j^{th} UAV, which is distributed according to $\mathscr{CN}(0,N_0)$, where N_0 is the noise power.

At the receiver side, each UAV applies the SIC to detect its message. Depending on the given decoding order, the j^{th} UAV might decode some messages intended for the other i^{th} UAV, for $i \in \mathscr{U}_n$, before decoding its own message. Here, we assume perfect SIC, where the decoding processing is always successful. We observe hence that the decoding order is coupled with the UAV's positions. To allow the determination of an appropriate decoding order, let us introduce binary variable $a_{n,i,j} \in \{0,1\}$, where $a_{n,i,j} = 0$ implies that the j^{th} UAV can decode the i^{th}

UAV's message s_{ni} in sector n and subtract it out of the total received signal; and $a_{n,i,j} = 1$ otherwise. Therefore, the achievable rate, computed in bps/Hz, after applying SIC at the j^{th} UAV to detect messages s_{nj} is $R_{n,j,j}\left(\mathbf{a}_n, \mathbf{w}_n, \delta_{nj}\right) = \log\left(1 + \mathsf{SINR}_{n,j,j}\right)$, where

$$\mathsf{SINR}_{n,j,j} = \frac{w_{nj}\left|\tilde{h}_{nj}\delta_{nj}\right|^2}{\sum\limits_{l \in \mathscr{U}_n \setminus j} a_{n,l,j} w_{nl}\left|\tilde{h}_{nj}\delta_{nj}\right|^2 + N_0} \tag{6.8}$$

and $\mathbf{w}_n = \left\{w_{nj}, \forall j \in \mathscr{U}_n\right\}$, $\mathbf{a}_n = \left\{a_{n,l,j}, \forall l, j \in \mathscr{U}_n\right\}$. Let us denote $\mathbf{w} = \{\mathbf{w}_n, \forall n\}$ and $\mathbf{a} = \{\mathbf{a}_n, \forall n\}$. Note that $a_{n,j,j} = 0, \forall j \in \mathscr{U}_n$ and $a_{n,j,i} + a_{n,i,j} = 1, \forall i, j \in \mathscr{U}_n, i \neq j$. In addition, in sector n, if the j^{th} UAV can decode the i^{th} UAV's message s_{ni}, $\forall i \neq j$, the following constraints must hold

$$\delta_{nj}|\tilde{h}_{nj}| \geq (1 - a_{n,i,j})\delta_{ni}|\tilde{h}_{ni}| \tag{6.9}$$

$$\delta_{nj}|\tilde{h}_{nj}| \leq (1 - a_{n,i,j})M + \delta_{ni}|\tilde{h}_{ni}|. \tag{6.10}$$

where $M \gg 1$. The characteristic of (6.9)–(6.10) can be explained as follows. If the j^{th} UAV can decode the i^{th} UAV's message, we have $a_{n,i,j} = 0$ so that from (6.9)–(6.10), we have $\delta_{nj}\left|\tilde{h}_{nj}\right| \geq \delta_{ni}\left|\tilde{h}_{ni}\right|$ and $\delta_{nj} \leq M + \delta_{ni}|\tilde{h}_{ni}|$. Reversely, if the i^{th} UAV can decode the j^{th} UAV's message, we have $a_{n,i,j} = 1$ so that from (6.9)–(6.10), we have $\delta_{nj}\left|\tilde{h}_{nj}\right| \geq 0$ and $\delta_{nj}\left|\tilde{h}_{nj}\right| \leq \delta_{ni}\left|\tilde{h}_{ni}\right|$.

6.2.3.2 Novel Cooperative Access Transmissions from UAVs to Users

The details of the proposed cooperative transmission scheme can be elaborated by analyzing the following example, as illustrated in Fig. 6.2. *An example case for $U_n = 3, \forall n$*: We assume that in each sector n, the UAV decoding order is ranked 1, 2, and 3. Following this, UAV 1 decodes only s_{n1}, UAV 2 decodes s_{n1} and s_{n2} and UAV 3 decodes s_{n1}, s_{n2}, and s_{n3}. Here, the cooperation between UAVs is that UAVs 2 and 3 can simultaneously transmit message s_{n1}, in addition to their own messages s_{n2} and s_{n3}, respectively, to help UAV 1 improve' the signal reception at user 1. Similarly, UAV 3 can cooperate with UAV 2 to send s_{n2} in addition to s_{n3} and s_{n1}. By assuming that $\mathbf{v}_{n1} = [v_{n11}, v_{21}, v_{31}], \mathbf{v}_{n2} = [v_{n12}, v_{n22}, v_{n32}], \mathbf{v}_{n3} = [v_{n13}, v_{n23}, v_{n33}]$

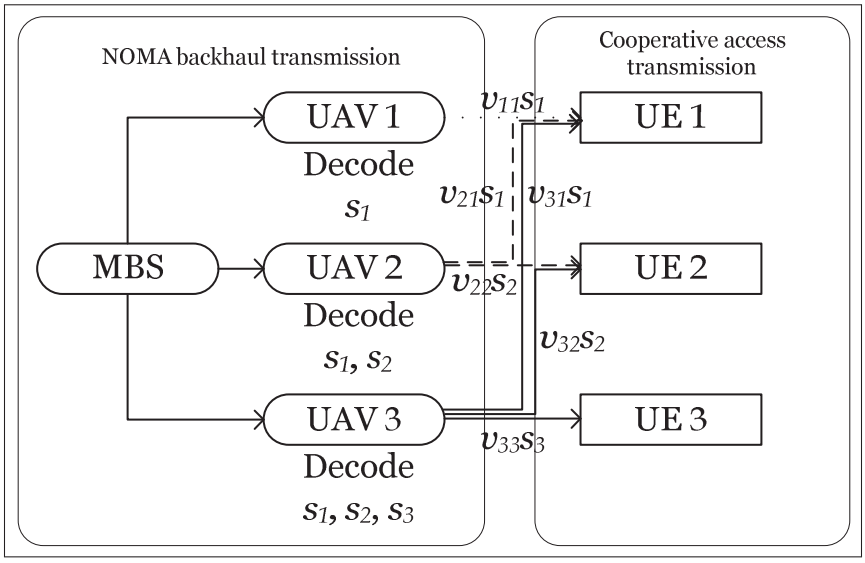

Figure 6.2 Proposed cooperative NOMA scheme.

are the transmit beamformers to users 1, 2, and 3 in sector n, respectively, the received signal at each user is

$$y_{ni} = \mathbf{v}_{ni}^H \mathbf{g}_{nni} s_{ni} + \sum_{k \neq i} \mathbf{v}_{nk}^H \mathbf{g}_{nni} s_{nk} + \sum_{m \neq n} \sum_{l \in \mathscr{U}_m} \mathbf{v}_{ml}^H \mathbf{g}_{mni} s_{ml} + \tilde{n}_{ni}, \tag{6.11}$$

where $\mathbf{g}_{mni} = [g_{mn1i}(\Delta_{mn1i}), g_{mn2i}(\Delta_{mn2i}), g_{mn3i}(\Delta_{mn3i})]^T$ is the channel vector from all the UAVs in sector m to the i^{th} user in sector n. s_{ni} is the message for the i^{th} user in sector n with unit average power, e.g., $\mathbb{E}\{s_{ni} s_{ni}^*\} = 1$, $\tilde{n}_{ni} \sim \mathscr{CN}(0, N_0)$ is an AWGN at the i^{th} user in sector n. Note that we associate the cooperation capability at each UAV by coupling the binary variable $a_{n,i,j}$ with the weight v_{nij} using the following inequalities

$$\begin{cases} \left|v_{nij}\right|^2 \leq (1 - a_{n,j,i}) \lambda_{nij} & ,\forall i, j = 1,2,3, \\ \sum_{j=1}^{3} \lambda_{nij} \leq p^{\max} & ,\forall i = 1,2,3, \end{cases} \tag{6.12}$$

where $\lambda_{nij} \in \mathbb{R}^+$ represents the soft power level (c.f. (Tervo *et al.*, 2015b)) that UAV i applies at the j^{th} user. In our example, $a_{n,2,1} = a_{n,3,1} = a_{n,3,2} = 1$ where the remaining elements of $\mathbf{a}_n$

are 0. This means that $v_{n12}, v_{n13}, v_{n23}$ are enforced to 0, where the remaining elements of each vector $\mathbf{v}_{ni}, \forall i = 1,2,3$ can take any complex value whose norm is from 0 to λ_{nij}.

In the general case, let us denote $\mathbf{v}_{ni} = [v_{n1i}, \ldots, v_{nUi}]^T$, $\mathbf{v} = \{\mathbf{v}_{ni}, \forall i \in \mathscr{U}_n, \forall n = 1,2,3\}$ and $\mathbf{g}_{mni}(\Delta_{ni}) = [g_{mn1i}(\Delta_{mn1i}), \ldots, g_{mnUi}(\Delta_{mnUi})]^T$ and $\mathbf{g}_{ni}(\Delta_{ni}) = [\mathbf{g}_{1ni}(\Delta_{1ni}), \mathbf{g}_{2ni}(\Delta_{2ni}), \mathbf{g}_{3ni}(\Delta_{3ni})]^T$ as the arrays of weights and channel vectors from all the UAVs to the i^{th} user. Treating interference as noise, the achievable rate $r_{ni}(\mathbf{v}, \Delta_{ni})$ at the i^{th} user in sector n is $r_{ni}(\mathbf{v}, \Delta_{ni}) = \log(1 + \mathsf{SINR}_{ni})$, where

$$\mathsf{SINR}_{ni} = \frac{\left|\mathbf{v}_{ni}^H \mathbf{g}_{nni}(\Delta_{nni})\right|^2}{\sum_{(m,l)\neq(n,i)} \left|\mathbf{v}_{ml}^H \mathbf{g}_{mni}(\Delta_{mni})\right|^2 + N_0} \tag{6.13}$$

6.3 Mathematical Problem Formulation

Observing that the achievable rate of each user depends on the design of transmit beamforming $\mathbf{w}$, $\mathbf{v}$, λ the decoding order $\mathbf{a}$, the placement of each UAV $\mathbf{p}$, and the path-loss factor δ and Δ, we aim at formulating a problem which jointly optimizes $\mathbf{w}, \mathbf{v}, \lambda, \mathbf{a}, \mathbf{p}$, δ, and Δ to maximize the user sum rate. The optimization problem can be formulated as

$$\max_{\substack{\mathbf{a},\lambda,\mathbf{w},\mathbf{v},\\ \mathbf{p},\delta,\Delta}} \sum_{n=1}^{3} \sum_{j\in\mathscr{U}_n} r_{nj}(\mathbf{v}, \Delta_{nj}) \tag{6.14a}$$

$$\text{s.t.} R_{n,j,j}\left(\mathbf{a}_n, \mathbf{w}_n, \delta_{nj}\right) \geq r_{nj}(\mathbf{v}, \Delta_{nj}) \tag{6.14b}$$

$$\delta_{nj}\left|\tilde{h}_{nj}\right| \geq \left(1 - a_{n,i,j}\right)\delta_{ni}\left|\tilde{h}_{ni}\right| \tag{6.14c}$$

$$\delta_{nj}\left|\tilde{h}_{nj}\right| \leq \left(1 - a_{n,i,j}\right)M + \delta_{ni}\left|\tilde{h}_{ni}\right| \tag{6.14d}$$

$$\delta_{nj}^{-1} \geq \sqrt{(\bar{x}_b - x_{nj})^2 + (\bar{y}_b - y_{nj})^2 + (H_b - H)^2} \tag{6.14e}$$

$$\Delta_{nmji}^{-1} \geq \sqrt{(x_{nj} - \bar{x}_{mi})^2 + (y_{nj} - \bar{y}_{mi})^2 + H^2} \tag{6.14f}$$

$$\sum_{n=1}^{3} \sum_{j\in\mathscr{U}_n} |w_{nj}|^2 \leq P^{\max} \tag{6.14g}$$

$$\left|v_{nij}\right|^2 \leq (1 - a_{n,j,i})\lambda_{nij}; \sum_{j\in\mathscr{U}_n} \lambda_{nij} \leq p^{\max} \tag{6.14h}$$

$$a_{n,i,j} + a_{n,j,i} = 1; a_{n,j,j} = 0; a_{n,i,j} \in \{0,1\} \tag{6.14i}$$

for $i, j \in \mathscr{U}_n, i \neq j$ and $n = 1,2,3$. Here, constraints (6.14b) ensure that the WB rate at the j^{th} UAV be greater than or equal to the WA rate for the j^{th} user in sector n (Wang *et al.*, 2016, Eq. (13)). (6.14b) also implies a complicated coupling between variables $\mathbf{w}, \mathbf{v}, \mathbf{a}$, δ, and Δ by non-concave non-convex functions, i.e., $R_{n,j,j}\left(\mathbf{a}_n, \mathbf{w}_n, \delta_{nj}\right), r_{nj}(\mathbf{v}, \Delta_{nj}), \forall j \in \mathscr{U}_n, \forall n$. Note that (6.14e) and (6.14f) consider the relaxed version of the relationship in (6.14e) and (6.14f) to reduce the complexity of the formulated problem. Indeed, (6.14e) and (6.14f) become equality at optimality (which can be proved similar to Appendix I in (Nguyen *et al.*, 2016b)). (6.14g) and (6.14h) are the maximum power budget at the MBS and each UAV. To reflect the impact of decoding order, we also present the binary constraints (6.14i).

We first observe that solving (6.14) requires the knowledge of CSI, e.g., $\tilde{h}_{ni}, \tilde{\mathbf{g}}_{ni}, \forall i \in \mathscr{U}_n, \forall n = 1,2,3$. However, CSI varies when the location of the UAV changes. Since we cannot determine the location of UAVs initially, we cannot obtain the accurate CSI, and thus cannot properly solve (6.14). To overcome this obstacle, in the following, we first propose an approach to decouple the design of UAV's location $\mathbf{p}, \delta, \Delta$ from the remaining variables $\mathbf{a}, \mathbf{w}, \mathbf{v}, \lambda$. To serve the practicability of this method, we assume that the MBS can store a collection of past CSI, denoted by $\tilde{h}_{ni,t}, \tilde{\mathbf{g}}_{ni,t}, \forall t = 1, \ldots, T$ by collecting the copy of each obtained CSI starting from the T^{th} previous time slot. We use these past CSI to first solve for the UAV's location apart from $\mathbf{a}, \lambda, \mathbf{w}, \mathbf{v}$. Then, we command the UAV to stay at the determined positions to attain the real CSI (via training) and proceed to solve for $\mathbf{a}, \lambda, \mathbf{w}, \mathbf{v}$.

6.4 Proposed Solution Approach

As discussed above, we assume that the MBS has a collection of past CSI $\tilde{h}_{ni,t}, \tilde{\mathbf{g}}_{ni,t}, \forall t = 1, \ldots, T$. Using these past CSI, we solve for the location of UAVs $\mathbf{p}, \delta, \Delta$ jointly with the variables $\tilde{\mathbf{A}}, \tilde{\mathbf{W}}, \tilde{\mathbf{v}}, \tilde{\Lambda}$ in the following problem

$$\max_{\substack{\tilde{\mathbf{A}}, \tilde{\Lambda}, \tilde{\mathbf{W}}, \tilde{\mathbf{v}}, \\ \mathbf{p}, \delta, \Delta}} \quad \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{3} \sum_{j \in \mathscr{U}_n} r_{nj}(\tilde{\mathbf{v}}_t, \Delta_{nj}) \tag{6.15a}$$

$$\text{s.t.:} \quad R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t}, \tilde{\mathbf{w}}_{n,t}, \delta_{nj}\right) \geq r_{nj}(\tilde{\mathbf{v}}_t, \Delta_{nj}) \tag{6.15b}$$

$$\delta_{nj}\left|\tilde{h}_{nj,t}\right| \geq \left(1-a_{n,i,j,t}\right)\delta_{ni}\left|\tilde{h}_{ni,t}\right| \tag{6.15c}$$

$$\delta_{nj}\left|\tilde{h}_{nj,t}\right| \leq \left(1-a_{n,i,j,t}\right)M+\delta_{ni}\left|\tilde{h}_{ni,t}\right| \tag{6.15d}$$

$$\sum_{n=1}^{3}\sum_{j\in\mathscr{U}_n}|\tilde{w}_{nj,t}|^2 \leq P^{\max} \tag{6.15e}$$

$$\left|v_{nij,t}\right|^2 \leq (1-a_{n,j,i,t})\lambda_{nij,t}; \sum_{j\in\mathscr{U}_n}\lambda_{nij,t} \leq p^{\max} \tag{6.15f}$$

$$a_{n,i,j,t}+a_{n,j,i,t}=1; a_{n,j,j,t}=0; a_{n,i,j,t}\in\{0,1\} \tag{6.15g}$$

$$\text{(6.14e)},\text{(6.14f)}. \tag{6.15h}$$

for $i,j\in\mathscr{U}_n, i\neq j$, $n=1,2,3$, and $t=1,\ldots,T$. We denote $\tilde{\mathbf{A}}=\{\tilde{\mathbf{a}}_t,\forall t=1,\ldots,T\}, \tilde{\mathbf{W}}=\{\tilde{\mathbf{w}}_t,\forall t=1,\ldots,T\}$, $\tilde{\mathbf{V}}=\{\tilde{\mathbf{v}}_t,\forall t=1,\ldots,T\}$ and $\tilde{\Lambda}=\{\tilde{\lambda}_t,\forall t=1,\ldots,T\}$. Note that variables $\mathbf{p}$, δ, and Δ in (6.14) and (6.15) are similar, while variables $\tilde{\mathbf{A}},\tilde{\mathbf{W}},\tilde{\mathbf{v}},\tilde{\Lambda}$ in (6.15) are different from $\mathbf{a},\mathbf{w},\mathbf{v},\lambda$ in (6.14) in the index t. The purpose of using $\tilde{\mathbf{A}},\tilde{\mathbf{W}},\tilde{\mathbf{v}},\tilde{\Lambda}$ is to jointly solve them together with $\mathbf{p}$, δ, and Δ. The physical meaning of (6.15a), which is computed as the average sum rate evaluated from the past CSI $\tilde{h}_{ni,t},\tilde{\mathbf{g}}_{ni,t},\forall n,i,t$ represents the long-term overall network performance. Here, the reasons that we choose this objective function are two main folds. First, the available collection of past CSI can represent any channel realization $h_{ni,t},\mathbf{g}_{ni,t},\forall n,i,t$ drawn from a random distribution. This means that we can use these channel realization value at any location of UAVs and users within the considered region. Using this collection of past CSI to maximize the average sum rate can be considered as the optimization of the long-term network performance, which is reasonable for the determination of UAV's location. Second, since collecting CSI can be done in any previous time before the current transmission period, solving (6.15) can also be executed offline before the current transmission.

Next, once we attain the solution $\tilde{\mathbf{A}}^\circ,\tilde{\Lambda}^\circ,\tilde{\mathbf{W}}^\circ,\tilde{\mathbf{v}}^\circ,\mathbf{p}^\circ,\delta^\circ,\Delta^\circ$, we command the UAV to stay at those location corresponding to the value of $\mathbf{p}^\circ$. Then, we can obtain the real CSI associated with location $\mathbf{p}^\circ$ and proceed to solve for the remaining variables $\mathbf{a},\lambda,\mathbf{w},\mathbf{v}$ through the

following problem.

$$\max_{\mathbf{a},\mathbf{w},\mathbf{v},\lambda} \quad \sum_{n=1}^{3} \sum_{j \in \mathscr{U}_n} r_{nj}(\mathbf{v}, \Delta_{nj}) \tag{6.16a}$$

$$\text{subject to : (6.14b)} - \text{(6.14i)} \tag{6.16b}$$

We observe that (6.15) and (6.16) are both general mixed-integer non-linear programs, which are generally NP-hard and difficult to solve. Note that (6.16) is a simpler version of (6.15) with fixed value of $\mathbf{p}$, δ, Δ and no index t. Here, the main difficulty of solving (6.15) is to deal with the binary-related constraints (6.14i). Moreover, even if we relax constraint (6.14i) to make $\tilde{a}_{ni,j,t}$ continuous within the interval $[0,1]$, the relaxed version of (6.15) is still non-convex. The non-convexity of (6.14) is due to the existence of the non-convex non-concave rate functions $r_{nj}(\tilde{\mathbf{v}}_t, \Delta_{nj})$ and $R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t}, \tilde{\mathbf{w}}_{n,t}, \delta_{nj}\right)$ which appear in the objective function and constraint (6.14b). Finding a global solution for (6.15) often requires a high–complexity exhaustive search algorithm, which is impractical. Motivated by this, we only aim at developing a low-complexity algorithm to attain a sub-optimal solution within a polynomial time.

6.5 Low-Complexity Solution Approach

In this section, we present a method to transform (6.15) into a tractable form which is more amenable to develop a low-complexity algorithm based on the principle of DC programming. Note that our proposed method introduces a minimum number of slack variables to attain the solution at most efficiency, which is more efficient than (Nguyen *et al.*, 2016b). The developed algorithm from this approach can be used to solve for the simpler problem (6.16) as well.

6.5.1 DC-based Transformations

We first equivalently rewrite the binary constraint in (6.15g) into the continuous DC form constraints as

$$\tilde{a}_{n,i,j,t} - \tilde{a}_{n,i,j,t}^2 \leq 0 \tag{6.17a}$$

$$0 \leq \tilde{a}_{n,i,j,t} \leq 1. \tag{6.17b}$$

Since (6.17a) implies that $\tilde{a}_{n,i,j,t} \leq 0$ or $\tilde{a}_{n,i,j,t} \geq 1$, the joint condition of (6.17a) and (6.17b) results in $\tilde{a}_{n,i,j,t} = 0$ or 1, which is equivalent to (6.15g). After the above transformation, the difficulties in solving (6.14) still remain because of the continuous but non-convex non-concave rate functions $r_{nj}(\tilde{\mathbf{v}}_t, \Delta_{nj})$, $R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t}, \tilde{\mathbf{w}}_{n,t}, \delta_{nj}\right)$, with respect to their variables in (6.14a), (6.14b), the non-convex non-concave function $\left(1 - \tilde{a}_{n,i,j,t}\right)\delta_{ni}$ in (6.15c), the convex function δ_{nj}^{-1} and Δ_{nmji}^{-1} lie in the greater side of (6.14e) and (6.14f), and the concave function in (6.17a). Based on the concept of DC programming, we will express each of the non-convex non-concave functions $r_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$ and $R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t}, \tilde{\mathbf{w}}_{nt}, \delta_{nj}\right)$ as the difference of two convex or concave ones. To illustrate this, we rewrite $r_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$ as

$$r_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni}) = \underbrace{r_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni}) + \xi_{ni}(\|\tilde{\mathbf{v}}_t\|^2 + \|\Delta_{ni}\|^2)}_{f_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})} - \xi_{ni}(\|\tilde{\mathbf{v}}_t\|^2 + \|\Delta_{ni}\|^2) \tag{6.18}$$

$$= \underbrace{r_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni}) - \zeta_{ni}(\|\tilde{\mathbf{v}}_t\|^2 + \|\Delta_{ni}\|^2)}_{g_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})} + \zeta_{ni}(\|\tilde{\mathbf{v}}_t\|^2 + \|\Delta_{ni}\|^2) \tag{6.19}$$

for any $\xi_{ni}, \zeta_{ni} \geq 0$. If ξ_{ni} (or ζ_{ni}) is chosen sufficiently large, function $f_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$ (or $g_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$) becomes convex (or concave) with respect to variables $\tilde{\mathbf{v}}_t, \Delta_{ni}$ due to the dominance of the strongly convex quadratic function $\xi_{ni}\left(\|\tilde{\mathbf{v}}_t\|^2 + \|\Delta_{ni}\|^2\right)$ (or concave function $-\zeta_{ni}(\|\tilde{\mathbf{v}}_t\|^2 + \|\Delta_{ni}\|^2)$). Thus, the DC form of function $r_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$ is more clearly exposed. Although this DC decomposition approach was previously visited (Kha *et al.*, 2011), we remark that the problem of finding a proper value for ξ_{ni}, ζ_{ni} to make (6.19) the appropriate DC expression is very challenging. Under this observation, we provide the following theorem to characterize the computation of ξ_{ni}, ζ_{ni}.

Theorem 1. *For $\xi_{ni} \geq \xi_0, \zeta_{ni} \geq \zeta_0$, where ξ_0, and ζ_0 can be determined via Appendix 10, $f_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$ and $g_{ni}(\tilde{\mathbf{v}}_t, \Delta_{ni})$ are strongly convex and concave, respectively.*

Proof. The proof is given in Appendix 10. □

Similarly, we can also alter the DC decomposition of $R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj}\right)$ as

$$R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj}\right) = -\psi_{n,j,j}(\|\tilde{\mathbf{w}}_{n,t}\|^2+\|\tilde{\mathbf{a}}_{n,t}\|^2+\left|\delta_{nj}\right|^2) + \underbrace{R_{n,j,j}\left(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj}\right)+\psi_{n,j,j}(\|\tilde{\mathbf{w}}_{n,t}\|^2+\|\tilde{\mathbf{a}}_{n,t}\|^2+\left|\delta_{nj}\right|^2)}_{u_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj})} \quad (6.20)$$

for any $\psi_{nj,j}\geq 0$. Again, if $\psi_{n,j,j}$ are chosen sufficiently large, function $u_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj})$ become convex due to the dominance of strongly convex functions $\psi_{n,j,j}(\|\tilde{\mathbf{w}}_{n,t}\|^2+\|\tilde{\mathbf{a}}_{n,t}\|^2+\left|\delta_{nj}\right|^2)$. The convexity of $u_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj})$ is characterized in the following theorem.

Theorem 2. *For $\psi_{n,j,j}\geq\psi_0$, where ψ_0 can be derived following similar steps in Appendix 10, $u_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj})$ is strongly convex.*

Proof. The proof for Theorem 2 is similar to Appendix 10, which is omit due to space constraint. □

6.5.2 Proposed Relaxation of (6.14)

After applying the DC-based and binary transformation, we remark that the joint existence of dual constraint (6.17a)–(6.17b) will make the subsequent iterative DC-based algorithm may fail to converge, due technical difficulties associated with existence of feasible solutions (Vu *et al.*, 2016b). Inspired by (Vu *et al.*, 2016b), we relax the constraint (6.17a) into the objective function by introducing a new slack variable $\tilde{\mathbf{C}}=\left\{c_{n,i,j,t}\geq 0,\forall n,i,j,t\right\}$ and reformulate (6.14) as

$$\max_{\substack{\tilde{\mathbf{A}},\tilde{\Lambda},\tilde{\mathbf{W}},\tilde{\mathbf{v}},\\ \mathbf{p},\delta,\Delta,\tilde{\mathbf{C}}}} \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{3}\sum_{j\in\mathscr{U}_n}\left[f_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj})-\xi_{nj}(\|\tilde{\mathbf{v}}_t\|^2+\left\|\Delta_{nj}\right\|^2)\right]-A\sum_{t=1}^{T}\sum_{n=1}^{3}\sum_{i,j\in\mathscr{U}_n}c_{n,i,j,t} \quad (6.21a)$$

subject to:

$$u_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj})-\psi_{n,j,j}(||\tilde{\mathbf{w}}_{n,t}||^2+||\tilde{\mathbf{a}}_{n,t}||^2+|\delta_{nj}|^2)\geq g_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj})+\zeta_{nj}(||\tilde{\mathbf{v}}_t||^2+||\Delta_{nj}||^2) \quad (6.21b)$$

$$\delta_{nj}\left|\tilde{h}_{nj,t}\right| \geq 0.25(1-\tilde{a}_{n,i,j,t}+\delta_{ni}|\tilde{h}_{ni,t}|)^2 - 0.25(1-\tilde{a}_{n,i,j,t}-\delta_{ni}|\tilde{h}_{ni,t}|)^2 \tag{6.21c}$$

$$\delta_{nj}\left|\tilde{h}_{nj,t}\right| \leq \left(1-\tilde{a}_{n,i,j,t}\right)M + \delta_{ni}\left|\tilde{h}_{ni,t}\right| \tag{6.21d}$$

$$\tilde{a}_{n,i,j,t} - \tilde{a}^2_{n,i,j,t} \leq c_{n,i,j,t} \tag{6.21e}$$

$$(6.14e), (6.14f), (6.14g) - (6.14i), (6.17b) \tag{6.21f}$$

$\forall i, j, n, t$, where $A > 0$ is a penalty parameter. Note that we have equivalently rewritten (6.15c) as (6.21c) by some simple algebraic manipulations. It is obvious that (6.21) and (6.15) are equivalent when $c_{n,i,j,t} = 0, \forall n, i, j, t$.

6.5.3 DC-based Approximation Method

In this section, we propose the DC-based method to solve (6.21). To better describe this method, we visit the following general DC constraint using a slight abuse of notation

$$f(\mathbf{x}) - g(\mathbf{x}) \leq 0 \tag{6.22}$$

where $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex with respect to $\mathbf{x}$. It is easy to observe that the concave function $-g(\mathbf{x})$ is the factor that makes (6.22) non-convex. Assuming that $g(\mathbf{x})$ is differentiable, the DC-based method approximates $g(\mathbf{x})$ around the point $\mathbf{x}^{[n]}$ to result in the following convex approximate constraint:

$$f(\mathbf{x}) - g(\mathbf{x}^{[n]}) - \left\langle \nabla g(\mathbf{x}^{[n]}), \mathbf{x} - \mathbf{x}^{[n]} \right\rangle \leq 0 \tag{6.23}$$

Note that in the above approximation, we have upper-bounded the concave function $-g(\mathbf{x})$ by its linearization $-g(\mathbf{x}^{[n]}) - \left\langle \nabla g(\mathbf{x}^{[n]}), \mathbf{x} - \mathbf{x}^{[n]} \right\rangle$, which is a convex (and concave) function of variable $\mathbf{x}$. Using the principle from this general example, let us first deal with the objective function (6.21a). We quickly observe that the non-convexity of the objective function (6.21a) is due to the maximization over the convex functions $f_{nj}(\tilde{\mathbf{v}}_t, \Delta_{nj})$. Thus, we can respectively approximate function $f_{nj}(\tilde{\mathbf{v}}_t, \Delta_{nj})$ by its first order Taylor linearization around the point $\tilde{\mathbf{v}}_t^{[n]}$ and

$\Delta_{nj}^{[\ell]}$ as

$$F_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) = f_{nj}(\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) + \underbrace{\hat{f}_{nj}(\tilde{\mathbf{v}}_t;\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) + \breve{f}_{nj}(\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]})}_{\tilde{F}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]})} + 2\xi_{nj}\underbrace{\mathrm{Re}\left(\tilde{\mathbf{v}}_t^{[\ell]H}\tilde{\mathbf{v}}_t - \left\|\tilde{\mathbf{v}}_t^{[\ell]}\right\|^2 + \Delta_{nj}^{[\ell]H}\Delta_{nj} - \left\|\Delta_{nj}^{[\ell]}\right\|^2\right)}_{\bar{F}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]})} \quad (6.24)$$

where we denote $\hat{f}_{nj}(\tilde{\mathbf{v}}_t;\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]})$ as

$$\hat{f}_{nj}(\tilde{\mathbf{v}}_t;\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) = \frac{2\mathrm{Re}\left(\tilde{\mathbf{v}}_t^{[\ell]H}\mathbf{G}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t - \tilde{\mathbf{v}}_t^{[\ell]H}\mathbf{G}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t^{[\ell]}\right)}{\tilde{\mathbf{v}}_t^{[\ell]H}\mathbf{G}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t^{[\ell]} + N_0} - \frac{2\mathrm{Re}\left(\tilde{\mathbf{v}}_t^{[\ell]H}\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t - \tilde{\mathbf{v}}_t^{[\ell]H}\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t^{[\ell]}\right)}{\tilde{\mathbf{v}}_t^{[\ell]H}\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t^{[\ell]} + N_0}, \quad (6.25)$$

using the derivation in (A I-79), and $\breve{f}_{nj}(\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]})$ as

$$\breve{f}_{nj}(\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) = \frac{2\mathrm{Re}\left(\Delta_{nj}^{[\ell]H}\Omega_{nj}\left(\tilde{\mathbf{v}}_t^{[\ell]}\right)\Delta_{nj} - \Delta_{nj}^{[\ell]H}\Omega_{nj}\left(\tilde{\mathbf{v}}_t^{[\ell]}\right)\Delta_{nj}^{[\ell]}\right)}{\tilde{\mathbf{v}}_t^{[\ell]H}\mathbf{G}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t^{[\ell]} + N_0} - \frac{2\mathrm{Re}\left(\Delta_{nj}^{[\ell]H}\tilde{\Omega}_{nj}\left(\tilde{\mathbf{v}}_t^{[\ell]}\right)\Delta_{nj} - \Delta_{nj}^{[\ell]H}\tilde{\Omega}_{nj}\left(\tilde{\mathbf{v}}_t^{[\ell]}\right)\Delta_{nj}^{[\ell]}\right)}{\tilde{\mathbf{v}}_t^{[\ell]H}\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}^{[\ell]}\right)\tilde{\mathbf{v}}_t^{[\ell]} + N_0} \quad (6.26)$$

using the derivation in (A I-80). Similarly, we note that the non-convexity of (6.21b) is because of the concave function $g_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj})$ on the lesser side and the convex function $u_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj})$ on the greater side of the inequality. Thus, we can approximate function $g_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj})$ around the point $\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}$ as

$$G_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) = g_{nj}(\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) + \tilde{F}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) - 2\xi_{nj}\bar{F}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) \quad (6.27)$$

and function $u_{n,j,j}(\tilde{\mathbf{a}}_{n,t}, \tilde{\mathbf{w}}_{n,t}, \delta_{nj})$ around the point $\tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}$ as

$$\begin{aligned} U_{n,j,j}(\tilde{\mathbf{a}}_{n,t}, \tilde{\mathbf{w}}_{n,t}, \delta_{nj}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) &= u_{n,j,j}(\tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) \\ &+ 2\psi_{n,j,j}(\tilde{\mathbf{w}}_{n,t}^{[\ell]H}\tilde{\mathbf{w}}_{n,t} - ||\tilde{\mathbf{w}}_{n,t}^{[\ell]}||^2 + \tilde{\mathbf{a}}_{n,t}^{[\ell]H}\tilde{\mathbf{a}}_{n,t} - ||\tilde{\mathbf{a}}_{n,t}^{[\ell]}||^2 + \delta_{nj}^{[\ell]}\delta_{nj} - (\delta_{nj}^{[\ell]})^2) \\ &+ \hat{u}_{n,j,j}(\tilde{\mathbf{a}}_{n,t}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) + \breve{u}_{n,j,j}(\tilde{\mathbf{w}}_{n,t}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) + \bar{u}_{n,j,j}(\delta_{nj}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) \end{aligned} \quad (6.28)$$

where we denote $\hat{u}_{n,j,j}(\tilde{\mathbf{a}}_{n,t}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})$ as

$$\begin{aligned} \hat{u}_{n,j,j}(\tilde{\mathbf{a}}_{n,t}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) = &\frac{\tilde{\pi}_{nj}(\tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})\left(\tilde{\mathbf{a}}_{n,j,t} - \tilde{\mathbf{a}}_{n,j,t}^{[\ell]}\right)}{\pi_{nj}^H(\tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})\mathbf{a}_{n,j,t}^{[\ell]} + N_0} - \\ &\frac{\tilde{\pi}_{nj}(\tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})\left(\tilde{\mathbf{a}}_{n,j,t} - \tilde{\mathbf{a}}_{n,j,t}^{[\ell]}\right)}{\tilde{\pi}_{nj}(\tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})\tilde{\mathbf{a}}_{n,j,t}^{[\ell]} + N_0}, \end{aligned} \quad (6.29)$$

where

$$\pi_{nj}(\tilde{\mathbf{w}}_{n,t}, \delta_{nj}) = [|\tilde{w}_{n,1,t}^H \tilde{h}_{nj,t}\delta_{nj}|^2, \ldots, |\tilde{w}_{n,U,t}^H \tilde{h}_{nj,t}\delta_{nj}|^2]^T.$$

In addition, we have

$$\tilde{\mathbf{a}}_{n,j,t} = [\tilde{a}_{n,1,j,t}, \ldots, 1, \ldots, \tilde{a}_{n,U,j,t}]^T,$$

and

$$\tilde{\pi}_{nj}(\tilde{\mathbf{w}}_{n,t}, \delta_{nj}) = [|\tilde{w}_{n,1,t}^H \tilde{h}_{n,j,t}\delta_{nj}|^2, \ldots, 0, \ldots, |\tilde{w}_{n,U,t}^H \tilde{h}_{n,j,t}\delta_{nj}|^2]^T,$$

where 0 appears at the j^{th} element. Moreover, we denote $\breve{u}_{n,j,j}(\tilde{\mathbf{w}}_{n,t}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})$ as

$$\begin{aligned} \breve{u}_{n,j,j}(\tilde{\mathbf{w}}_{n,t}; \tilde{\mathbf{a}}_{n,t}^{[\ell]}, \tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]}) = &\frac{2\text{Re}\left(\tilde{\mathbf{w}}_{n,t}^{[\ell]H}\theta(\tilde{\mathbf{a}}_{n,j,t}^{[\ell]})\mathbf{H}_{n,j,t}\delta_{nj}^{[\ell]}\left(\tilde{\mathbf{w}}_{n,t} - \tilde{\mathbf{w}}_{n,t}^{[\ell]}\right)\right)}{\pi_{nj}^H(\tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})\mathbf{a}_{n,j,t}^{[\ell]} + N_0} - \\ &\frac{2\text{Re}\left(\tilde{\mathbf{w}}_{n,t}^{[\ell]H}\theta(\tilde{\mathbf{a}}_{n,j,t}^{[\ell]})\tilde{\mathbf{H}}_{n,j,t}\delta_{nj}^{[\ell]}\left(\tilde{\mathbf{w}}_{n,t} - \tilde{\mathbf{w}}_{n,t}^{[\ell]}\right)\right)}{\tilde{\pi}_{nj}(\tilde{\mathbf{w}}_{n,t}^{[\ell]}, \delta_{nj}^{[\ell]})\tilde{\mathbf{a}}_{n,j,t}^{[\ell]} + N_0}, \end{aligned} \quad (6.30)$$

where

$$\theta(\tilde{\mathbf{a}}_{n,j,t}) = [\tilde{a}_{n,1,j,t}\mathbf{1}^T, \ldots, 0^T, \ldots, a_{n,U,j,t}\mathbf{1}^T]^T,$$

where $\mathbf{0}$ appears at the j^{th} element. In addition, we have denoted $\hat{\mathbf{H}}_{nj,t} = \tilde{\mathbf{h}}_{nj,t}\tilde{\mathbf{h}}_{nj,t}^H$ and $\mathbf{H}_{nj,t} = \text{Bdiag}\left(\hat{\mathbf{H}}_{nj,t},\ldots,\hat{\mathbf{H}}_{nj,t}\right)$, and $\tilde{\mathbf{H}}_{nj,t} = \text{Bdiag}\left(\hat{\mathbf{H}}_{nj,t},\ldots,0,\ldots,\hat{\mathbf{H}}_{nj,t}\right)$. Finally, we have denoted $\bar{u}_{n,j,j}(\delta_{nj};\tilde{\mathbf{a}}_{n,t}^{[\ell]},\tilde{\mathbf{w}}_{n,t}^{[\ell]},\delta_{nj}^{[\ell]})$ as

$$\bar{u}_{n,j,j}(\delta_{nj};\tilde{\mathbf{a}}_{n,t}^{[\ell]},\tilde{\mathbf{w}}_{n,t}^{[\ell]},\delta_{nj}^{[\ell]}) = \frac{2\delta_{nj}^{[\ell]}\phi_{nj}^H(\tilde{\mathbf{w}}_{n,t}^{[\ell]})\tilde{\mathbf{a}}_{n,j,t}^{[\ell]}(\delta_{nj}-\delta_{nj}^{[\ell]})}{\pi_{nj}^H(\tilde{\mathbf{w}}_{n,t}^{[\ell]},\delta_{nj}^{[\ell]})\tilde{\mathbf{a}}_{n,j,t}^{[\ell]}+N_0} - \frac{2\delta_{nj}^{[\ell]}\tilde{\phi}_{nj}^H(\tilde{\mathbf{w}}_{n,t}^{[\ell]})\tilde{\mathbf{a}}_{n,j,t}^{[\ell]}(\delta_{nj}-\delta_{nj}^{[\ell]})}{\tilde{\pi}_{nj}(\tilde{\mathbf{w}}_{n,t}^{[\ell]},\delta_{nj}^{[\ell]})\tilde{\mathbf{a}}_{n,j,t}^{[\ell]}+N_0} \quad (6.31)$$

where

$$\phi_{nj}(\tilde{\mathbf{w}}_{n,t}) = [|\tilde{w}_{n,1,t}\tilde{h}_{n,j,t}|^2,\ldots,|\tilde{w}_{U,t}\tilde{h}_{j,t}|^2]^T$$

and

$$\tilde{\phi}_{nj}(\tilde{\mathbf{w}}_{n,t}) = [|\tilde{w}_{n,1,t}\tilde{h}_{n,j,t}|^2,\ldots,0,\ldots,|\tilde{w}_{n,U,t}^H\tilde{h}_{n,j,t}|^2]^T,$$

where 0 appears at the j^{th} element. By applying the above approximations, we can formulate the convex approximate of (6.21) at the ℓ^{th} iteration as

$$\max_{\substack{\tilde{\mathbf{A}},\tilde{\Lambda},\tilde{\mathbf{W}},\tilde{\mathbf{y}},\\ p,\delta,\Delta,\tilde{\mathbf{C}}}} \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{3}\sum_{j\in\mathscr{U}_n}\left[F_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) - \xi_{nj}(\|\tilde{\mathbf{v}}_t\|^2+\left\|\Delta_{nj}\right\|^2)\right] - A\sum_{t=1}^{T}\sum_{n=1}^{3}\sum_{i,j\in\mathscr{U}_n} c_{n,i,j,t} \quad (6.32a)$$

subject to:

$$U_{n,j,j}(\tilde{\mathbf{a}}_{n,t},\tilde{\mathbf{w}}_{n,t},\delta_{nj};\tilde{\mathbf{a}}_{n,t}^{[\ell]},\tilde{\mathbf{w}}_{n,t}^{[\ell]},\delta_{nj}^{[n]}) - \psi_{n,j,j}(\|\tilde{\mathbf{w}}_{n,t}\|^2+||\tilde{\mathbf{a}}_{n,t}||^2+|\delta_{nj}|^2) \geq G_{nj}(\tilde{\mathbf{v}}_t,\Delta_{nj};\tilde{\mathbf{v}}_t^{[\ell]},\Delta_{nj}^{[\ell]}) + \zeta_{nj}(\|\tilde{\mathbf{v}}_t\|^2+\left\|\Delta_{nj}\right\|^2) \quad (6.32b)$$

$$\delta_{nj} \geq 0.25\left(1-\tilde{a}_{n,i,j,t}+\delta_{ni}\right)^2 - 0.25\left(1-\tilde{a}_{n,i,j,t}^{[\ell]}-\delta_{ni}^{[\ell]}\right)^2 - \frac{1-\tilde{a}_{n,i,j,t}^{[\ell]}-\delta_{ni}^{[\ell]}}{2}\left(\tilde{a}_{n,i,j,t}^{[\ell]}+\delta_{ni}^{[\ell]}-\tilde{a}_{n,i,j,t}-\delta_{ni}\right) \quad (6.32c)$$

$$\frac{2\delta_{nj}^{[\ell]}-\delta_{nj}}{(\delta_{nj}^{[\ell]})^2} \geq \sqrt{(\bar{x}_b-x_{nj})^2+(\bar{y}_b-y_{nj})^2+(H_b-H)^2} \quad (6.32d)$$

$$\frac{2\Delta_{nmji}^{[\ell]}-\Delta_{nmji}}{(\Delta_{nmji}^{[\ell]})^2} \geq \sqrt{(x_{nj}-\bar{x}_{mi})^2+(y_{nj}-\bar{y}_{mi})^2+H^2} \tag{6.32e}$$

$$\tilde{a}_{n,i,j,t}-2\tilde{a}_{n,i,j,t}^{[\ell]}\tilde{a}_{n,i,j,t}+(\tilde{a}_{n,i,j,t}^{[\ell]})^2) \leq c_{n,i,j,t} \tag{6.32f}$$

$$(6.14g)-(6.14i),(6.17b) \tag{6.32g}$$

for $i,j \in \mathscr{U}$, where $\tilde{\mathbf{A}}^{[\ell]},\tilde{\mathbf{W}}^{[\ell]},\tilde{\mathbf{V}}^{[\ell]},\delta^{[\ell]},\Delta^{[\ell]}$ are not the optimization variables but the parameters. Note that we approximated the concave function $-\left(1-\tilde{a}_{n,i,j,t}-\delta_{ni}\right)^2$ by the second and third terms on the right side of (6.32c). Similarly, we also approximated δ_{nj}^{-1} and Δ_{nmji}^{-1} by the left side of (6.32d) and (6.32e), respectively. Algorithm 7 outlines the DC–method to solve (6.21). First, we initially set $A^{[0]}$ at sufficient small value. Then after each iteration n, we increase $A^{[\ell]}$ using a constant $\varepsilon > 1$ to ensure that $\sum_{\forall n,j,i,t} c_{n,i,j,t}^{\star} = 0$ when $\ell \to \infty$. The proof that Algorithm 7 converges after a finite number of iteration is similar to (Vu *et al.*, 2016b), which is omitted here.

After using Algorithm 7 to solve (6.21), the value of UAV's positions $\mathbf{p}^{\circ}$ is determined. At this point, we command the UAV to fly to $\mathbf{p}^{\circ}$ to attain the real CSI there and use these CSI as the input for problem (6.16). Then, we can develop a low-complexity algorithm similar to Algorithm 7 to solve for the value of variables $\mathbf{a},\lambda,\mathbf{w},\mathbf{v}$. We omit the details of this proceess here for brevity.

Algorithm 7:

1: Initialize starting points of $\tilde{\mathbf{A}}^{[\ell]},\tilde{\mathbf{W}}^{[\ell]},\tilde{\mathbf{V}}^{[\ell]},\delta^{[n]},\Delta^{[n]}$ and $A^{[\ell]}$;
2: Set $\ell := 0$;
3: **repeat**
4: Optimally solve the convex problem in (6.32) to achieve $\tilde{\mathbf{A}}^{\star},\tilde{\mathbf{W}}^{\star},\tilde{\mathbf{v}}^{\star},\tilde{\Lambda}^{\star},\mathbf{p}^{\star},\delta^{\star},\Delta^{\star},\tilde{\mathbf{C}}^{\star}$;
5: Set $\ell := \ell+1$;
6: Update $\tilde{\mathbf{A}}^{[\ell]}=\tilde{\mathbf{A}}^{\star},\tilde{\mathbf{W}}^{[\ell]}=\tilde{\mathbf{W}}^{\star},\tilde{\mathbf{V}}^{[\ell]}=\tilde{\mathbf{V}}^{\star},\delta^{[\ell]}=\delta^{\star},\Delta^{[\ell]}=\Delta^{\star}$;
7: Update $A^{[\ell]}=\min\left\{\varepsilon A^{[\ell-1]},A_{\max}\right\}$;
8: **until** Convergence of the objective (6.32a);
9: Assign $\mathbf{p}^{\circ}=\mathbf{p}^{\star}$.

6.6 Numerical Results

We consider a circular coverage of radius $20d_0$ ($d_0 = 10^3$ meter is the reference distance) centered at the MBS with coordinate $\{0,0,0\}$. The coverage is equally divided into 3 sectors, each contains 4 users and 4 UAVs. The users in sector n are evenly placed on the circumference of radius $10d_0$ away from the MBS within sector n. Each UAV's position in sector n is bounded inside this sector area. Unless otherwise mentioned, we set $P_{\max} = 43$ dBm, $p_{\max} = 23$ dBm, $N_0 = -120$ dB, $K_b = 4$ dB, $K_u = 3$ dB, $T = 100$. After solving for $\mathbf{p}^\circ$, we consider the random Rician channels $\tilde{h}_{nj}$ and $\tilde{\mathbf{g}}_{nj}$ as the input to solve (6.16) for $\mathbf{a}, \mathbf{w}, \mathbf{v}, \lambda$.

In Fig. 6.3, we show the convergence of Algorithm 7 when solving (6.15) to obtain $\mathbf{p}^\circ$. Then, we fix $\mathbf{p} = \mathbf{p}^\circ$ and examine the convergence when runing Algorithm 7 to solve (6.16) at three choices of initial setup. The curves Initial Point 1 and 3 have the same value of $\mathbf{w}^{[\ell]}, \mathbf{v}^{[\ell]}A^{[\ell]}$, but the values of ξ_0, ζ_0, ψ_0 for Initial Point 1 uses much smaller than Initial Point 3, where we compute ξ_0, ζ_0, ψ_0 for Initial Point 3 according to Appendix 10. However, the curves Initial Point 1 and 2 have the same value of ξ_0, ζ_0, ψ_0, but the values of $\mathbf{w}^{[\ell]}, \mathbf{v}^{[\ell]}A^{[\ell]}$ are chosen differently. We observe that the objective value (6.32a) always improves and converges after some number of iterations. Here, we also observe that the curves Initial Point 1 and 2 only require around 50 iterations to convergence, but the curve Initial Point 3 requires up to 300 iterations. In fact, the choice of $\mathbf{a}^{[\ell]}, \mathbf{w}^{[\ell]}, \mathbf{v}^{[\ell]}A^{[\ell]}$ does not have a significant impact on Algorithm 7 performance, but a loose choice of ξ_0, ζ_0, ψ_0 can lead to slow convergence behavior. As reflected in Fig. 6.3, there also exist tighter bounds for ξ_0, ζ_0, ψ_0 than the results in Appendix 10. In Fig. 6.3, we validate this argument by heuristically choosing a better choice of ξ_0, ζ_0, ψ_0 for Initial Point 1 and 2. Note that theoretically deriving such bounds is challenging so that we search for these bounds by gradually reducing the value of ξ_0, ζ_0, ψ_0 until Algorithm 7 does not converge.

In Fig. 6.4, we compare the achievable sum rate of users considering different transmission schemes. In particular, the problem of NOMA no Cooperation Scheme is similar to (6.14), except that there is no cooperation between the UAVs, e.g., we additionally consider the constraint $v_{ij} = 0, \forall i \neq j$. The NOMA no Cooperation fixed Position Scheme is similar to the NOMA

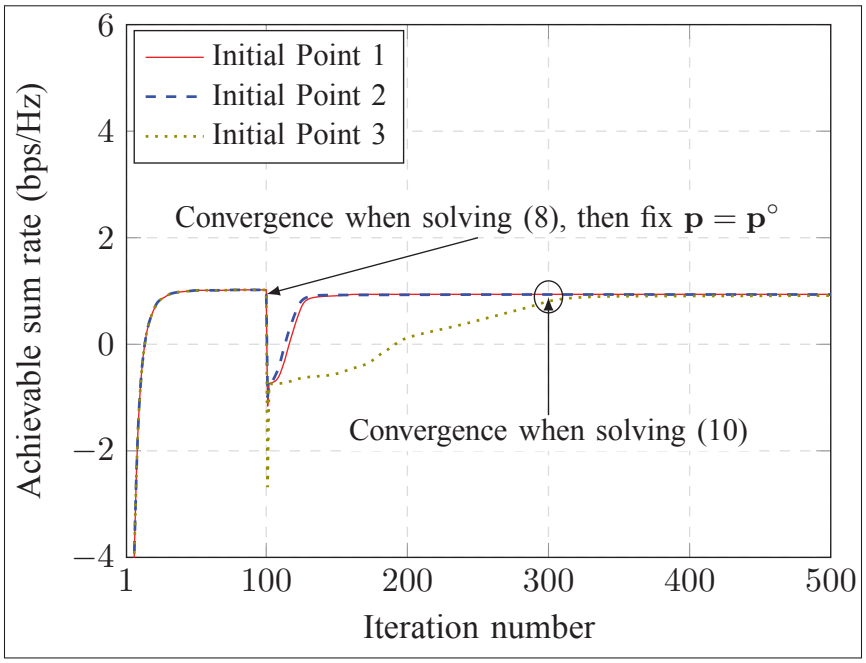

Figure 6.3 Convergence of the proposed algorithm.

no Cooperation Scheme, where we choose the fixed position of all the considered UAVs. In particular, the coordinate of UAV i in sector n is chosen as $x_{ni} = \bar{x}_{ni}/2$ and $y_{ni} = \bar{y}_{ni}/2, \forall i \in \mathcal{U}_n$. On the other hand, in the no NOMA Scheme, there is no application of NOMA for the WB transmission, and consequently there is no cooperation between the UAVs. We can easily confirm the similarity of this no NOMA design with (Mozaffari *et al.*, 2017). In this case, we can formulate the no NOMA problem similar to (6.14) where $\tilde{a}_{n,i,j,t} = 1, \forall n, i, j$, and (6.14i) is removed. Similarly, in the no NOMA fixed Position Scheme, we apply the aforementioned fixed position strategy for the no NOMA Scheme. When $p^{\max}$ increases, all the schemes increase and saturate at high value $p^{\max}$, where the Proposed Scheme and NOMA no Cooperation Scheme always outperforms the other schemes. This is because the NOMA process improves the WB rates, so that when the UAVs have higher $p^{\max}$, the system can achieve higher rate. Moreover, when combining the NOMA process with the cooperation between the UAVs, the system can achieve a higher gain, as shown by the gap between the Proposed Scheme and the NOMA no Cooperation Scheme. However, without the NOMA process, the WB rate is limited and hence WA rate is also limited according to constraint (6.14b). In the NOMA no Cooperation and no

NOMA with fixed Position Schemes, it is important to note that fixing the position also reflects a fixed decoding order, which also limits the benefit of NOMA over the WB transmissions.

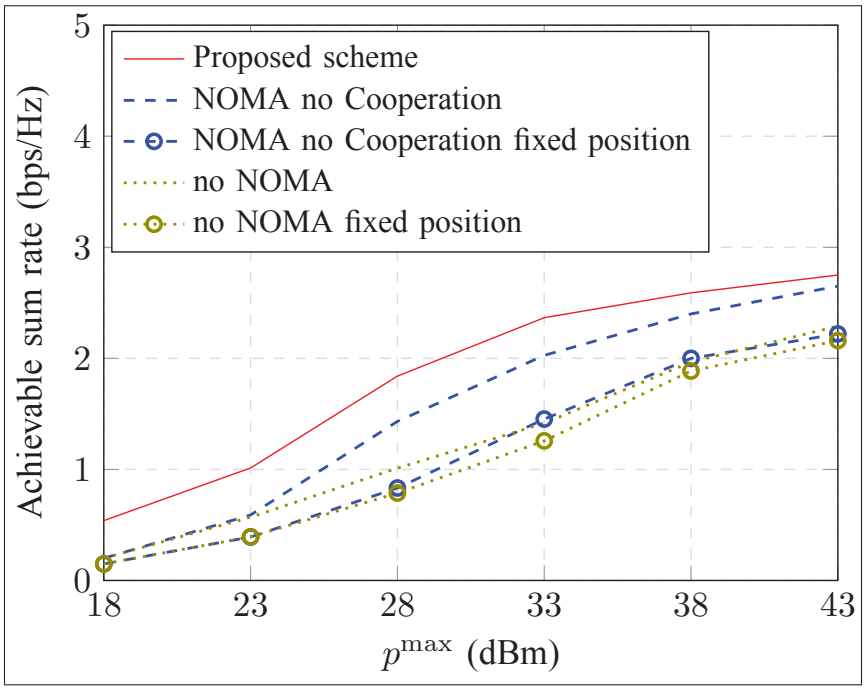

Figure 6.4 Comparison of the achieved sum rate with respect to $p_{\max}$.

In Fig. 6.5, we show the achievable sum rate of the users with respect to the MBS's maximum power budget $P^{\max}$. We again observe that the Proposed Scheme always significantly outperforms the other schemes, which is consistent with the analysis of Fig. 6.4. Beside, we also observe that when $P^{\max}$ increases, the achievable rate of all the schemes increase and saturate at high power $P^{\max}$. This can be explained as when $P^{\max}$ increases, the MBS can allocate more power to increase the WB rate, which in turn increases the WA rate. However, when $P^{\max}$ is high, the MBS does not allocate all of its available power to the UAVs to avoid increasing the co-channel interference between the UAVs. This results in a saturation of WB rate, and subsequently leads to the saturation of the WA achievable sum rate.

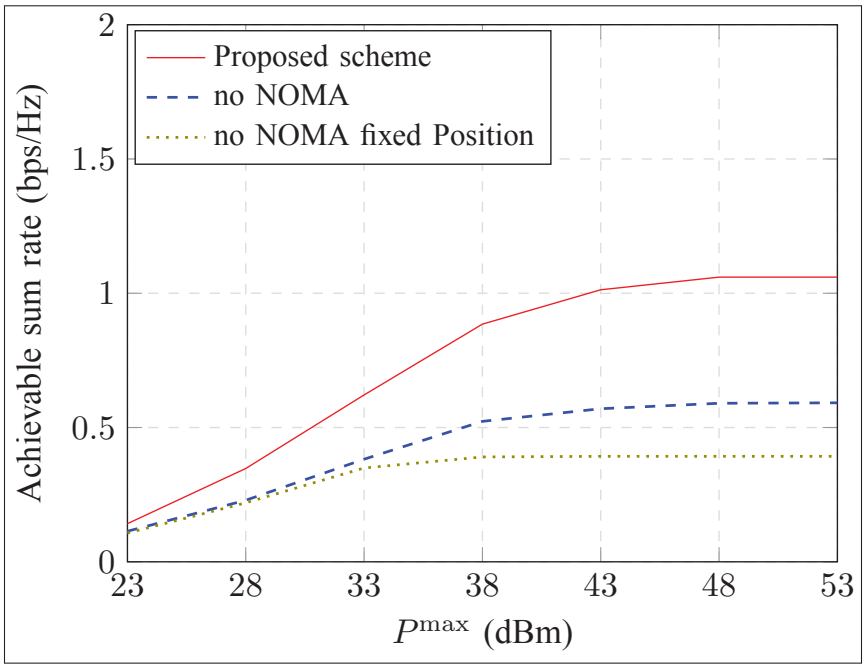

Figure 6.5 Comparison of the achieved sum rate with respect to $P^{\max}$.

Finally, in Fig. 6.6, we show the performance of the Proposed Scheme when the number of UAV and the number of cluster of the users is $U = 4, 7, 10$. Again, we observe the when $p^{\max}$ increases, the achievable rate of the Proposed Scheme in different scenarios increase, which is consistent with Fig. 6.4. However, when U becomes greater, we can achieve higher sum rate. This is because under our proposed cooperative NOMA scheme, more UAVs can cooperatively provides higher degree of freedom to transmit to the users. Also, we notice that the 4 UAVs scheme significantly outperforms the 1 (and 2) UAV(s) schemes. This is because it is less beneficial with fewer UAVs to cooperate under the cooperative NOMA scheme. Consequently, inter-user interference is poorly coordinated, leading to a rapid saturation at low user sum rate at the 1 (and 2) UAV(s) schemes. This corroborates the improvement of our proposed scheme.

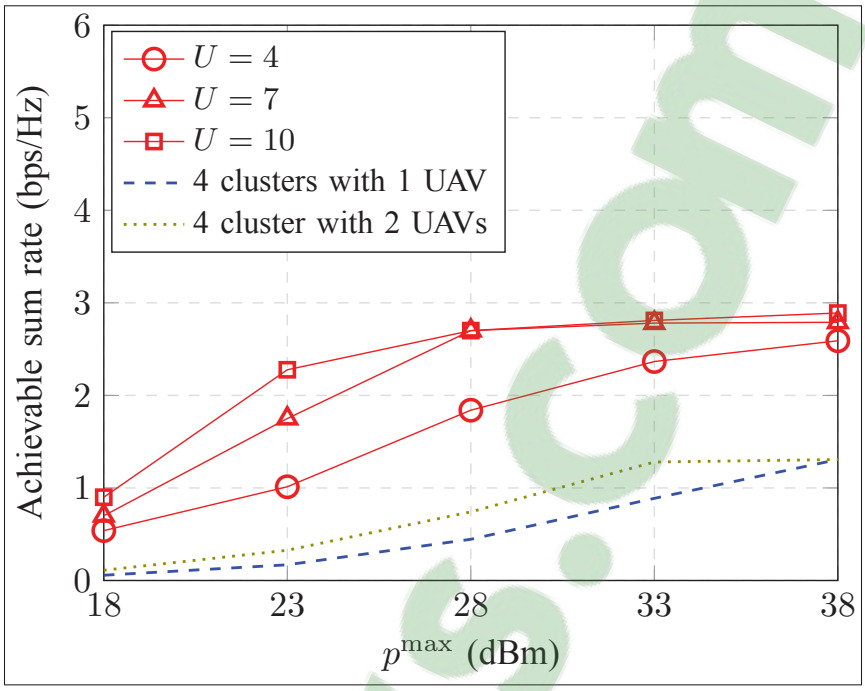

Figure 6.6 Comparison of the achieved sum rate with respect to $p_{\max}$ between the number of UAV.

6.7 Concluding Remarks

In this paper, we investigated the downlink transmissions in the UAV–assisted WB networks. We proposed a novel cooperative NOMA strategy to simultaneously boost the system performance. Observing that the employed NOMA scheme is affected by the change of UAV's position, we jointly optimized the radio resource allocation at the MBS and UAVs along with the decoding order of the NOMA process and the positions of the UAVs to maximize the user sum rate. To overcome the difficulty in achieving the optimal solution, we proposed a framework based on the method of DC program characterized by the Lipschitz continuity to transform the original problem in an approximate sense, into a series of convex optimization problems, leading to a low-complexity algorithm that may be iterated until convergence. Numerical results showed that our achieved solution, under the proposed model and developed algorithm, can outperform the other designs which aim at optimizing without using cooperative NOMA or do not optimize the UAV position. For future work, it is interesting to analyze the theoretical gain

of the proposed NOMA scheme and to study the impact of imperfect SIC in (Tabassum *et al.*, 2017) on the UAV's cooperation and overall network performance.

CONCLUSION AND RECOMMENDATIONS

7.1 Conclusions

This dissertation studied the performance of PtMP WB HetNets. As being raised by the academic and industrial research, WB plays an important role in the 5G perspective to sustain the future 5G's operation. However, jointly designing the WB together with the WA communications is generally challenging due to their fundamental bottleneck relationship. The main theme of this dissertation was to design the multi-dimensional resource allocation in PtMP WB HetNets under an explicit consideration of WB and WA transmission relationship. Via the joint optimization, this dissertation improved performance of the WB HetNets from different approaches. In particular, Chapter 2 addressed the coupling between the uplink and downlink transmissions and designed a joint radio resource allocation of beamforming, power, and spectrum partitioning which maximizes small cell sum rate on both uplink and downlink sides. On the other hand, Chapter 3 concentrated on the energy-efficiency side of the WB HetNets. Then, a joint optimization of downlink transmit beamforming and power allocation which maximizes the access energy-efficiency was formulated. By developing a low-complexity iterative algorithm based on SCA and SOCP, each of these work arrived at a high-quality solution of resource allocation which outperforms the conventional work and achieve a close-to-optimal solution achieved by the exhaustive search algorithm. Chapter 4 further considered a finite-size buffer at each small cell which can flexibly store and release backhaul data. Via jointly optimizing the radio resource allocation together with data management in the buffer which maximizes the total throughput of small cell user over time, this work derived an appropriate design of buffer usage which can achieve a better network performance compared to previous design without properly exploiting the dynamics of the buffer queue. optimizing the radio resource allocation together with data management in the buffer which maximizes the total throughput of small cell user over time. Chapter 5 enhanced the WB HetNets by proposing a novel co-

operative NOMA to improve both the achievable WB and WA rates. This work proposed to solve for a joint solution of NOMA's decoding order, small cell cooperation policy, transmit beamforming at the macro cell BS and power allocation at the SAPs which maximizes the total small cell and macro cell user sum rate. By employing a novel low-complexity algorithm based on DC programming and Lipschitz continuity, the achieved results of this work show a significant gain of system throughput under the proposed cooperative NOMA compared to the literature. Finally, Chapter 6 presented a novel idea of using UAVs combined with the cooperative NOMA to assist the WB communications in WB networks. Chapter 6 proposed to jointly optimize the radio resource allocation, NOMA's decoding order, UAV cooperation policy, together with the UAVs' locations which maximizes the user sum rate. Via developing a novel, more advanced with lower complexity algorithm based on the DC programming and Lipschitz continuity, achieved results show that the new system of WB UAV networks can significantly boost the network performance compared to traditional WB designs.

7.2 Recommendations

For future work, we recommend to move beyond than the focus on optimizing the resource allocation of PtMP NLoS WB HetNets. In fact, WB can be accommodated on mmWave spectrum band or can be combined with advanced 5G technological candidates. According to the results in Chapters 4, 5, and 6, proper exploitation of small cell buffering, NOMA or UAV-assisted communications can significantly improves the network performance compared to sole resource allocation design. This implies that there might exist some other approaches to improve the WB operation in 5G dense networks. In particular, small cell and macrocell BSs can indeed flexibly switch between the mmWave and sub-6 GHz modes to harness both the NLoS and LoS communications for WB and WA transmissions. Small cell BSs can either communicate with macrocell BS using NLoS or LoS, depending on their geometrical distances; or they can communicate with each other using LoS combined with multi-hop communications. The

challenge is how to determine which NLoS and LoS communications should be allocated for each WB communication in order to achieve the best network performance.

Furthermore, equipping a local cache beside using finite-size buffer at each small cell can potentially improve the user sum rate compared to the achieved results in Section 4. Caching content at the small cell BS prior to be requested by local users can efficiently reduce WB overload, thus preserve more communication channel for needed WB transmission to achieve higher data rate. However, the prediction process to cache data is not always perfect, so that content must be transmitted through WB links and buffered whenever a cache miss occurs. Therefore, it is interesting to appropriately manage the cached and buffered content through jointly with the resource allocation to improve the performance of the WB networks.

Cache is recently proposed for UAV-assisted communications and is obviously a prominent solution for UAV-assisted WB networks. Cache usage at each UAV can be combined with the advantages of the proposed cooperative NOMA and the flexible movement of each UAV in order to exploit more cooperation between UAVs. However, it is challenging to jointly optimize the cached content management, the decoding policy at each UAV, the cooperation between the UAVs, and the positions of UAV in order to leverage the overall network performance.

Finally, we can concentrate on the energy-efficiency characteristic of the proposed cooperative NOMA on WB HetNets used in Chapter 5. Although NOMA can boost the achievable rate of each user, its implementation is complicated due to the complexity nature of the SIC process. The SIC principle involves in the decoding of other users' messages, which in turn requires more energy consumption at each small cell BSs. Beside, some small cell BSs must cooperate to transmit more messages to other users beside its own user, therefore, increases the amount of energy consumption. To achieve the greenest communication, it is interesting to determine whether a particular small cell BS will participate in the NOMA or not to save more energy while still maintains a good network performance.

APPENDIX I

1. Proof of equivalence between problem transformations in Chapter 2

1.1 Equivalence between the relaxation of (2.7) and (2.9)

To prove that (2.7) and (2.9) are equivalent by adding more slack variables and constraints, we have to prove that at optimality, all the newly introduced constraints (2.10), (2.11) are active. This can be easily proved by contradiction. Let us assume that at the optimum solution $\alpha^\star, \mathbf{v}^\star, \rho^\star, \mathbf{p}^\star, \mathbf{q}^\star, \mathbf{t}^\star$, some constraints in (2.10) and (2.11) are strictly inequalities, e.g. the ith constraint. Observing from (2.10), we can slightly increase each optimum solution $t_i^{o\star}$ until equality occurs. Consequently, we can increase $q_i^{o\star}$ in (2.11) until equality occurs. However, increasing $q_i^{o\star}$ results in an improvement of objective function (2.9a), which contradicts the assumption of the optimality. Thus, we conclude that at optimality, all the constraints in (2.10), (2.11) occur at equalities.

1.2 Equivalence in the transformation in (2.12)

The equivalence between these transformations can again be proved by contradiction. Assuming that $\mathbf{z}$ and the constraints (2.12) are added into (2.9). Without any loss of generality, assuming that at optimum solution of $\alpha^\star, \mathbf{v}^\star, \rho^\star, \mathbf{p}^\star, \mathbf{q}^\star, \mathbf{t}^\star, \mathbf{z}^\star$, some of the constraints in the system (2.12) are strictly inequalities, e.g., the ith constraint with $o = d$. From there, we can form a pair of optimum solution $\left\{z_i^{d\star}/c, t_i^{d\star}c\right\}$ for $c > 1$ so that (2.12b) occurs at equality, where the product of $z_i^{d\star}/c \times t_i^{d\star}c$ remains the same as $z_i^{d\star}t_i^{d\star}$. Then, one can improve the objective function by slightly increasing $t_i^{d\star}$ until (2.12a) occurs at equality. As shown in (1.1), increasing $t_i^{d\star}$ is equivalent to increasing the objective function. This leads to the contradiction

of the assumption of optimality and validates the equivalence in the transformation from (2.9b) to (2.12).

1.3 Equivalence in the transformation in (2.13)

Similar to the contradiction method in Appendices 1.1 and 1.3, the equivalence of these transformations in (2.13) can be proved by proving that at optimality, all the inequalities in (2.13) are active. Without loss of generality, assuming that at optimality with the optimum solution $\alpha^\star, \mathbf{v}^\star, \rho^\star, \mathbf{p}^\star, \mathbf{q}^\star, \mathbf{t}^\star, \mathbf{z}^\star, \mathbf{a}^\star, \mathbf{b}^\star$, there are some constraints with index $k \in \mathscr{F}$ with $o = d$ and $l \in \mathscr{F}$ with $o = u$ in (2.13) occur at strict inequalities, while the rests are equalities. Consider the kth constraints of the inequalities system in (2.13) with $o = d$. There exists a feasible solution of $\tilde{\mathbf{v}}_k = \varepsilon \mathbf{v}_k^\star$, where $\varepsilon < 1$, such that $a_k^{d\star} = \log\left(1 + \Gamma_k^d\left(\tilde{\mathbf{v}}_k, \mathbf{v}_{\sim k}^\star, \alpha^{d\star}\right)\right)$. In addition, there also exists $\tilde{b}_k^d = \varepsilon b_k^{d\star}$ where $\varepsilon < 1$ such that $\tilde{b}_k^d = \log\left(1 + \gamma_k^d\left(\mathbf{p}^{d\star}, 1 - \alpha^{u\star}\right)\right)$. Applying a similar argument for the lth constraints of the inequalities system in (2.13) with $o = u$, there exists a pair of $\tilde{\rho}_l, \tilde{b}_l^u$ such that $a_l^{u\star} = \log\left(1 + \Gamma_l^u\left(\tilde{\rho}_l, \rho_{\sim l}^\star, \alpha^{u\star}\right)\right)$ and $\tilde{b}_l^u = \log\left(1 + \gamma_l^d\left(\mathbf{p}^{u\star}, 1 - \alpha^{d\star}\right)\right)$. With the new values of $\tilde{\mathbf{v}}_k, \tilde{\rho}_l$ less than $\mathbf{v}_k^\star, \rho_l^\star$, respectively, all the constraints in (2.13a), for $i \neq k$ with $o = d$ and $i \neq l$ with $o = u$ result in a higher value of $a_i^{o\star}$ and become strict inequalities. From there, we can slightly increase $1 - \alpha^{d\star}$ and $1 - \alpha^{u\star}$ by new feasible values $1 - \tilde{\alpha}^d$ and $1 - \tilde{\alpha}^u$ and achieve a higher value of the objective function. This contradicts with the assumption of optimality. Therefore, at optimality, all the constraints in (2.13) occur at equalities.

2. Proof of Derivation in (2.28)

We complete the proof by following the similar framework as in (Ben-Tal & Nemirovski, 2001a). We rewrite $e^{x/t}$ as $\left[\exp\left(2^{-m}x/t\right)\right]^{2^m}$ and write its fourth order Taylor approximation

$(f_m(x,t))^{2^m}$ as

$$f_m(x,t) = 1 + \left(2^{-m}\frac{x}{t}\right) + \frac{\left(2^{-m}\frac{x}{t}\right)^2}{2} + \frac{\left(2^{-m}\frac{x}{t}\right)^3}{6} + \frac{\left(2^{-m}\frac{x}{t}\right)^4}{24}. \tag{A I-1}$$

It is obvious that $f_m(x,t)$ is a good approximation of $e^{x/t}$ since at large m, $(f_m(x,t))^{2^m} \approx [\exp(2^{-m}x/t)]^{2^m} = e^{x/t}$. To further extract $y \geq t\,(f_m(x,t))^{2^m}$ into SOC constraints, we first introduce variable κ_{m+4} and rewrite this constraint as

$$\kappa_{m+4} \leq y, \tag{A I-2}$$

$$t\,(f_m(x,t))^{2^m} \leq \kappa_{m+4} \tag{A I-3}$$

Multiplying t on both sides of (A I-3) and introducing slack variable κ_{m+3}, we can rewrite (A I-3) as

$$\kappa_{m+3}^2 \leq t\,\kappa_{m+4}\,;\, t\,(f_m(x,t))^{2^{m-1}} \leq \kappa_{m+3} \tag{A I-4}$$

Thus, by applying the same rule of multiplying t and introducing more slack variables m times, we can rewrite the constraint $y \geq t\,(f_m(x,t))^{2^m}$ as

$$\kappa_{m+4} \leq y, \tag{A I-5}$$

$$\kappa_{m+3}^2 \leq t\,\kappa_{m+4}, \tag{A I-6}$$

$$\kappa_{i-1}^2 \leq t\,\kappa_i, i = 5, \ldots m+3, \tag{A I-7}$$

$$t f_m(x,t) \leq \kappa_4. \tag{A I-8}$$

where $\kappa_4, \kappa_5, \ldots, \kappa_{m+4}$ are the newly introduced variables. Taking further manipulation, we rewrite the term $t f_m(x,t)$ in the last inequality as follow

$$\frac{19}{72}t + \frac{\left(5/6t + x2^{-m-1}\right)^2}{t} + \frac{(t + x2^{-m})^4}{24t^3} \leq \kappa_4. \tag{A I-9}$$

To complete the proof, we can equivalent decompose the above constraint by the introduction of new variables $\kappa_1, \kappa_2, \kappa_3$ as

$$\left(t + x2^{-m}\right)^2 \leq \kappa_1 t, \tag{A I-10}$$

$$\left(5/6t + x2^{-m-1}\right)^2 \leq \kappa_2 t, \tag{A I-11}$$

$$\kappa_1^2 \leq \kappa_3 t, \tag{A I-12}$$

$$(19/72)t + \kappa_2 + (1/24)\,\kappa_3 \leq \kappa_4, \tag{A I-13}$$

where we noted that the first 3 constraints can be rewritten in the form of SOC as

$$\left\| \left[2t + x/2^{m-1} \;\; t - \kappa_1\right] \right\|_2 \leq t + \kappa_1, \tag{A I-14}$$

$$\left\| [5t/3 + x/2^m \;\; t - \kappa_2] \right\|_2 \leq t + \kappa_2, \tag{A I-15}$$

$$\left\| [2\kappa_1 \;\; t - \kappa_3] \right\|_2 \leq t + \kappa_3. \tag{A I-16}$$

3. Proof of Proposition 2 in Chapter 3

To prove that (3.16) and (3.10) are equivalent, we must prove that at optimality of (3.16), (3.16b)–(3.16g) occur at equality. This can be proved by contradiction. Assuming that at the optimal solution $\Pi^\star$, (3.16b)–(3.16g) occur at strict inequalities. However, we can always find $\tilde{q} > q^\star$ which results in a higher objective function but still makes constraint (3.16b) equal. In addition, at (3.16c), we can always find another value $\tilde{\tau} > \tau^\star$ so that (3.16c) becomes equal. Moreover, this increment event of variable $\tau^\star$ to $\tilde{\tau}$ allows $q^\star$ to increase to achieve a better valued objective, which contradicts the assumption of optimality. Similarly, assuming that (3.16d)–(3.16g) occur at strict inequalities at some index $(i, j) \in (\mathscr{S}_a, \mathscr{S}_u)$. At (3.16f), we can find another $\tilde{t}_{ij}^s > t_{ij}^{s\star}$ and $\tilde{s}_{ij} < s_{ij}^\star$ so that (3.16f) occurs at equality. The increment in $t_{ij}^{s\star}$ to $\tilde{t}_{ij}^s$ again offers the possibility of objective function improvement, which contradicts the assumption of optimality. Moreover, with the reduction of $s_{ij}^\star$ to $\tilde{s}_{ij}$, we can also find $\tilde{t}_i^m < t_i^{m\star}$

so that (3.16e) occurs at equality. This implies that $v_i^\star, \mathbf{v}_i^\star, i \in \mathscr{S}_a$ should be reduced to $\tilde{v}_i = \beta v_i^\star, \tilde{\mathbf{v}}_i = \alpha \mathbf{v}_i^\star; \alpha, \beta < 1$ so that (3.16d) and (3.16g) occur at equality with $i \in \mathscr{S}_a$. Finally, the reduction of $\mathbf{v}_i^\star, i \in \mathscr{S}_a$ will lead to the reduction of $\tilde{\tilde{P}}_{\text{tot}}$ and we can find a larger $\tilde{\tau}$ and $\tilde{q}$ to let (3.16c) and (3.16b) occur at equality, which again contradicts the assumption. Therefore, we conclude that at optimality, (3.16b)–(3.16g) should all occur at equalities.

4. Proof of Convergence in Section 3.3.3 of Chapter 3

4.1 Conditions for regular feasible solution (Rockafellar & Wets, 1998)

Given a general non-convex optimization problem as

$$(\mathscr{Q}) : \max_{x \in \Omega} \{ f_0(x) \,|\, f_i(x) \leq 0, g_j(x) \leq 0, i = 1, \ldots, I; j = 1, \ldots J \}, \tag{A I-17}$$

where $\Omega \subset \mathbb{R}^n$, $f_0(x)$ is concave, $f_i(x)$ is convex $\forall i = 1, \ldots, I$, $g_j(x)$ is nonconvex $\forall j = 1, \ldots, J$, a feasible solution $\hat{x}$ of problem $(\mathscr{Q})$ is called regular if the Mangasarian-Fromovitz Constraint Qualification (MFCQ) holds at $\hat{x}$. Particularly, MFCQ holds at $\hat{x}$ if there exists a vector $\omega \in \mathbb{R}^n$ that satisfies

$$\nabla f_i(\hat{x})^T \omega < 0, \text{for } i \in I(\hat{x}) \tag{A I-18}$$

$$\nabla f_i(\hat{x})^T \omega < 0, \text{for } i \in I(\hat{x}) \tag{A I-19}$$

where $I(\hat{x}) = \{ i : f_i(\hat{x}) = 0 \}$ and $J(\hat{x}) = \{ j : g(\hat{x}) = 0 \}$.

4.2 Slater Conditions and Convergence of Algorithm 2

- Under the assumption that all the feasible points of problem (3.17) are regular, and the fact that all the convex approximation used in Section 3.3.3 satisfy the last two conditions

in (Marks & Wright) , any choice of initial point $\mathbf{v}^{(0)}, \nu^{(0)}, \mathbf{p}^{(0)}, \tau^{(0)}$ from the regular feasible solution of (3.16) implies that the approximated problem at the 1st iteration $\left(\tilde{\mathscr{P}}^{(1)}\right)$ also contains all regular feasible solution. According to (Scutari *et al.*, 2017), Slater condition holds for $\left(\tilde{\mathscr{P}}^{(1)}\right)$ since it is a convex problem contains all regular feasible solutions. Similarly, assuming that $\mathbf{v}^{\star}, \mathbf{p}^{\star}, \nu^{\star}, \mathbf{s}^{\star}, \mathbf{t}^{\star}, \tau^{\star}, q^{\star}$ is the optimal solution of $\left(\tilde{\mathscr{P}}^{(1)}\right)$ and by the updating rule in Step 6, $\mathbf{v}^{(1)}, \nu^{(1)}, \mathbf{p}^{(1)}, \tau^{(1)}$ is chosen from the optimal solution $\mathbf{v}^{\star}, \mathbf{p}^{\star}, \nu^{\star}, \mathbf{s}^{\star}, \mathbf{t}^{\star}, \tau^{\star}, q^{\star}$ as the feasible point parameter of $\left(\tilde{\mathscr{P}}^{(2)}\right)$. This also implies that $\left(\tilde{\mathscr{P}}^{(2)}\right)$ contains a regular point which satisfies MFCQ and Slater CQ. By repeating this step, we conclude that Slater condition holds for all the approximated problems.

- Let $f^{(n)}$ denote the optimal objective value and $\Pi^{(n)}$ denote the optimal solution set at the $(n-1)$th iteration of Algorithm 2. Due to the convex approximation in Section 3.3.3, the updating rules in Algorithm 2 and the second condition in (Marks & Wright) ensure that the solution set $\Pi^{(n)}$ is a feasible solution to problem (3.24) at step $n+1$. This subsequently leads to the results of $f^{(n+1)} \geq f^{(n)}$, which means that Algorithm 2 generates a non-decreasing sequence of objective function values. Due to the limited power constraints, the sequence of $f^{(n)}, n = 1, 2, \ldots$ is bounded above and therefore, Algorithm 2 guarantees that the objective converges.

5. Proof of NP-hardness of (3.9) in Chapter 3

We will apply the reduction from the maximum independent set problem to problem (2.7). Since the former problem is NP-hard, this will imply the NP-hardness of (2.7). Suppose that $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$ is an undirected graph. An independent set of $\mathscr{G}$ is a subset $\mathscr{U} \subset \mathscr{V}$ such that no two nodes in $\mathscr{U}$ are connected: for any two vertices $u_i, u_j \in \mathscr{U}$, we have $(u_i, u_j) \notin \mathscr{E}$. To find an independent set with a given size is NP-hard (Luo & Zhang, 2008).

Consider a connected graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$ with S_a vertices, i.e., $|\mathscr{V}| = S_a$. Then we choose the instance of (2.7) with $S_u = 1$. For convenience, we denote the channel gain from the ith SAP to the SUE in the jth SAP as h_{ij}, where we choose $h_{ii} = 1$. For each $u_i \in \mathscr{V}$, let

$$h_{ij} = \begin{cases} \sqrt{KS_a^3} & \text{, if } u_j \text{ is adjacent to } u_i \\ 0 & \text{, otherwise} \end{cases} \tag{A I-20}$$

where we choose $K > \frac{5(S_a^4+2)^5+2S_a^4(S_a^4+2)}{S_a^3}$. Beside, we choose an arbitrary value of N, M and

- $\mathbf{h}_j = \sqrt{K} \times \underbrace{[1,0,\ldots,0]^T}_{N \text{ components}}, \forall j \in \mathscr{S}_a$ and $\mathbf{h}_k = \underbrace{[0,\ldots,0]^T}_{N \text{ components}}, \forall k \in \mathscr{M}$,
- $N_0 = K$, $p_{\max} = 1$, $P_{\max} = S_a + M$, and $r^m_{\min} = r^s_{\min} = 0$.
- $\kappa_m = \kappa_s = 2$, $\alpha_i = 1, \forall i \in \mathscr{S}_a$ and $P_{\text{circ}} + P_{\text{ad}} = P_0$, where P_0 is chosen sufficiently large compared to K,

$$\begin{cases} P_0 > \frac{4}{K} \frac{(S_a+2)^2(S_a+1)^2}{2S_a+3} \\ P_0 > 4S_a(K+1)^2 \\ P_0 > \frac{2S_a(K+1)^2}{r(y)} \\ P_0 > \frac{8S_a^4(K+1)(KS_a^3+K)(KS_a^3+K+1)}{(S_a^3+1)^2(K-1)} + S_a \end{cases} \tag{A I-21}$$

We claim that $\mathscr{G}$ has a maximum independent set of size $|\mathscr{I}|$ if and only if (3.9) has an optimal value $\eta^\star$ that satisfies

$$\xi(|\mathscr{I}|) \leq \eta^\star < \xi(|\mathscr{I}|+1) \tag{A I-22}$$

where

$$\xi(|\mathscr{I}|) = \frac{|\mathscr{I}| \log(1+\frac{1}{K})}{2(|\mathscr{I}| + \frac{|\mathscr{I}|}{K-|\mathscr{I}|+1}) + P_0 + |\mathscr{I}| \log(1+1/K)} \tag{A I-23}$$

We prove "⇒": If $\mathscr{G}$ has a maximum independent set $\mathscr{I}$, then we can choose the value of $\mathbf{p}$

$$p_i = \begin{cases} 1 & , \text{ if } u_i \in \mathscr{I} \\ 0 & , \text{ otherwise} \end{cases} \tag{A I-24}$$

and the value of $\mathbf{v}$ as solving the following problem with the chosen $\mathbf{p}$

$$\min_{\mathbf{v}} \sum_{i \in \mathscr{F}} \|\mathbf{v}_i\|^2 \tag{A I-25a}$$

$$\text{s.t. (3.9b), (3.9c), (3.9e).} \tag{A I-25b}$$

Here, we note that when $\mathbf{p}$ is fixed, (A I-25) can be rewritten as a standard second order cone programming, which is a convex optimization problem with respect to variable $\mathbf{v}$. Moreover, we can remark that by the chosen structure of channel $\mathbf{h}_i, \forall i \in \{\mathscr{S}_a \cup \mathscr{M}\}$ and $r^m_{\min} = 0$, at optimality of (A I-25), all constraints (3.9b) occur at equality (the proof for this is similar to Proposition 2 in the Chapter 3). Beside, at optimality of (A I-25), we have the optimal $\mathbf{v}_k = 0, \forall k \in \mathscr{M}$, $\mathbf{v}_j = 0, \forall j \in \mathscr{S}_a \setminus \mathscr{I}$ and $v_{ij} = 0, \forall i \in \mathscr{I}, \forall j = 2, \ldots, N$ and $v_{ij} \neq 0$ otherwise. Thus, we can easily compute $\sum_{i \in \mathscr{F}} \|\mathbf{v}_i\|^2 = \sum_{i \in \mathscr{I}} |v_{i1}|^2 = \frac{|\mathscr{I}|}{K - |\mathscr{I}| + 1}$ and $R_i(\mathbf{v}) = \log(1 + 1/K), \forall i \in \mathscr{I}$. It is obvious that the computed $\mathbf{p}$ and $\mathbf{v}$ are a feasible solution to (3.9). Therefore, we have a solution for (3.9) with an objective value equal to $\xi(|\mathscr{I}|)$ which satisfies (A I-22).

We prove "⇐": Suppose that we have an optimal solution $\mathbf{p}^\star$ and $\mathbf{v}^\star$ with the optimal objective value $\eta^\star$. Following the same step in (Luo & Zhang, 2008), we compute the first and second derivatives of the objective $\eta(\mathbf{v}, \mathbf{p})$ of (3.9) with respect to each component $p_i, \forall i \in \mathscr{S}_a$. The following claim is provided to characterize the form of the optimal solution $\mathbf{p}^\star$.

Proposition 7. *$\eta(\mathbf{p})$ is either strictly increasing or convex with respect to each variable $p_i, \forall i \in \mathscr{S}_a$ at all the feasible value of $\mathbf{p} \in [0,1]^{S_a \times 1}$ when we choose $P_0 > K$ and $K > \frac{5(S_a^4+2)^5+2S_a^4(S_a^4+2)}{S_a^3}$.*

Proof. The proof is given at the end of this section. □

Since the maximum of a convex function with respect to variable p_i is always attainable at a vertex, we can assume that $p_i^\star$ must be either 0 or 1. Beside, since the maximum of a strictly increasing function is attainable at its highest value of p_i, we can also assume that $p_i^\star$ must be 1. Therefore, it follows that we can assume that $\mathbf{p}^\star$ contains either 0 or 1 elements.

Let us construct a set $\mathscr{A}$ from the optimal variable $\mathbf{p}^\star$ as follow

$$\mathscr{A} := \{u_i | p_i^\star = 1, \forall i \in \mathscr{S}_a\} \tag{A I-26}$$

Let $\mathscr{I}$ be a maximum independent set contained in $\mathscr{A}$. For convenient purpose, we denote $x_i = \frac{1}{KS_a^3 \sum_{j \in \mathscr{A} \setminus i : (u_j, u_i) \in \mathscr{E}} 1 + K}$, $y = \frac{1}{KS_a^3 + K}$, $r(x) = \log(1+x)$. Then we have

$$\eta(\mathbf{v}^\star, \mathbf{p}^\star) \tag{A I-27}$$

$$= \frac{\sum_{i:u_i \in \mathscr{A}} \log\left(1 + \frac{p_i^\star}{KS_a^3 \sum_{j \in \mathscr{A} \setminus i:(u_j,u_i) \in \mathscr{E}} p_j^\star + K}\right)}{2\sum_{i:u_i \in \mathscr{A}} p_i^\star + 2\sum_{i \in \mathscr{F}} \|\mathbf{v}_i^\star\|^2 + \sum_{i \in \mathscr{S}_a} R_i(\mathbf{v}^\star) + P_0} \tag{A I-28}$$

$$\overset{(a)}{\leq} \frac{|\mathscr{I}| r(1/K) + \sum_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} r(x_i)}{2\sum\limits_{i:u_i \in \mathscr{A}} p_i^\star + 2\frac{|\mathscr{I}| + K\sum_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} x_i}{K - (|\mathscr{I}|-1) - K\sum_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} x_i} + |\mathscr{I}| r(1/K) + \sum\limits_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} r(x_i) + P_0} \tag{A I-29}$$

$$\overset{(b)}{<} \frac{|\mathscr{I}| r(1/K) + \sum_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} r(y)}{2(|\mathscr{I}|+1) + 2\frac{|\mathscr{I}| + K\sum_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} y}{K - (|\mathscr{I}|-1) - K\sum_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} y} + |\mathscr{I}| r(1/K) + \sum\limits_{i:u_i \in \mathscr{A} \setminus \mathscr{I}} r(y) + P_0} \tag{A I-30}$$

$$= \frac{|\mathscr{I}| r(1/K) + (|\mathscr{A}| - |\mathscr{I}|) r(y)}{2(|\mathscr{I}|+1) + 2\frac{|\mathscr{I}| + K(|\mathscr{A}| - |\mathscr{I}|)y}{K - (|\mathscr{I}|-1) - K(|\mathscr{A}| - |\mathscr{I}|)y} + |\mathscr{I}| r(1/K) + (|\mathscr{A}| - |\mathscr{I}|) r(y) + P_0} \tag{A I-31}$$

$$\overset{(c)}{<} \frac{|\mathscr{I}|r(1/K)+S_a r\left(\frac{1}{KS_a^3+K}\right)}{2\left(|\mathscr{I}|+1\right)+2\frac{|\mathscr{I}|+K\frac{S_a}{KS_a^3+K}}{K-(|\mathscr{I}|-1)-K\frac{S_a}{KS_a^3+K}}+|\mathscr{I}|r(1/K)+S_a r\left(\frac{1}{KS_a^3+K}\right)+P_0} \tag{A I-32}$$

$$\overset{(d)}{<} \frac{|\mathscr{I}|r(1/K)+r(\frac{1}{2K})}{2\left(|\mathscr{I}|+1\right)+2\frac{|\mathscr{I}|+K\frac{1}{2K}}{K-(|\mathscr{I}|-1)-K\frac{1}{2K}}+|\mathscr{I}|r(1/K)+r(\frac{1}{2K})+P_0} \tag{A I-33}$$

$$\overset{(e)}{<} \frac{|\mathscr{I}|r(1/K)+r(\frac{1}{K})}{2\left(|\mathscr{I}|+1\right)+2\frac{|\mathscr{I}|+K\frac{1}{K}}{K-(|\mathscr{I}|-1)-K\frac{1}{K}}+|\mathscr{I}|r(1/K)+r(\frac{1}{K})+P_0} \tag{A I-34}$$

$$= \frac{(|\mathscr{I}|+1)\log\left(1+\frac{1}{K}\right)}{2(|J|+1)+2\frac{|\mathscr{I}|+1}{K-|\mathscr{I}}+(|\mathscr{I}|+1)\log\left(1+\frac{1}{K}\right)+P_0} = \xi(|\mathscr{I}|+1) = \xi(|\mathscr{I}|+1) \tag{A I-35}$$

which establishes our claim above.

5.1 Proof of Proposition 7

Let us monitor the computation of first and second derivative of function $\eta(\mathbf{p}) = f(\mathbf{p})/g(\mathbf{p})$ with respect to each p_i, which is given by

$$\frac{\partial\left(f(\mathbf{p})/g(\mathbf{p})\right)}{\partial p_i} = \frac{\frac{\partial f(\mathbf{p})}{\partial p_i}g(\mathbf{p}) - f(\mathbf{p})\frac{\partial g(\mathbf{p})}{\partial p_i}}{g^2(\mathbf{p})} \tag{A I-36}$$

$$\frac{\partial^2\left(f(\mathbf{p})/g(\mathbf{p})\right)}{\partial p_i^2} = \frac{\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2}g^2(\mathbf{p}) - f(\mathbf{p})g(\mathbf{p})\frac{\partial^2 g(\mathbf{p})}{\partial p_i^2} - 2\frac{\partial f(\mathbf{p})}{\partial p_i}\frac{\partial g(\mathbf{p})}{\partial p_i}g(\mathbf{p}) + 2f(\mathbf{p})\left(\frac{\partial g(\mathbf{p})}{\partial p_i}\right)^2}{g^3(\mathbf{p})} \tag{A I-37}$$

where we can derive and bound $\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2}$ as

$$\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2} = -\frac{1}{(K+\sum_{k:(u_i,u_k)\in\mathscr{E}} KS_a^3 p_k + p_i)^2} + \sum_{j:(u_j,u_i)\in\mathscr{E}} \frac{(KS_a^3)^2 p_j(p_j + 2\sum_{l:(u_l,u_j)\in\mathscr{E}} KS_a^3 p_l + 2K)}{(p_j+\sum_{l:(u_l,u_j)\in\mathscr{E}} KS_a^3 p_l + K)^2(\sum_{l:(u_l,u_j)\in\mathscr{E}} KS_a^3 p_l + K)^2} \tag{A I-38}$$

$$\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2} > -\frac{1}{K^2} + \sum_{j:(u_j,u_i)\in\mathscr{E}} \frac{K(KS_a^3)^2 p_j}{K^4(S_a^4+2)^4} = \frac{\sum_{j:(u_j,u_i)\in\mathscr{E}} K(KS_a^3)^2 p_j - K^2(S_a^4+2)^4}{K^4(S_a^4+2)^4} \tag{A I-39}$$

Similarly, we can derive and bound $\frac{\partial f(\mathbf{p})}{\partial p_i}$ as

$$\frac{\partial f(\mathbf{p})}{\partial p_i} = \frac{1}{K+\sum_{k:(u_i,u_k)\in\mathscr{E}} KS_a^3 p_k + p_i} + \sum_{j:(u_j,u_i)\in\mathscr{E}} \frac{-KS_a^3 p_j}{(p_j+\sum_{l:(u_l,u_j)\in\mathscr{E}} KS_a^3 p_l + K)(\sum_{l:(u_l,u_j)\in\mathscr{E}} KS_a^3 p_l + K)} \tag{A I-40}$$

$$\frac{1}{K} > \frac{\partial f(\mathbf{p})}{\partial p_i} > \frac{1}{K(S_a^4+2)} - \frac{\sum_{j:(u_j,u_i)\in\mathscr{E}} KS_a^3 p_j}{K^2} \tag{A I-41}$$

$$f(\mathbf{p}) \leq \frac{S_a}{K} \tag{A I-42}$$

Finally, we can also derive and bound $\frac{\partial^2 g(\mathbf{p})}{\partial p_i^2}$ as

$$\frac{\partial^2 g(\mathbf{p})}{\partial p_i^2} = 0; \frac{\partial g(\mathbf{p})}{\partial p_i} = 2; g(\mathbf{p}) \geq P_0 \tag{A I-43}$$

The following special cases are in order:

- In case $\sum_{j:(u_j,u_i)\in\mathscr{E}} p_j > \frac{(S_a^4+2)^4 P_0 + 4K(S_a^4+2)^4}{KS_a^6 P_0}$, we can bound

$$\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2} > \frac{K^2(S_a^4+2)^4 + \frac{4K^3(S_a^4+2)^4}{P_0} - K^2(S_a^4+2)^4}{K^4(S_a^4+2)^4} = \frac{4}{K} > 0 \tag{A I-44}$$

$$\frac{\partial^2 (f(\mathbf{p})/g(\mathbf{p}))}{\partial p_i^2} > \frac{\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2} g(\mathbf{p}) - 2\frac{\partial f(\mathbf{p})}{\partial p_i}\frac{\partial g(\mathbf{p})}{\partial p_i}}{g^2(\mathbf{p})} > \frac{\frac{4}{K} - \frac{4}{K}}{g^2(\mathbf{p})} = 0 \tag{A I-45}$$

  Thus, $\frac{\partial^2 (f(\mathbf{p})/g(\mathbf{p}))}{\partial p_i^2}$ is positive, and $\eta(\mathbf{p})$ is convex with respect to p_i in the region of

$$\sum_{j:(u_j,u_i)\in\mathscr{E}} p_j > \frac{(S_a^4+2)^4 P_0 + 4K(S_a^4+2)^4}{KS_a^6 P_0}$$

.

- In case $\sum_{j:(u_j,u_i)\in\mathscr{E}} p_j \leq \frac{(S_a^4+2)^4P_0+4K(S_a^4+2)^4}{KS_a^6P_0}$, if we choose $P_0 > K$, $K > \frac{5(S_a^4+2)^5}{S_a^3}$ and $K > 5(S_a^4+2)^5 + 2S_a^4(S_a^4+2)$, then we have

$$\sum_{j:(u_j,u_i)\in\mathscr{E}} p_j \leq \frac{(S_a^4+2)^4P_0+4K(S_a^4+2)^4}{KS_a^6P_0} < \frac{5P_0(S_a^4+2)^4}{KS_a^6P_0} = \frac{5(S_a^4+2)^4}{KS_a^6} \tag{A I-46}$$

$$\frac{\partial f(\mathbf{p})}{\partial p_i} > \frac{1}{K(S_a^4+2)} - \frac{KS_a^3\frac{5(S_a^4+2)^4}{KS_a^6}}{K^2} = \frac{1}{K(S_a^4+2)} - \frac{5(S_a^4+2)^4}{K^2S_a^3} = \frac{KS_a^3-5(S_a^4+2)^5}{K^2S_a^3(S_a^4+2)} > 0 \tag{A I-47}$$

$$\begin{aligned}\frac{\partial\left(f(\mathbf{p})/g(\mathbf{p})\right)}{\partial p_i} &> \frac{\frac{KS_a^3-5(S_a^4+2)^5}{K^2S_a^3(S_a^4+2)}P_0 - 2\frac{S_a}{K}}{g^2(\mathbf{p})} \\ &> \frac{\frac{KS_a^3-5(S_a^4+2)^5}{P_0KS_a^3(S_a^4+2)}P_0 - 2\frac{S_a}{K}}{g^2(\mathbf{p})} = \frac{\frac{KS_a^3-5(S_a^4+2)^5-2S_a^4(S_a^4+2)}{KS_a^3(S_a^4+2)}}{g^2(\mathbf{p})} > 0\end{aligned} \tag{A I-48}$$

  Thus, $\frac{\partial(f(\mathbf{p})/g(\mathbf{p}))}{\partial p_i}$ is positive, and $\eta(\mathbf{p})$ is strictly increasing with respect to p_i in the region of

$$\sum_{j:(u_j,u_i)\in\mathscr{E}} p_j \leq \frac{(S_a^4+2)^4P_0+4K(S_a^4+2)^4}{KS_a^6P_0}.$$

By repeating the similar step, we also have the same results for the derivatives of $\eta(\mathbf{p})$ with respect to other variable $p_j, j \neq i$. In conclusion, $\eta(\mathbf{p})$ is either strictly increasing or convex with respect to each variable $p_i, \forall i \in \mathscr{S}_a$ at all the feasible value of $\mathbf{p} \in [0,1]^{S_a\times 1}$ when we choose $P_0 > K$ and $K > \frac{5(S_a^4+2)^5+2S_a^4(S_a^4+2)}{S_a^3}$. This completes the proof.

5.2 Explanations for inequality (a)-(e) in (A I-27)

- (a) is because the set of the optimal solution $\mathbf{v}^\star$ must satisfy the following optimization problem given the value of $\mathbf{p}^\star$

$$\min_{\mathbf{v}} 2\sum_{i\in\mathscr{F}} \|\mathbf{v}_i\|^2 + \sum_{i\in\mathscr{S}_a} R_i(\mathbf{v}) \tag{A I-49a}$$

$$\text{s.t. } R_i \geq \log(1+1/K), \forall i \in \mathscr{I} \tag{A I-49b}$$

$$R_j \geq \log(1+x_j), \forall j \in \mathscr{A} \setminus \mathscr{I} \tag{A I-49c}$$

$$R_k \geq 0, \forall k \in \{\mathscr{S}_a \setminus \mathscr{A}\} \cup \mathscr{M} \tag{A I-49d}$$

$$\text{(2.7e)} \tag{A I-49e}$$

We quickly note at optimality of (A I-49), all constraints (A I-49b)-(A I-49d) occur at equalities. We can prove this by following the proof for Proposition 1 in our manuscript. Thus problem (A I-49) is equivalent to

$$\min_{\mathbf{v}} \sum_{i \in \mathscr{F}} \|\mathbf{v}_i\|^2 \tag{A I-50a}$$

$$\text{s.t. } K|v_{i1}|^2 = \left(K \sum_{j \in \mathscr{S}_a \setminus i} |v_{j1}|^2 + K\right) 1/K, i \in \mathscr{I} \tag{A I-50b}$$

$$K|v_{j1}|^2 = \left(K \sum_{l \in \mathscr{S}_a \setminus j} |v_{j1}|^2 + K\right) x_j, j \in \mathscr{A} \setminus \mathscr{I} \tag{A I-50c}$$

$$K|v_{k1}|^2 = 0, \forall k \in \{\mathscr{S}_a \setminus \mathscr{A}\} \cup \mathscr{M} \tag{A I-50d}$$

$$\text{(2.7e)} \tag{A I-50e}$$

it can be seen that because the channel gain is chosen as $\mathbf{h}_i = \sqrt{K}[1,0,\ldots,0], \forall i \in \mathscr{S}_a$, each vector $\mathbf{v}_i$ only involves in its first component v_{i1}. Beside, $v_{ij} = 0, \forall i \in \mathscr{S}_a \cup \mathscr{M}, \forall j = 2,\ldots,N$ due to the minimization problem on the objective of (A I-50). In conclusion, the optimal $\mathbf{v}^\star$ should contain $v_{i1} \neq 0$ and $v_{ij} = 0, \forall i \in \mathscr{A}, \forall j = 2,\ldots,N$, and $\mathbf{v}_k = 0, \forall k \in \{\mathscr{S}_a \setminus \mathscr{A}\} \cup \mathscr{M}$. Therefore, we have

$$\sum_{i \in \mathscr{F}} \|\mathbf{v}_i^\star\|^2 = \sum_{i \in \mathscr{A}} |v_{i1}^\star|^2 = \sum_{i \in \mathscr{I}} \left(\sum_{j \in \mathscr{S}_a \setminus i} |v_{j1}^\star|^2 + 1\right) \frac{1}{K} + \sum_{i \in \mathscr{A} \setminus \mathscr{I}} \left(\sum_{j \in \mathscr{S}_a \setminus i} |v_{j1}^\star|^2 + 1\right) x_i \tag{A I-51}$$

$$= \sum_{i\in\mathscr{A}} \frac{(|\mathscr{I}|-1)|v_{i1}|^2}{K} + \frac{I}{K} + \sum_{k\in\mathscr{A}\setminus\mathscr{I}} \frac{|v_{k1}|^2}{K} + \sum_{k\in\mathscr{A}\setminus\mathscr{I}} \left(\sum_{i\in\mathscr{A}\setminus k} |v_{i1}|^2 \right) x_k + \sum_{k\in\mathscr{A}\setminus\mathscr{I}} x_k \tag{A I-52}$$

$$\overset{(\alpha)}{>} \sum_{i\in\mathscr{A}} \frac{(|\mathscr{I}|-1)|v_{i1}|^2}{K} + \frac{I}{K} + \sum_{k\in\mathscr{A}\setminus\mathscr{I}} \left(\sum_{i\in\mathscr{A}} |v_{i1}|^2 \right) x_k + \sum_{k\in\mathscr{A}\setminus\mathscr{I}} x_k \tag{A I-53}$$

where in (α), we have $\sum_{k\in\mathscr{A}\setminus\mathscr{I}} \frac{|v_{k1}|^2}{K} > \sum_{k\in\mathscr{A}\setminus\mathscr{I}} |v_{k1}|^2 x_k$ since $1/K \geq x_k, \forall k \in \mathscr{A}\setminus\mathscr{I}$. This subsequently leads to

$$\sum_{i\in\mathscr{A}} |v_{i1}^\star|^2 \left(K - (|\mathscr{I}|-1) - K \sum_{k\in\mathscr{A}\setminus\mathscr{I}} x_k \right) \geq |\mathscr{I}| + K \sum_{k\in\mathscr{A}\setminus\mathscr{I}} x_k \tag{A I-54}$$

$$\Leftrightarrow \sum_{i\in\mathscr{A}} |v_{i1}^\star|^2 \geq \frac{|\mathscr{I}| + K\sum_{k\in\mathscr{A}\setminus\mathscr{I}} x_k}{K - (|\mathscr{I}|-1) - K\sum_{k\in\mathscr{A}\setminus\mathscr{I}} x_k} \tag{A I-55}$$

- (b) is because the function on the right side of (a) is a monotonical increasing function with respect to each x_i. We can verify this property by seeing its first derivative with respect to each x_i is greater than

$$\frac{\frac{P_0}{1+x_i} - S_a r(1/K) \left(2\frac{K(K+1)}{(K-(|\mathscr{I}|-1)-K\sum_i x_i)^2} + \frac{1}{1+x_i} \right)}{\left(2(|\mathscr{I}|+1) + 2\frac{|\mathscr{I}|+K\sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} x_i}{K-(|\mathscr{I}|-1)-K\sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} x_i} + |\mathscr{I}| r(1/K) + \sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} r(x_i) + P_0 \right)^2} >$$
$$\frac{\frac{P_0}{2} - 2S_a(K+1)^2}{\left(2(|\mathscr{I}|+1) + 2\frac{|\mathscr{I}|+K\sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} x_i}{K-(|\mathscr{I}|-1)-K\sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} x_i} + |\mathscr{I}| r(1/K) + \sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} r(x_i) + P_0 \right)^2} \tag{A I-56}$$

since we treat $r(x_i) < r(1/K) < 1$, $0 < x_i < 1/K$, and $K - (|\mathscr{I}|-1) - K\sum_{i:u_i\in\mathscr{A}\setminus\mathscr{I}} x_i > 1$. The right side of (A I-56) is greater than 0 if we choose $P_0 > 4S_a(K+1)^2$. Thus, function on the right side of (a) is a monotonical increasing function with respect to each x_i and with $x_i < y$, we obtain the inequality (b).

- To explain (c), let us treat $|\mathscr{A}|-|\mathscr{I}|$ in (A I-31) as a variable z where $|\mathscr{A}|-|\mathscr{I}| \leq z \leq S_a$. The function in (A I-31) is indeed a monotonical increasing function with respect to z. We can verify this by seeing its first derivative with respect to z is greater than

$$\frac{P_0 r(y) - S_a r(1/K)\left(2\frac{Ky(K+1)}{(K-(|\mathscr{I}|-1)-Kzy)^2} + r(y)\right)}{\left(2(|\mathscr{I}|+1) + 2\frac{|\mathscr{I}|+Kzy}{K-(|\mathscr{I}|-1)-Kzy} + |\mathscr{I}|r(1/K) + zr(y) + P_0\right)^2} > \frac{P_0 r(y) - 2S_a(K+1)^2}{\left(2(|\mathscr{I}|+1) + 2\frac{|\mathscr{I}|+Kzy}{K-(|\mathscr{I}|-1)-Kzy} + |\mathscr{I}|r(1/K) + zr(y) + P_0\right)^2} \tag{A I-57}$$

since we treat $r(y) < r(1/K) < 1$, $y < \frac{1}{K}$ and $K-(|\mathscr{I}|-1)-Kzy > K-|\mathscr{A}|+1 > 1$. The right side of (A I-57) is greater than 0 if we choose $P_0 > \frac{2S_a(K+1)^2}{r(y)}$. Thus, the function in (A I-31) is indeed a monotonical increasing with respect to z and since $|\mathscr{A}|-|\mathscr{I}| < S_a$, we obtain the inequality (c).

- (d) is because the function on the right side of (c) is monotonically decreasing with respect to S_a in the region $S_a \geq 1$ and P_0 is chosen large [R.0-4, Appendix]. By denoting the function on the right side of (c) as $\eta_1(S_a) = \frac{f_1(S_a)}{g_1(S_a)}$ as the function with respect to S_a, we can bound its first derivative as follow

$$f_1'(S_a) = \left(S_a r\left(\frac{1}{KS_a^3+K}\right)\right)' = r\left(\frac{1}{KS_a^3+K}\right) - \frac{3KS_a^3}{(KS_a^3+K)(KS_a^3+k+1)} \tag{A I-58}$$

$$\overset{(i)}{<} r\left(\frac{1}{KS_a^3+K}\right) - \frac{KS_a^3+2K}{(KS_a^3+K)(KS_a^3+k+1)} \overset{(ii)}{<} \frac{-(K-1)}{(KS_a^3+K)(KS_a^3+k+1)} < 0 \tag{A I-59}$$

$$g_1'(S_a) = \left(S_a r\left(\frac{1}{KS_a^3+K}\right)\right)' + 2\left(\frac{I+\frac{S_a}{S_a^3+1}}{K-|\mathscr{I}|+1-\frac{S_a}{S_a^3+1}}\right)' \tag{A I-60}$$

$$= \left(S_a r\left(\frac{1}{KS_a^3+K}\right)\right)' + 2\frac{\frac{(1-2S_a^3)(K+1)}{(S_a^3+1)^2}}{(K-|\mathscr{I}|+1-\frac{S_a}{S_a^3+1})^2} \tag{A I-61}$$

$$\eta_1'(S_a) = \frac{f_1'(S_a)g_1(S_a) - f_1(S_a)g_1'(S_a)}{g_1^2(S_a)} \tag{A I-62}$$

$$< \frac{\left(S_a r\left(\frac{1}{KS_a^3+K}\right)\right)'\left(P_0 + S_a r\left(\frac{1}{KS_a^3+K}\right)\right) - f_1(S_a)g_1'(S_a)}{g_1^2(S_a)} \tag{A I-63}$$

$$= \frac{\left(S_a r\left(\frac{1}{KS_a^3+K}\right)\right)'\left(P_0 - |\mathscr{I}|r(\frac{1}{K})\right) - 2\frac{f_1(S_a)(1-2S_a^3)(K+1)}{(S_a^3+1)^2(K-|\mathscr{I}|+1-\frac{S_a}{S_a^3+1})^2}}{g_1^2(S_a)} \tag{A I-64}$$

$$= \frac{\frac{-(K-1)}{(KS_a^3+K)(KS_a^3+k+1)}\left(P_0 - |\mathscr{I}|r(\frac{1}{K})\right) + 2\frac{f_1(S_a)(2S_a^3-1)(K+1)}{(S_a^3+1)^2(K-|\mathscr{I}|+1-\frac{S_a}{S_a^3+1})^2}}{g_1^2(S_a)} \tag{A I-65}$$

$$\overset{(iii)}{<} \frac{\frac{-(K-1)}{(KS_a^3+K)(KS_a^3+k+1)}(P_0 - S_a)) + \frac{4S_a 2S_a^3(K+1)}{(S_a^3+1)^2}}{g_1^2(S_a)} \tag{A I-66}$$

since we treat $3KS_a^3 > KS_a^3 + 2KS_a^3 > KS_a^3 + 2K$ in (i) because $S_a \geq 1$. Beside we have $r\left(\frac{1}{KS_a^3+K}\right) < \frac{1}{KS_a^3+K} < 1$ in (ii). In (iii), we have $|\mathscr{I}| \leq S_a$ and $r(1/K) < 1$ and finally $K - |\mathscr{I}| + 1 - K\frac{S_a}{KS_a^3+K} > K - |\mathscr{I}| > 1$. $\eta_1(S_a)$ is negative if we choose

$$P_0 > \frac{8S_a^4(K+1)(KS_a^3+K)(KS_a^3+K+1)}{(S_a^3+1)^2(K-1)} + S_a.$$

Thus, the function on the right side of (c) is strictly decreasing with respect to $S_a \geq 1$, so that its maximum value is at $S_a = 1$ and we achieve the function on the right side of (d).

- (e) can be explained similar like (b). If we treat $1/(2K)$ as a variable t, we can argue that the function on the right side of (d) is monotonically increasing with respect to variable t by following the similar step as for (b). Thus, since $1/2K < 1/K$, we obtain the function on the right side of (e).

6. Proof of the Equivalence between (4.8) and (4.9) in Chapter 4

To prove that problems (4.8) and (4.9) are equivalent, we must prove that any feasible solution of (4.8) is also a feasible solution of (4.9). Conversely, from any feasible solution of (4.9), we can always find a feasible solution of (4.8). Assuming that $\bar{\mathscr{V}}, \bar{\mathscr{P}}, \bar{\mathscr{U}}, \bar{\mathscr{W}}, \bar{\mathscr{X}}, \bar{\mathscr{Y}}$ is a feasible solution of (4.9). It is easy to remark that since $\bar{\mathscr{V}}, \bar{\mathscr{P}}, \bar{\mathscr{U}}, \bar{\mathscr{W}}, \bar{\mathscr{X}}, \bar{\mathscr{Y}}$ satisfy all constraints

(4.9b)–(4.9e), (4.7d)–(4.7g), this subsequently leads to the fact that $\bar{\mathscr{V}}, \bar{\mathscr{P}}$ satisfy all constraints (2.7b)–(4.8d) of problem (4.8). On the other hand, assume that $\check{\mathscr{V}}, \check{\mathscr{P}}$ is a feasible solution of (2.7). Since $\check{\mathscr{V}}, \check{\mathscr{P}}$ satisfy all constraints (4.8b)–(4.8d), then we can find

$$\check{\mathrm{v}}_{ij}(t) = \check{\mu}_{ij}(t) = \log\left[1 + \frac{\check{p}_{ij}(t)\left|h_{iij}(t)\right|^2}{\check{I}_{ij}(\check{\mathscr{P}}(t))}\right] \tag{A I-67}$$

$$\check{w}_i(t) = \check{u}_i(t) = \log\left[1 + \frac{\left|\check{\mathbf{v}}_j(t)^H \mathbf{h}_j(t)\right|^2}{\sum\limits_{k\in\mathscr{F}\setminus j}\left|\check{\mathbf{v}}_k(t)^H \mathbf{h}_j(t)\right|^2 + N_{01}}\right]. \tag{A I-68}$$

It is obvious that $\check{\mathscr{V}}, \check{\mathscr{P}}$ and $\check{\mathscr{U}}, \check{\mathscr{W}}, \check{\mathscr{X}}, \check{\mathscr{Y}}$ also satisfy constraints (4.9b)–(4.9e), (4.7d)–(4.7g). Thus, $\check{\mathscr{V}}, \check{\mathscr{P}}, \check{\mathscr{U}}, \check{\mathscr{W}}, \check{\mathscr{X}}, \check{\mathscr{Y}}$ is a feasible solution of (4.9). This completes the proof.

7. Proof of Proposition 4 in Chapter 4

The proof of Proposition 4 is provided by observing that the set of feasible solutions of problem $(\mathscr{P}_{\text{con}})$ in (4.7) is a subset of the feasible solution of of problem $(\mathscr{P}_{\text{off}})$ in (4.8). To prove this, we consider two possible cases as follows:

- Case 1: Assume that $(\mathscr{P}_{\text{con}})$ has an optimal solution $\mathscr{V}(t)^\star, \mathscr{P}(t)^\star$ at time slot t such that this solution makes all constraints (4.8c) in $(\mathscr{P}_{\text{off}})$ all satisfied. It is easy to check that if $\mathscr{V}(t)^\star, \mathscr{P}(t)^\star$ satisfies all constraints (4.7b), (4.7d)–(4.7g), it also satisfies all constraints (4.8b), (4.8d). Thus, we can state that an optimal solution of $(\mathscr{P}_{\text{con}})$ is a feasible solution of $(\mathscr{P}_{\text{off}})$. In other words, the set of feasible solutions of problem $(\mathscr{P}_{\text{con}})$ is a subset of the feasible solution of $(\mathscr{P}_{\text{off}})$.

- Case 2: Assume that $(\mathscr{P}_{\text{con}})$ has an optimal solution $\mathscr{V}(t)^\star, \mathscr{P}(t)^\star$ at time slot t such that this solution makes at least one constraint (4.8c) in $(\mathscr{P}_{\text{off}})$ violated. Without loss of generality, we observe that the violation only occurs at $t \geq 3$. This can be verified since at time slot $t = 1$, both problems $(\mathscr{P}_{\text{con}})$ and $(\mathscr{P}_{\text{off}})$ are the same so that the violation cannot occur at time slot

1. Thus, let us assume that the violation occurs at constraint (4.8c) of problem $(\mathscr{P}_{\text{off}})$ at time index $t = 3$ and at some small cell index $i_1 \in \mathscr{S}_a$ (the same proof can be applied to other time and small cell indices), e.g.,

$$D_{i_1}(\mathscr{V}(3)^\star) + Q_{i_1}(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star) > C_{\max}, i_1 \in \mathscr{S}_a, \tag{A I-69}$$

$$D_l(\mathscr{V}(3)^\star) + Q_l(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star) \leq C_{\max}, l \in \mathscr{S}_a \setminus i_1. \tag{A I-70}$$

Then, we provide the following claim to support further proof.

Claim 1. *We claim that if* (A I-69)-(A I-70) *occur, problem* $\left(\mathscr{P}_{off}\right)$ *must contain solution* $\mathscr{V}(3)^\circ$ *such that*

$$\begin{cases} D_{i_1}(\mathscr{V}(3)^\circ) = C_{\max} - Q_{i_1}(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star) < D_{i_1}(\mathscr{V}(3)^\star), & i_1 \in \mathscr{S}_a, \\ D_l(\mathscr{V}(3)^\circ) = D_l(\mathscr{V}(3)^\star) \leq C_{\max} - Q_l(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star), & l \in \mathscr{S}_a \setminus i_1. \end{cases}$$

We can easily prove Claim 1 by examining the feasibility of the two following problems

$$\left(\mathscr{P}^1_{\min}\right): \min_{\mathscr{V}(3)} \quad \sum_{i \in \mathscr{F}} \|\mathbf{v}_i(3)\|^2 \tag{A I-71a}$$

$$s.t. \quad D_{i_1}(\mathscr{V}(3)) \geq C_{\max} - Q_{i_1}(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star), i_1 \in \mathscr{S}_a, \tag{A I-71b}$$

$$D_l(\mathscr{V}(3)) \geq D_l(\mathscr{V}(3)^\star), l \in \mathscr{S}_a \setminus i_1. \tag{A I-71c}$$

$$R_k(\mathscr{V}(3)) \geq R_{\min}, \forall k \in \mathscr{M}, \tag{A I-71d}$$

$$\sum_{k \in \mathscr{F}} \|\mathbf{v}_k(3)\|^2 \leq P_m, \tag{A I-71e}$$

and

$$\left(\mathscr{P}^2_{\min}\right): \min_{\mathscr{V}(3)} \quad \sum_{i \in \mathscr{F}} \|\mathbf{v}_i(3)\|^2 \tag{A I-72a}$$

$$s.t. \quad D_{i_1}(\mathscr{V}(3)) \geq D_{i_1}(\mathscr{V}(3)^\star), i_1 \in \mathscr{S}_a, \tag{A I-72b}$$

(A I-71c) – (A I-71e) (A I-72c)

We remark that the feasible solution set of problem (A I-72) *is a subset of the feasible solution set of problem* (A I-71). *This can be explained because problem* (A I-72) *has a set of higher minimum rate requirement than problem* (A I-25)*, e.g.,* $D_{i_1}(\mathscr{V}(3)^\star) > C_{\max} - Q_{i_1}(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star)$, $i_1 \in \mathscr{S}_a$*, so that any feasible solution of problem* (A I-72) *is also feasible to problem* (A I-71). *Beside, the optimal solution of problem* (A I-72) *is* $\mathscr{V}(3)^\star$*, implies that problem* (A I-72) *is feasible, which subsequently leads to the feasibility of problem* (A I-71). *Thus, problem* (A I-71) *is feasible and at the optimality, constraints* (A I-71b)-(A I-71d) *occur at equalities. This establishes Claim 1.*

The result of Claim 1 allows us to recast problems $(\mathscr{P}_{\text{off}})$ and $(\mathscr{P}_{\text{con}})$ up to time index 3 as

$$\left(\tilde{\mathscr{P}}_{\text{off}}\right): \max_{\mathscr{P}(3)} \quad \sum_{(i,j)\in(\mathscr{S}_a,\mathscr{S}_u)} d_{ij}(\mathscr{P}(3)) \tag{A I-73a}$$

$$\text{s.t.} \quad C_{\max} \geq \sum_{j\in\mathscr{S}_u} d_{i_1 j}(\mathscr{P}(3)), i_1 \in \mathscr{S}_a, \tag{A I-73b}$$

$$D_l(\mathscr{V}(3)^\circ) + Q_l(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star) \geq \sum_{j\in\mathscr{S}_u} d_{lj}(\mathscr{P}(3)), l \in \mathscr{S}_a \setminus i_1 \tag{A I-73c}$$

$$\sum_{j\in\mathscr{S}_u} p_{ij}(3) \leq p_m, \forall\,(i,j) \in (\mathscr{S}_a, \mathscr{S}_u), \forall k \in \mathscr{M}, \tag{A I-73d}$$

and

$$\left(\tilde{\mathscr{P}}_{\text{con}}\right): \max_{\mathscr{P}(3)} \quad \sum_{(i,j)\in(\mathscr{S}_a,\mathscr{S}_u)} d_{ij}(\mathscr{P}(3)) \tag{A I-74a}$$

$$\text{s.t.} \quad C_{\max} \geq D_{i_1}(\mathscr{V}(3)^\star) \geq \sum_{j\in\mathscr{S}_u} d_{i_1 j}(\mathscr{P}(3)), i_1 \in \mathscr{S}_a, \tag{A I-74b}$$

$$D_l(\mathscr{V}(3)^\star) + Q_l(\mathscr{V}(1)^\star, \mathscr{P}(1)^\star) \geq$$

$$\sum_{j\in\mathscr{S}_u} d_{lj}(\mathscr{P}(3)), l \in \mathscr{S}_a \setminus i_1. \tag{A I-74c}$$

$$\sum_{j\in\mathscr{S}_u} p_{ij}(3) \leq p_m, \forall (i,j) \in (\mathscr{S}_a, \mathscr{S}_u), \forall k \in \mathscr{M}, \tag{A I-74d}$$

We quickly observe that both $\left(\tilde{\mathscr{P}}_{\text{off}}\right)$ and $\left(\tilde{\mathscr{P}}_{\text{con}}\right)$ are almost similar, except for (A I-73b) and (A I-74b). Particularly, we note from constraints (A I-73b) and (A I-74b) that $\left(\tilde{\mathscr{P}}_{\text{off}}\right)$ offers a higher buffer capacity $C_{\max}$ than the buffer capacity shown in constraint (A I-74b) of $\left(\tilde{\mathscr{P}}_{\text{con}}\right)$. Thus, it is obvious that any feasible solution of problem $\left(\tilde{\mathscr{P}}_{\text{con}}\right)$ is also feasible to $\left(\tilde{\mathscr{P}}_{\text{off}}\right)$. Therefore, the optimal objective of $\left(\tilde{\mathscr{P}}_{\text{off}}\right)$ is greater than or equal to the optimal objective of $\left(\tilde{\mathscr{P}}_{\text{con}}\right)$. This completes the proof.

8. Proof of Proposition 5 in Chapter 5

By examining the rate formula $\tilde{R}_{j,j}(\mathbf{w})$ in (5.16) and $R_{j,j}(\mathbf{w},\mathbf{a})$ in (5.4), where $j \in \mathscr{F}$, we observe that for all value of $\mathbf{w}$, $R_{j,j}(\mathbf{w},\mathbf{a}) \geq \tilde{R}_{j,j}(\mathbf{w}), \forall j \in \mathscr{S}$. Therefore, problem $(\mathscr{P}_1)$ can provide a tuple of backhaul rates no less than the optimal solution of $(\mathscr{P}_{\text{non}})$. Beside, by comparing the rate formula of $r_j(\mathbf{v}), \forall j \in \mathscr{S}$, we observe that when $v_{kj} = 0, \forall k, j \in \mathscr{S}, k \neq j$, the rate formula of $r_j(\mathbf{v})$ becomes similar to $\tilde{r}_j(\mathbf{v}), \forall j \in \mathscr{S}$. From this observation, let us denote the optimal solution of $(\mathscr{P}_{\text{non}})$ as $\mathbf{w}^\star, \mathbf{v}^\star$, where the set of power allocation $v^\star_{kj} = 0, \forall k, j \in \mathscr{S}, k \neq j$. We easily examine that this optimal solution of $(\mathscr{P}_{\text{non}})$ is a feasible solution of $(\mathscr{P}_1)$. However, the optimal solution of $(\mathscr{P}_1)$ does not always satisfy the constraints of $(\mathscr{P}_{\text{non}})$. This means that the set of feasible solution of $(\mathscr{P}_{\text{non}})$ is always a subset of the set of feasible solution of $(\mathscr{P}_1)$. Therefore, the optimal solution of $(\mathscr{P}_1)$ is greater or equal to the optimal solution of $(\mathscr{P}_{\text{non}})$.

9. Proof of Proposition 6 in Chapter 5

To prove Proposition 6, let us review the smoothness property of a continuous function $f(x)$: $\mathbb{R}^B \to \mathbb{R}$, where B is the dimension of the vector of variable x. According to (Sohrab, 2014), $f(x)$ is L-smooth, where L is the Lipschitz constant of $\partial f(x)$ (derivative of function $f(x)$ with respect to x), if and only if

$$\|\partial f(x_1) - \partial f(x_2)\|_2 \leq L \|x_1 - x_2\|_2, \forall x_1, x_2 \in \mathbb{R}^A. \tag{A I-75}$$

Based on this property, we show that $\ell(x) = \log(1+x), \forall x \in R$ is a smooth function with Lipschitz constant $L = 1$. By taking the first derivative of $\ell(x)$, we obtain

$$|\partial f(x_1) - \partial f(x_2)| = \left| \frac{1}{1+x_1} - \frac{1}{1+x_2} \right| = \left| \frac{x_2 - x_1}{(1+x_1)(1+x_2)} \right| \leq |x_1 - x_2| \tag{A I-76}$$

since we have $(1+x_1)(1+x_2) > 1$. Then, from the result of (Parikh & Boyd, 2014a), we have

$$\log(1+x) \geq \log\left(1+x^{(n)}\right) + \frac{x - x^{(n)}}{1+x^{(n)}} - \frac{L}{2}\left(x - x^{(n)}\right)^2 \tag{A I-77}$$

where $x, x^{(n)} \in \mathbb{R}$ and $L = 1$. This completes the proof.

10. Proof of Theorem 1 in Chapter 6

Let us replace $\tilde{\mathbf{v}}_t$ by $\mathbf{v}$ in $r_{nj}\left(\tilde{\mathbf{v}}_t, \Delta_{nj}\right)$ for simplicity and rewrite $r_{nj}\left(\mathbf{v}, \Delta_{nj}\right)$ as

$$r_{nj}\left(\mathbf{v}, \Delta_{nj}\right) = \log\left(\sum_{\substack{l \in \mathscr{U}_m \\ m=1,2,3}} \left|\mathbf{v}_{ml}^H \mathbf{g}_{mnj}(\Delta_{mnj})\right|^2 + N_0 \right) - \log\left(\sum_{(m,l) \neq (n,i)} \left|\mathbf{v}_{ml}^H \mathbf{g}_{mnj}(\Delta_{mnj})\right|^2 + N_0 \right) \tag{A I-78}$$

The derivative of $r_{nj}\left(\mathbf{v},\Delta_{nj}\right)$ is defined as $\nabla r_j\left(\mathbf{v},\Delta_{nj}\right)=\left[\nabla_{\mathbf{v}}r_j\left(\mathbf{v},\Delta_{nj}\right);\nabla_{\Delta_{nj}}r_j\left(\mathbf{v},\Delta_{nj}\right)\right]$, which are given by

$$\nabla_{\mathbf{v}}r_j\left(\mathbf{v},\Delta_{nj}\right)=\underbrace{\frac{2\mathbf{v}^H\mathbf{G}_{nj}\left(\Delta_{nj}\right)}{\mathbf{v}^H\mathbf{G}_{nj}\left(\Delta_{nj}\right)\mathbf{v}+N_0}}_{\bar{h}_1(\mathbf{v},\Delta_{nj})\bar{h}_2(\mathbf{v},\Delta_{nj})}-\underbrace{\frac{2\mathbf{v}^H\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)}{\mathbf{v}^H\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)\mathbf{v}+N_0}}_{\underline{\mathrm{h}}_1(\mathbf{v},\Delta_{nj})\underline{\mathrm{h}}_2(\mathbf{v},\Delta_{nj})} \tag{A I-79}$$

$$\nabla_{\Delta_{nj}}r_j\left(\mathbf{v},\Delta_{nj}\right)=\underbrace{\frac{2\Delta_{nj}^H\Omega_{nj}\left(\mathbf{v}\right)}{\mathbf{v}^H\mathbf{G}_{nj}\left(\Delta_{nj}\right)\mathbf{v}+N_0}}_{\bar{h}_3(\mathbf{v},\Delta_{nj})\bar{h}_2(\mathbf{v},\Delta_{nj})}-\underbrace{\frac{2\Delta_{nj}^H\tilde{\Omega}_{nj}\left(\mathbf{v}\right)}{\mathbf{v}^H\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)\mathbf{v}+N_0}}_{\underline{\mathrm{h}}_3(\mathbf{v},\Delta_{nj})\underline{\mathrm{h}}_2(\mathbf{v},\Delta_{nj})} \tag{A I-80}$$

where in (A I-79) and (A I-80), we denote $\hat{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)=\mathbf{g}_{nj}(\Delta_{nj})\left(\mathbf{g}_{nj}(\Delta_{nj})\right)^H$ to support the notation of

$$\mathbf{G}_{nj}\left(\Delta_{nj}\right)=\mathrm{Bdiag}\underbrace{\left(\hat{\mathbf{G}}_{nj}\left(\Delta_{nj}\right),\ldots,\hat{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)\right)}_{U\text{ elements}}$$

and

$$\tilde{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)=\mathrm{Bdiag}(\hat{\mathbf{G}}_{nj}\left(\Delta_{nj}\right),...,0,...,\hat{\mathbf{G}}_{nj}\left(\Delta_{nj}\right)),$$

where 0 appears at the j^{th} element. Beside, in (A I-80), we denote $\omega_{n,j,k}(\mathbf{v})=\mathbf{v}_k\circ\tilde{\mathbf{g}}_{nj}$, $\Omega_{n,j,k}(\mathbf{v})=\omega_{n,j,k}(\mathbf{v})\left(\omega_{n,j,k}(\mathbf{v})\right)^H$. Then, we denote $\Omega_{nj}(\mathbf{v})=\sum_{k\in\mathscr{U}_n}\Omega_{n,j,k}(\mathbf{v})$ and $\tilde{\Omega}_{nj}\left(\mathbf{v}\right)=\sum_{k\in\mathscr{U}\setminus j}\Omega_{n,j,k}(\mathbf{v})$. Next, we will find the Lipschitz constants for $\nabla_{\mathbf{v}}r_j\left(\mathbf{v},\Delta_{nj}\right)$ and $\nabla_{\Delta_{nj}}r_j\left(\mathbf{v},\Delta_{nj}\right)$. First, we have

$$\begin{aligned}
&\left\|\bar{h}_1(\mathbf{v},\Delta_{nj})-\bar{h}_1(\mathbf{v}^{(0)},\Delta_{nj}^{(0)})\right\|\\
&=\left\|2\mathbf{v}^H\mathbf{G}_{nj}(\Delta_{nj})-2\mathbf{v}^H\mathbf{G}_{nj}(\Delta_{nj}^{(0)})+2\mathbf{v}^H\mathbf{G}_{nj}(\Delta_{nj}^{(0)})-2\mathbf{v}^{(0)H}\mathbf{G}_{nj}(\Delta_{nj}^{(0)})\right\|\\
&\leq 2\left\|\mathbf{v}\right\|\left\|\mathrm{blkdiag}\left(\left(\mathbf{g}_{nj}(\Delta_{nj})-\mathbf{g}_{nj}(\Delta_{nj}^{(0)})\right)\left(\mathbf{g}_{nj}(\Delta_{nj})\right)^H\right)\right\|_F\\
&+2\left\|\mathbf{v}\right\|\left\|\mathrm{blkdiag}\left(\mathbf{g}_{nj}(\Delta_{nj}^{(0)})\left(\mathbf{g}_{nj}(\Delta_{nj})-\mathbf{g}_{nj}(\Delta_{nj}^{(0)})\right)^H\right)\right\|_F+2\left\|\mathbf{v}-\mathbf{v}^{(0)}\right\|\left\|\mathbf{G}_{nj}\left(1/H\right)\right\|_F\\
&\leq 2\left\|\mathbf{v}\right\|2U\left\|\mathbf{g}_{nj}(1/H)\right\|\left\|\tilde{\mathbf{g}}_{nj}\right\|_\infty\left\|\Delta_{nj}-\Delta_{nj}^{(0)}\right\|+2\left\|\mathbf{v}-\mathbf{v}^{(0)}\right\|\left\|\mathbf{G}_{nj}\left(1/H\right)\right\|_F\\
&\leq A_1\left(\left\|\Delta_{nj}-\Delta_{nj}^{(0)}\right\|+\left\|\mathbf{v}-\mathbf{v}^{(0)}\right\|\right)
\end{aligned} \tag{A I-81}$$

where $A_1 = \max\{4U^2 p^{\max} ||\mathbf{g}_{nj}(1/H)|| ||\tilde{\mathbf{g}}_{nj}||_\infty,\ 2||\mathbf{G}_{nj}(1/H)||_F\}$. In other words, the Lipschitz constant of $\bar{h}_1(\mathbf{v}, \Delta_{nj})$ is A_1 (Parikh & Boyd, 2014a). Similarly, we can derive

$$\left\| \bar{h}_2(\mathbf{v}, \Delta_{nj}) - \bar{h}_2(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) \right\| \leq B_1 \left(\left\| \Delta_{nj} - \Delta_{nj}^{(0)} \right\| + \left\| \mathbf{v} - \mathbf{v}^{(0)} \right\| \right) \tag{A I-82}$$

where $B_1 = 1/N_0^2 (A_1/2 + U p^{\max} \left\| \mathbf{G}_{nj}(1/H) \right\|_F$. Using (A I-81) and (A I-82), we can find the Lipschitz constant $\mathscr{A}_1$ of $\bar{h}_1(\mathbf{v}, \Delta_{nj}) \bar{h}_2(\mathbf{v}, \Delta_{nj})$ as

$$\begin{aligned} &\left\| \bar{h}_1(\mathbf{v}, \Delta_{nj}) \bar{h}_2(\mathbf{v}, \Delta_{nj}) - \bar{h}_1(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) \bar{h}_2(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) \right\| \\ &\leq \underbrace{\left(2U p^{\max} \left\| \mathbf{G}_{nj}(1/H) \right\|_F B_1 + 1/N_0 A_1 \right)}_{\mathscr{A}_1} \left(\left\| \Delta_{nj} - \Delta_{nj}^{(0)} \right\| + \left\| \left(\mathbf{v} - \mathbf{v}^{(0)} \right) \right\| \right) \end{aligned} \tag{A I-83}$$

Next, we can derive the Lipschitz constant of $\bar{h}_3(\mathbf{v}, \Delta_{nj})$ as

$$\left\| \bar{h}_3(\mathbf{v}, \Delta_{nj}) - \bar{h}_3(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) \right\| \leq C_1 \left(\| \Delta_{nj} - \Delta_{nj}^{(0)} \| + \left\| \left(\mathbf{v} - \mathbf{v}^{(0)} \right) \right\| \right) \tag{A I-84}$$

where $C_1 = \max \left\{ 4U^2/H \left\| \tilde{\mathbf{g}}_{nj} \right\|_\infty^2 p_{\max}, 2U p_{\max} \left\| \tilde{\mathbf{g}}_{nj} \right\|_\infty \right\}$. Then, we can also find the Lipschitz constant $\mathscr{B}_1$ of of $\bar{h}_3(\mathbf{v}, \Delta_{nj}) \bar{h}_2(\mathbf{v}, \Delta_{nj})$ as

$$\begin{aligned} &\left\| \bar{h}_3(\mathbf{v}, \Delta_{nj}) \bar{h}_2(\mathbf{v}, \Delta_{nj}) - \bar{h}_3(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) \bar{h}_2(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) \right\| \\ &\qquad \leq \underbrace{\left(2U/H 2U p^{\max} \left\| \tilde{\mathbf{g}}_{nj} \right\|_\infty B_1 + 1/N_0 C_1 \right)}_{\mathscr{B}_1} \left(\left\| \Delta_{nj} - \Delta_{nj}^{(0)} \right\| + \left\| \left(\mathbf{v} - \mathbf{v}^{(0)} \right) \right\| \right) \end{aligned} \tag{A I-85}$$

By following the same steps, we can also conclude that the Lipschitz constants for the terms $\underline{\mathrm{h}}_k(\mathbf{v}, \Delta_{nj}), k = 1 \ldots, 3$ are A_1, B_1, C_1, respectively. Finally, we have

$$\left\| \nabla r_{nj}\left(\mathbf{v}, \Delta_{nj}\right) - \nabla r_{nj}\left(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}\right) \right\| \leq \xi_0 \left(\left\| \Delta_{nj} - \Delta_{nj}^{(0)} \right\| + \left\| \left(\mathbf{v} - \mathbf{v}^{(0)} \right) \right\| \right) \tag{A I-86}$$

where $\xi_0 = 2 \times \max\{\mathscr{A}_1, \mathscr{B}_1\}$. We now show that $f_i(\mathbf{v}, \delta_i)$ is strongly convex. According to (A I-86), $r_j\left(\mathbf{v}, \Delta_{nj}\right)$ is ξ_0–smooth. Thus, we have

$$\left| r_{nj}(\mathbf{v}, \Delta_{nj}) - r_{nj}(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) - \nabla r_{nj}(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)})^T([\mathbf{v}, \Delta_{nj}] - [\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}]) \right| \leq \frac{\xi_0}{2}(||\Delta_{nj} - \Delta_{nj}^{(0)}||^2 + ||\mathbf{v} - \mathbf{v}^{(0)}||^2) \quad \text{(A I-87)}$$

and

$$r_{nj}\left(\mathbf{v}, \Delta_{nj}\right) \geq -\frac{\xi_0}{2}\left(\left\|\Delta_{nj} - \Delta_{nj}^{(0)}\right\|^2 + \left\|\left(\mathbf{v} - \mathbf{v}^{(0)}\right)\right\|^2\right) + r_{nj}(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) + \nabla r_{nj}(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)})^T\left([\mathbf{v}, \Delta_{nj}] - [\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}]\right) \quad \text{(A I-88)}$$

Due to the strong convexity of $\xi_{ni}(\|\mathbf{v}\|^2 + \left\|\Delta_{nj}\right\|^2)$, we have

$$\xi_{nj}(\|\mathbf{v}\|^2 + \left\|\Delta_{nj}\right\|^2) \geq \xi_{nj}\left(\left\|\mathbf{v}^{(0)}\right\|^2 + \left\|\Delta_{nj}^{(0)}\right\|^2\right) + 2\xi_{nj}[\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}]^T\left([\mathbf{v}, \Delta_{nj}] - [\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}]\right) + \frac{\xi_{nj}}{2}\left(\left\|\Delta_{nj} - \Delta_{nj}^{(0)}\right\|^2 + \left\|\left(\mathbf{v} - \mathbf{v}^{(0)}\right)\right\|^2\right) \quad \text{(A I-89)}$$

Summing these two inequalities (A I-88)–(A I-89), we obtain

$$f_{nj}(\mathbf{v}, \Delta_{nj}) \geq \frac{\xi_{nj} - \xi_0}{2}\left(\left\|\Delta_{nj} - \Delta_{nj}^{(0)}\right\|^2 + \left\|\mathbf{v} - \mathbf{v}^{(0)}\right\|^2\right) + f_{nj}(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}) + \nabla f_{nj}(\mathbf{v}^{(0)}, \Delta_{nj}^{(0)})^T\left([\mathbf{v}, \Delta_{nj}] - [\mathbf{v}^{(0)}, \Delta_{nj}^{(0)}]\right) \quad \text{(A I-90)}$$

which implies that $f_{nj}(\mathbf{v}, \Delta_{nj})$ is strongly convex with factor $\xi_{nj} - \xi_0$. Similarly, we also conclude that $g_{nj}(\mathbf{v}, \Delta_{nj})$ is strongly concave with factor $\xi_{nj} - \xi_0$. This completes the proof.

BIBLIOGRAPHY

Al-Imari, M., Xiao, P., Imran, M. A. & Tafazolli, R. (2014). Uplink non-orthogonal multiple access for 5G wireless networks. *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS 2014)*, pp. 781–785.

Andersin, M., Rosberg, Z. & Zander, J. (1995). Gradual removals in cellular PCS with constrained power control and noise. *Proc. IEEE Personal, Indoor and Mobile Radio Communications (PIMRC'95)*, pp. 56–60.

Andrews *et al.* (2014). What Will 5G Be? *IEEE J. Sel. Areas Commun.*, 32(6), 1065–1082.

Auer *et al.* (2011). How much energy is needed to run a wireless network? *IEEE Wireless Comm.*, 18(5), 1536-1284.

Bastug, E., Bennis, M. & Debbah, M. Living on the edge: The role of proactive caching in 5G wireless networks.

Beck, A., Ben-Tal, A. & Tetruashvili, L. (2010). A Sequential Parametric Convex Approximation Method with Application to Nonconvex Truss Topology Design Problem. *J. Global Opti.*, 47(1), 29–51.

Ben-Tal, A. & Nemirovski, A. (2001a). On the polyhedral approximations of the second-order cone. *Math. of Operations Research*, 26(2), 193–205.

Ben-Tal, A. & Nemirovski, A. (2001b). Lectures on modern convex optimization: analysis, algorithms, and engineering applications. *Society for Industrial and Applied Mathematics (SIAM)*.

Bharadia, D., McMilin, E. & Katti, S. (2013). Full duplex radios. *Proc. ACM SIGCOMM 2013*, pp. 375-386.

Bhushan *et al.* (2014). Network densification: The dominant theme for wireless evolution into 5G. *IEEE Commun. Mag.*, 52(2), 82–89.

Boccardi, F., Heath, R. W., Lozano, A., Marzetta, T. L. & Popovski, P. (2014). Five disruptive technology directions for 5G. *IEEE Commun. Mag.*, 52(2), 74–80.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge, UK: Cambridge University Press.

Cao, M., Wang, X., Kim, S.-J. & Madihian, M. (2007). Multi-hop wireless backhaul networks: a cross-layer design paradigm. *IEEE J. Sel. Areas Commun.*, 25(4), 738–748.

Chandrasekhar, V., Andrews, J. G. & Gatherer, A. (2008). Femtocell networks: A survey. *IEEE Commun. Mag.*, 46(9), 59-67.

Chandrasekhar, V. C., Kountouris, M. & Andrews, J. G. (2009). Coverage in multi-antenna two-tier networks. *IEEE Trans. Wireless Commun.*, 8(10), 5314-5327.

Chan. et al., P. (2006). The evolution path of 4G networks: FDD or TDD? *IEEE Commun. Mag.*, 44(12), 42–50.

Chen, D. C., Quek, T. Q. S. & Kountouris, M. (2015). Backhauling in Heterogeneous Cellular Networks: Modeling and Tradeoffs. *IEEE Trans. Wireless Commun.*, 14(6), 3194–3206.

Chen, M., Saad, W., Yin, C. & Debbah, M. (2017). Echo State Networks for Proactive Caching in Cloud-Based Radio Access Networks With Mobile Users. *IEEE Trans. Wireless Commun.*, 16(6), 3520–3535.

Cheng, Y., Pesavento, M. & Philipp, A. (2013). Joint Network Optimization and Downlink Beamforming for CoMP Transmissions Using Mixed Integer Conic Programming. *IEEE Trans. Signal Process.*, 61(16), 3972-3987.

Chiang, M., Tan, C. W., Palomar, D. P., O'Neill, D. & Julian, D. (2007). Power Control By Geometric Programming. *IEEE Trans. Wireless Commun.*, 6(7), 2640–2651.

Choi, J. (2015). Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems. *IEEE Trans. Commun.*, 63(3), 791–800.

Cui, J., Liu, Y., Ding, Z., Fan, P. & Nallanathan, A. (2018). Optimal User Scheduling and Power Allocation for Millimeter Wave NOMA Systems. *IEEE Trans. Wireless Commun.*, 17(3), 1502–1517.

Dahrouj, H. & Yu, W. (2010). Coordinated Beamforming for the Multicell Multi-Antenna Wireless System. *IEEE Trans. Wireless Commun.*, 9(5), 1748–1759.

Dahrouj *et al.* (2015). Cost-effective hybrid RF/FSO backhaul solution for next generation wireless systems. 22(5), 98-104.

Ding, Z., Yang, Z., Fan, P. & Poor, H. V. (2014). On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users. *IEEE Signal Process. Lett.*, 21(12), 1501–1505.

Ding, Z., Peng, M. & Poor, H. V. (2015). Cooperative Non-Orthogonal Multiple Access in 5G Systems. *IEEE Commun. Lett.*, 19(8), 1462–1465.

Ding, Z., Schober, R. & Poor, H. V. (2016). A General MIMO Framework for NOMA Downlink and Uplink Transmission Based on Signal Alignment. *IEEE Trans. Wireless Commun.*, 15(6), 4438–4454.

Dinkelbach, W. (1967). On linear fractional programming. *Bulletin of the Australian Mathematical Society*, 13, 492-498.

Foschini, G. J. & Miljanic, Z. (1993). A simple distributed autonomous power control algorithm and its convergence. *IEEE Trans. Veh. Technol.*, 42(4), 641–646.

Frew, E. W. & Brown, T. X. (2008). Airborne Communication Networks for Small Unmanned Aircraft Systems. *Proc. IEEE*, 96(12), 2008-2027.

Gao *et al.* (2015). MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network. 22(5), 13–21.

Ge, X., Cheng, H., Guizani, M. & Han, T. (2014). 5G wireless backhaul networks: challenges and research advances. *IEEE Network*, 28(6), 6–11.

Ge, X., Tu, S., Mao, G., Wang, C.-X. & Han, T. (2016). 5G Ultra-Dense Cellular Networks. 23(1), 72-79.

Georgiadis, L., Neely, M. J. & Tassiulas, L. (2006). Resource Allocation and Cross-Layer Control in Wireless Networks. *Foundations and Trends in Networking*, 1(1), 1–144.

Gershman, A. B., Sidiropoulos, N. D., Shahbazpanahi, S., Bengtsson, M. & Ottersten, B. (2010). Convex Optimization-Based Beamforming. *IEEE Signal Process. Mag.*, 27(3), 62-75.

Ghosh *et al.* (2012). Heterogeneous cellular networks: From theory to practice. *IEEE Commun. Mag.*, 50(6), 54-64.

Günlük, O. & Linderoth, J. (2010). Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Mathematical Programming*, 124(1-2), 183-205.

Goldsmith, A. (2005). *Wireless Communications*. Cambridge, UK: Cambridge University Press.

Golrezaei *et al.* (2012). FemtoCaching: Wireless video content delivery through distributed caching helpers. *Proc. IEEE INFOCOM 2012*.

Goonewardena, M. P. (2017). *Distributed Radio Resource Allocation in Wireless Heterogeneous Networks*. (Ph.D. thesis, École de technologie Superieure, Montréal, Canada).

Grippo, L. & Sciandrone, M. (2000). On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations Research Letters*, 26(3), 127–136.

Guo, Y., Yang, Q., Liu, J. & Kwak, K. S. Cross-Layer Rate Control and Resource Allocation in Spectrum-Sharing OFDMA Small-Cell Networks With Delay Constraints.

Guo, Y., Yang, Q., Liu, J. & Kwak, K. S. (2017). Cross-Layer Rate Control and Resource Allocation in Spectrum-Sharing OFDMA Small-Cell Networks With Delay Constraints. *IEEE Trans. Veh. Technol.*, 66(5), 4133 – 4146.

Gupta, L., Jain, R. & Vaszkun, G. (2016). Survey of Important Issues in UAV Communication Networks. *IEEE Commun. Surveys Tuts.*, 18(2), 1123–1152.

Ha, V. N., Nguyen, D. & Le, L. B. (2015). Sparse precoding design for Cloud-RANs sum-rate maximization. *Proc. IEEE Wireless Comm. Net. Conf. (WCNC'15)*, pp. 1648 - 1653.

Hamidouche, K., Saad, W., Debbah, M., Song, J. B. & Hong, C. S. (2017). The 5G Cellular Backhaul Management Dilemma: To Cache or to Serve. *IEEE Trans. Wireless Commun.*, 16(8), 4866–4879.

Hanif, M. F., Ding, Z., Ratnarajah, T. & Karagiannidis, G. K. (2016). A Minorization-Maximization Method for Optimizing Sum Rate in the Downlink of Non-Orthogonal Multiple Access Systems. *IEEE Trans. Signal Process.*, 64(1), 76–88.

Heath *et al.* (2016). An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems. 10(3), 436–453.

He *et al.* (2014). Leakage-Aware Energy-Efficient Beamforming for Heterogeneous Multicell Multiuser Systems. *IEEE J. Sel. Areas Commun.*, 32(6), 1268–1281.

Hong, M., Sun, R., Baligh, H. & Luo, Z.-Q. (2013). Joint Base Station Clustering and Beamformer Design for Partial Coordinated Transmission in Heterogeneous Networks. *IEEE J. Sel. Areas Commun.*, 31(2), 226–240.

Hou, X. & Yang, C. (2011). How much feedback overhead is required for base station cooperative transmission to outperform non-cooperative transmission? *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*.

Hoydis, J., Hosseini, K., ten Brink, S. & Debbah, M. (2013). Making smart use of excess antennas: Massive MIMO, Small cells, and TDD. *Bell Labs Tech. J.*, 18(2), 5–20.

Hu, B., Hua, C., Chen, C. & Guan, X. (2016). Multicast Beamforming for Wireless Backhaul with User-centric Clustering in Cloud-RANs. *Proc. IEEE Int. Conf. Communications (ICC'16)*, pp. 1-6.

Hu, R. Q. & Qian, Y. (2014). An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems. *IEEE Commun. Mag.*, 52(5), 94-101.

Huang, Y., He, S., Jin, S. & Chen, W. (2014). Decentralized energy-efficient coordinated beamforming for multicell systems. *IEEE Trans. Veh. Technol.*, 63(9), 4302-4314.

Huang, Y., Tan, C. W. & Rao, B. D. (2013). Joint Beamforming and Power Control in Coordinated Multicell: Max-Min Duality, Effective Network and Large System Transition. *IEEE Trans. Wireless Commun.*, 12(6), 2730–2742.

Hur, S., Kim, T., Love, D., Krogmeier, J., Thomas, T. & Ghosh, A. (2013a). Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans. Commun.*, 61(10), 4391–4403.

Hur, S., Kim, T., Love, D. J., Krogmeier, J. V., Thomas, T. A. & Ghosh, A. (2013b). Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans. Commun.*, 61(10), 4391–4403.

I *et al.*, C.-L. (2014). Toward Green and Soft: A 5G Perspective. *IEEE Commun. Mag.*, 52(2), 66-73.

Ikhlef, A., Michalopoulos, D. S. & Schober, R. (2012). Max-max Relay Selection for Relays with Buffers. *IEEE Trans. Wireless Commun.*, 11(3), 1124–1135.

Iserte, A., Palomar, D., Pérez-Neira, A. & Lagunas, M. (2006). A Robust Maximin Approach for MIMO Communications With Imperfect Channel State Information Based on Convex Optimization. *IEEE Trans. Signal Process.*, 54(1), 346–360.

Islam, M. N., Sampath, A., Maharshi, A., Koymen, O. & Mandayam, N. B. (2014). Wireless Backhaul Node Placement for Small Cell Networks. *48th Annual Conf. Info. Sciences and Systems (CISS 2014)*, pp. 1–6.

Islam, S. R., Avazov, N., Dobre, O. A. & Kwak, K.-S. (2017). Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges. *IEEE Commun. Surveys Tuts.*, 19(2), 721–742.

Jeong, S., Simeone, O. & Kang, J. (2018). Mobile Edge Computing via a UAV-Mounted Cloudlet: Optimization of Bit Allocation and Path Planning. *IEEE Trans. Veh. Technol.*, 67(3), 2049-2063.

Jeong, Y., Quek, T. Q. S. & Shin, H. (2011). Beamforming Optimization for Multiuser Two-Tier Networks. *J. Commun. and Networks*, 13(4), 327–338.

Kaleva, J., Tölli, A. & Juntti, M. J. (2016). Decentralized Sum Rate Maximization with QoS Constraints for Interfering Broadcast Channel via Successive Convex Approximation. *IEEE Trans. Signal Process.*, 64(11), 2788–2802.

Kela, P., Turkka, J. & Costa, M. Borderless Mobility in 5G Outdoor Ultra-Dense Networks. *IEEE Access*, 3.

Kha, H. H., Tuan, H. D. & Nguyen, H. H. (2011). Fast Global Optimal Power Allocation in Wireless Networks by Local D.C. Programming. *IEEE Trans. Wireless Commun.*, 11(2), 510–515.

Kha, H. H., Tuan, H. D. & Nguyen, H. H. (2012). Fast global optimal power allocation in wireless networks by local D.C. programming. *IEEE Trans. Wireless Commun.*, 11(2), 510-515.

Krikidis, I., Charalambous, T. & Thompson, J. (2012). Buffer-Aided Relay Selection for Cooperative Diversity Systems without Delay Constraints. *IEEE Trans. Wireless Commun.*, 11(5), 1957–1967.

Larsson *et al.* Massive MIMO for next generation wireless systems.

Lasaulce, S., Hayel, Y., Azouzi, R. E. & Debbah, M. (2009). Introducing hierarchy in energy games. *IEEE Trans. Wireless Commun.*, 8(7), 3833–3843.

Lau, V. & Cui, Y. (2010). Delay-optimal power and subcarrier allocation for OFDMA systems via stochastic approximation. *IEEE Trans. Wireless Commun.*, 9(1), 227–233.

Lee, S., Narlikar, G., Pal, M., Wilfong, G. & Zhang, L. Admission Control for Multihop Wireless Backhaul Networks with QoS Support.

Lee *et al.*, D. (2012). Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges. *IEEE Commun. Mag.*, 50(2), 148–155.

Li, G. Y., Xu, Z., Xiong, C., Yang, C., Zhang, S., Chen, Y. & Xu, S. (2011). Energy-efficient wireless communications: tutorial, survey, and open issues. *IEEE Wireless Comm.*, 18(6), 28–35.

Li, Y., Fan, P., Liu, L. & Yi, Y. (2018). Distributed MIMO Precoding for In-Band Full-Duplex Wireless Backhaul in Heterogeneous Networks. *IEEE Trans. Veh. Technol.*, 67(3), 2064–2076.

Li, Y., Sheng, M., Sun, Y. & Shi, Y. (2016). Joint Optimization of BS Operation, User Association, Subcarrier Assignment, and Power Allocation for Energy-Efficient HetNets. *IEEE J. Sel. Areas Commun.*, 34(12), 3339–3353.

Liao, J., Wong, K.-K., Zhang, Y., Zheng, Z. & Yang, K. (2017). Coding, Multicast, and Cooperation for Cache-Enabled Heterogeneous Small Cell Networks. *IEEE Trans. Wireless Commun.*, 16(10), 6838–6853.

Lipp, T. & Boyd, S. (2016). Variations and extension of the convex–concave procedure. *Opt. and Engineering*, 17(2), 263—287.

Liu, A. & Lau, V. K. (2013). Mixed-Timescale Precoding and Cache Control in Cached MIMO Interference Network. *IEEE Trans. Signal Process.*, 61(24), 6320–6332.

Liu, A. & Lau, V. K. (2014). Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems. *IEEE Trans. Signal Process.*, 62(2), 390–402.

Liu, G., Yu, F. R., Ji, H., Leung, V. C. M. & Li, X. (2015). In-Band Full-Duplex Relaying: A Survey, Research Issues and Challenges. *IEEE Commun. Surveys Tuts.*, 17(2), 500-524.

Liu, Y.-F., Dai, Y. H. & Luo, Z. Q. (2011a). Coordinated Beamforming for MISO Interference Channel: Complexity Analysis and Efficient Algorithms. *IEEE Signal Process. Lett.*, 59(3), 1142-117.

Liu, Y.-F., Dai, Y.-H. & Luo, Z.-Q. (2011b). Coordinated Beamforming for MISO Interference Channel: Complexity Analysis and Efficient Algorithms. *IEEE Trans. Signal Process.*, 59(3), 1142–1157.

Lobo, M., Vandenberghe, L., Boyd, S. & Lebret, H. (1998). Applications of Second-Order Cone Programming. *Lin. Alg. and its Applications*, 284, 193–228.

Lofberg, J. (2004). YALMIP: a toolbox for modeling and optimization in MATLAB. *Proc. CACSD Conf.*

Lopez-Perez, D., Chu, X., Vasilakos, A. V. & Claussen, H. (2013). Power Minimization Based Resource Allocation for Interference Mitigation in OFDMA Femtocell Networks. *IEEE J. Sel. Areas Commun.*, 32(2), 333–344.

Lucero, S. (2016). *IoT platforms: enabling the Internet of Things*. IHS Technology.

Luo, S., Zhang, R. & Lim, T. J. (2015). Downlink and Uplink Energy Minimization Through User Association and Beamforming in C-RAN. *IEEE Trans. Wireless Commun.*, 14(1), 494–508.

Luo, Z.-Q. & Zhang, S. (2008). Dynamic Spectrum Management: Complexity and Duality. 2(1), 57–73.

Luong, P., Tran, L.-N., Despins, C. & Gagnon, F. (2016a). Joint Beamforming and Remote Radio Head Selection in Limited Fronthaul C-RAN. *Proc. IEEE Veh. Technol. Conf. (VTC'16 Fall)*, pp. 1-5.

Luong, P., Nguyen, H., Wang, J., Lee, S. & Lee, Y.-K. (2011). Energy Efficiency Based on Quality of Data for Cyber Physical Systems. *International Conference on Internet of Things and International Conference on Cyber, Physical and Social Computing (2011)*, pp. 232–241.

Luong, P., Nguyen, T. M. & Le, L. B. (2014a). Throughput Analysis and Design for Coexisting WLAN and ZigBee Network. *Proc. IEEE Veh. Technol. Conf. (VTC'14 Fall)*, pp. 1-5.

Luong, P., Nguyen, T. M., Le, L. B. & Đào, N.-D. (2014b). Admission control design for integrated WLAN and OFDMA-based cellular networks. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC'14)*, pp. 1–6.

Luong, P., Nguyen, T. M. & Le, L. B. (2016b). Throughput analysis for coexisting IEEE 802.15.4 and 802.11 networks under unsaturated traffic. *EURASIP Journal on Wireless Communications and Networking*, 2016(1), 127–141.

Luong, P., Nguyen, T. M., Le, L. B., Đào, N.-D. & Hossain, E. (2016c). Energy-efficient WiFi offloading and network management in heterogeneous wireless networks. *IEEE Access*, 4, 10210–10227.

Luong, P., Despins, C., Gagnon, F. & Tran, L.-N. (2017a, May). Designing Green C-RAN with Limited Fronthaul via Mixed-Integer Second Order Cone Programming. *Proc. IEEE Int. Conf. Communications (ICC'17)*, pp. 1–6.

Luong, P., Despins, C., Gagnon, F. & Tran, L.-N. (2017b). Designing Green C-RAN with Limited Fronthaul via Mixed-Integer Second Order Cone Programming. *Proc. IEEE Int. Conf. Communications (ICC'17)*, pp. 1-6.

Luong, P., Gagnon, F., Despins, C. & Tran, L.-N. (2017c). Optimal Joint Remote Radio Head Selection and Beamforming Design for Limited Fronthaul C-RAN. *IEEE Trans. Signal Process.*, 65(21), 5605 - 5620.

Luong, P., Despins, C., Gagnon, F. & Tran, L.-N. (2018a). A Novel Energy-Efficient Resource Allocation Approach in Limited Fronthaul Virtualized C-RANs. *Proc. IEEE Veh. Technol. Conf. (VTC'18 Spring)*, pp. 1-6.

Luong, P., Gagnon, F., Despins, C. & Tran, L.-N. (2018b). Joint Virtual Computing and Radio Resource Allocation in Limited Fronthaul Green C-RANs. *IEEE Trans. Wireless Commun.*, 17(4), 2602-2617.

Lyu, J., Zeng, Y. & Zhang, R. (2018). UAV-aided offloading for cellular hotspot. *IEEE Trans. Wireless Commun.*, 17(6), 3988–4001.

Marks, B. R. & Wright, G. P. A general inner approximation algorithm for nonconvex mathematical programs. *Operations Research*, 26(4).

Marshoud, H., Kapinas, V. M., Karagiannidis, G. K. & Muhaidat, S. (2016). Non-orthogonalmultiple access for visible light communications. *IEEE Photon. Technol. Lett.*, 28(1), 51–54.

Miao, G., Himayat, N., Li, G. Y. & Talwar, S. (2011). Distributed Interference-Aware Energy-Efficient Power Optimization. *IEEE Trans. Wireless Commun.*, 10(4), 1323–1333.

Mozaffari, M., Saad, W., Bennis, M. & Debbah, M. (2017). Optimal Transport Theory for Cell Association in UAV-Enabled Cellular Networks. *IEEE Commun. Lett.*, 21(9), 2053–2056.

Mozaffari *et al.* (2016). Unmanned Aerial Vehicle With Underlaid Device-to-Device Communications: Performance and Tradeoffs. *IEEE Trans. Wireless Commun.*, 15(6), 3949–3963.

Mozaffari *et al.*, M. (2017). Mobile Unmanned Aerial Vehicles (UAVs) for Energy-Efficient Internet of Things Communications. *IEEE Trans. Wireless Commun.*, 16(11), 7574–7589.

Nasir, A. A., Ngo, D. T., Tuan, H. D. & Durrani, S. (2015). Iterative Optimization for Max-Min SINR in Dense Small-Cell Multiuser MISO SWIPT System. *Proc. IEEE Global Conference on Signal and Information Processing(GlobalSIP'15)*, pp. 1392-1396.

Ng, D. W. K., Lo, E. S. & Schober, R. (2012). Energy-Efficient resource allocation in OFDMA system with large numbers of base station antennas. *IEEE Trans. Wireless Commun.*, 11(9), 3292-3304.

Ngo, D. T., Khakurel, S. & Le-Ngoc, T. (2014a). Joint Subchannel assignment and Power Allocation for OFDMA Femtocell Networks. *IEEE Trans. Wireless Commun.*, 13(1), 342-355.

Ngo, D. T., Khakurel, S. & Le-Ngoc, T. (2014b). Joint subchannel assignment and power allocation for OFDMA femtocell networks. *IEEE Trans. Wireless Commun.*, 13(1), 342-355.

Ngo, H. Q., Larsson, E. G. & Marzetta, T. L. (2013). Energy and spectral efficiency of very large multiuser MIMO systems. *IEEE Trans. Commun.*, 61(4), 1436–1449.

Nguyen, D. H. N. & Le-Ngoc, T. (2014). Sum-rate maximization in the multicell MIMO broadcast channel with interference coordination. *IEEE Trans. Signal Process.*, 62(6), 1501–1513.

Nguyen, K.-G., Tran, L.-N., Tervo, O., Vu, Q.-D. & Juntti, M. (2015a). Achieving Energy Efficiency Fairness in Multicell MISO Downlink. *IEEE Commun. Lett.*, 19(8), 1426–1429.

Nguyen, T. M., Ajib, W. & Assi, C. A Novel Cooperative NOMA For Designing Unmanned Aerial Vehicle (UAV)–Assisted Wireless Backhaul Networks. *IEEE J. Sel. Areas Commun.*

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2016a). Achieving Energy-Efficiency in Two-Tier Wireless Backhaul HetNets. *Proc. IEEE Int. Conf. Communications (ICC'16)*, pp. 1-6.

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2016b). Resource Allocation in Two-Tier Wireless Backhaul Heterogeneous Networks. *IEEE Trans. Wireless Commun.*, 15(10), 6690-6704.

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2017a). Centralized and Distributed Energy Efficiency Designs in Wireless Backhaul HetNets. *IEEE Trans. Wireless Commun.*, 16(7), 4711-4726.

Nguyen, T. M., Ajib, W. & Assi, C. (2018a). Designing Wireless Backhaul Heterogeneous Networks with Small Cell Buffering. *IEEE Trans. Commun.*

Nguyen, T. M., Ajib, W. & Assi, C. (2018b). A Novel Cooperative Non-Orthogonal Multiple Access (NOMA) in Wireless Backhaul Two-Tier HetNets. *IEEE Trans. Wireless Commun.*, 17(7), 4873–4887.

Nguyen, T. M. & Le, L. B. (2014a). Cognitive spectrum access in femtocell networks exploiting nearest interferer information. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC' 14)*, pp. 1–6.

Nguyen, T. M. & Le, L. B. (2014b). Opportunistic spectrum sharing in Poisson femtocell networks. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC' 14)*, pp. 1–6.

Nguyen, T. M. & Le, L. B. (2015). Joint pilot assignment and resource allocation in multicell massive MIMO network: Throughput and energy efficiency maximization. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC' 15)*, pp. 1–6.

Nguyen, T. M., Quek, T. Q. S. & Shin, H. (2012a). Opportunistic interference alignment in MIMO femtocell networks. *Proc. IEEE Int. Symp. on Inf. Theory (ISIT' 2012)*, pp. 1–6.

Nguyen, T. M., Shin, H. & Quek, T. Q. S. (2012b). Network throughput and energy efficiency in MIMO femtocells. *Proc.18th European Wireless (EW' 2012)*, pp. 1–6.

Nguyen, T. M., Jeong, Y., Quek, T. Q. S., Tay, W. P. & Shin, H. (2013). Interference alignment in a Poisson field of MIMO femtocells. *IEEE Trans. Wireless Commun.*, 12(6), 2633–2645.

Nguyen, T. M., Ha, V. N. & Le, L. B. (2015b). Resource allocation optimization in multi-user multi-cell massive MIMO networks considering pilot contamination. *IEEE Access*, 3, 1272-1287.

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2016c). Energy Efficiency with Adaptive Decoding Power and Wireless Backhaul Small Cell Selection. *Proc. IEEE Global Telecom. Conf. (GLOBECOM 2016)*.

Nguyen, T. M., Ajib, W. & Assi, C. (2017b). Online Algorithm for Wireless Backhaul HetNets with Advanced Small Cell Buffering. *Proc.26th Int. Conf. on Comp. Comm. and Netw. (ICCCN'17)*, pp. 1–8.

Nikolich, P., I, C.-L., Korhonen, J., Marks, R., Tye, B., Li, G., Ni, J. & Zhang, S. (2017). Standards for 5G and Beyond: Their Use Cases and Applications. *IEEE 5G Tech Focus*, 1(2).

Niu *et al.* (2017). Energy Efficient Scheduling for mmWave Backhauling of Small Cells in Heterogeneous Cellular Networks. *IEEE Trans. Veh. Technol.*, 66(3), 2674–2687.

Nomikos, N., Charalambous, T., Krikidis, I., Skoutas, D. N., Vouyioukas, D. & Johansson, M. (2016). A Survey on Buffer–Aided Relay Selection. *IEEE Commun. Surveys Tuts.*, 18(2), 1073–1097.

Palomar, D. P. & Chiang, M. (2006a). A Tutorial on Decomposition Methods for Network Utility Maximization. *IEEE J. Sel. Areas Commun.*, 24(8), 1439-1451.

Palomar, D. P. & Chiang, M. (2006b). A Tutorial on Decomposition Methods for Network Utility Maximization. *IEEE J. Sel. Areas Commun.*, 24(8), 1439–1451.

Papandriopoulos, J. & Evans, J. S. (2009). SCALE: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks. *IEEE Trans. Inf. Theory*, 55(8), 3711-3724.

Parikh, N. & Boyd, S. (2014a). Proximal Algorithms. *Foundations and Trends in optimization*, 1(3).

Parikh, N. & Boyd, S. (2014b). Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3).

Peng, M., Yu, Y., Xiang, H. & Poor, H. V. (2016). Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks. *IEEE Trans. Multimedia*, 18(5), 879–892.

Phan, H., Zheng, F.-C. & Fitch, M. (2015). Wireless Backhaul Networks With Precoding Complex Field Network Coding. *IEEE Commun. Lett.*, 19(3), 447–450.

Pi, Z., Choi, J. & Heath, R. (2016). Millimeter-wave gigabit broadband evolution toward 5G: fixed access and backhaul. *IEEE Commun. Mag.*, 54(4), 138–144.

Pitaval *et al.*, R.-A. (2015). Full-duplex self-backhauling for small-cell 5G networks. 22(5), 83–89.

Qiao, D., Gursoy, M. C. & Velipasalar, S. (2013). Effective Capacity of Two–Hop Wireless Communication Systems. *IEEE Trans. Inf. Theory*, 59(2), 873–885.

Ranaweera *et al.* (2013). Design and optimization of fiber optic small-cell backhaul based on an existing fiber-to-the-node residential access network. *IEEE Commun. Mag.*, 51(9), 62–69.

Rao, J. B. & Fapojuwo, A. O. (2014). A Survey of Energy Efficient Resource Management Techniques for Multicell Cellular Networks. *IEEE Commun. Surveys Tuts.*, 16(1), 154–180.

Rashid-Farrokhi, F., Liu, K. J. R. & Tassiulas, L. (1998a). Transmit Beamforming and Power Control for Cellular Wireless Systems. *IEEE J. Sel. Areas Commun.*, 16(8), 1437–1450.

Rashid-Farrokhi, F., Tassiulas, L. & Liu, K. J. R. (1998b). Joint Optimal Power Control and Beamforming in Wireless Networks Using Antenna Arrays. *IEEE Trans. Wireless Commun.*, 46(10), 1313–1324.

Rockafellar, R. & Wets, J. (1998). *Variational Analysis*. Springer.

Rubio, J. & Pascual-Iserte, A. (2014). Energy-Aware Broadcast Multiuser-MIMO Precoder Design with Imperfect Channel and Battery Knowledge. *IEEE Trans. Wireless Commun.*, 13(6), 3137-3152.

Sabharwal*et al.*, A. (2014). In-Band Full-Duplex Wireless: Challenges and Opportunities. *IEEE J. Sel. Areas Commun.*, 32(9), 1637–1652.

Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A. & Higuchi, K. (2013). Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access. *Proc. VTC'13 Spring*.

Samarakoon, S., Bennis, M., Saad, W. & Latva-aho, M. (2013). Backhaul-Aware Interference Management in the Uplink of Wireless Small Cell Networks. *IEEE Trans. Wireless Commun.*, 12(11), 5813–5825.

Sanguinetti, L., Moustakas, A. L. & Debbah, M. (2015). Interference Management in 5G Reverse TDD HetNets with Wireless Backhaul: A Large System Analysis. *IEEE J. Sel. Areas Commun.*, 33(6), 1187–1200.

Sanjabi, M., Razaviyayn, M. & Luo, Z.-Q. (2014). Optimal Joint Base Station Assignment and Beamforming for Heterogeneous Networks. *IEEE Trans. Signal Process.*, 62(8), 1950–1961.

Scutari, G., Facchinei, F., Lampariello, L. & Song, P. (2017). Parallel and distributed methods for nonconvex optimization – Part I: Theory. *IEEE Trans. Signal Process.*, 65(8), 1929–1944.

Shanmugam, K., Golrezaei, N., Dimakis, A. G., Molisch, A. F. & Caire, G. (2013). Femto-Caching: Wireless Content Delivery Through Distributed Caching Helpers. *IEEE Trans. Inf. Theory*, 59(12), 8402–8413.

Shi, Q., Razaviyayn, M., Luo, Z.-Q. & He, C. (2011). An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel. *IEEE Trans. Signal Process.*, 59(9), 4331-4340.

Shi, S., Yang, L. & Zhu, H. (2016). Outage Balancing in Downlink Nonorthogonal Multiple Access With Statistical Channel State Information. *IEEE Trans. Wireless Commun.*, 15(7), 4718–4731.

Shi *et al.* (2014). Joint Beamforming and Power Splitting for MISO Interference Channel With SWIPT: An SOCP Relaxation and Decentralized Algorithm. *IEEE Trans. Signal Process.*, 62(3), 6194–6208.

Siddique, U., Tabassum, H., Hossain, E. & Kim, D. I. (2015a). Wireless backhauling of 5G small cells: challenges and solution approaches. *IEEE Wireless Commun.*, 22(5), 22–31.

Siddique, U., Tabassum, H., Hossain, E. & Kim, D. I. (2015b). Wireless backhauling of 5G small cells: Challenges and solution approaches. 22(5), 22-31.

Siddique, U., Tabassum, H. & Hossain, E. (2017). Downlink spectrum allocation for in-band and out-band wireless backhauling of full-duplex small cells. *IEEE Trans. Commun.*, 65(8), 3538–3554.

Singh, S., Kulkarni, M. N., Ghosh, A. & Andrews, J. G.

Soh, Y. S., Quek, T. Q. S., Kountouris, M. & Shin, H. (2013). Energy efficient heterogeneous cellular networks. *IEEE J. Sel. Areas Commun.*, 31(5), 840–850.

Sohrab, H. H. (2014). *Basic Real Analysis*. New York, NY: Springer New York.

Sturm, J. F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimiz. Methods Softw.*, 11-12, 625-653.

Sun, Q., Han, S., I, C.-L. & Pan, Z. (2015a). On the ergodic capacity of MIMO NOMA systems. *IEEE Wireless Commun. Lett.*, 4(4), 405–408.

Sun, R., Hong, M. & Luo, Z.-Q. (2015b). Joint Downlink Base Station Association and Power Control for Max-Min Fairness: Computation and Complexity. *IEEE J. Sel. Areas Commun.*, 33(6), 1040–1054.

Sun, Y., Ng, D. W. K., Ding, Z. & Schober, R. (2017a). Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems. *IEEE Trans. Commun.*, 65(3), 1077–1091.

Sun, Y., Ng, D. W. K. & Schober, R. (2017b). Optimal resource allocation for multicarrier MISO-NOMA systems. *Proc. IEEE Int. Conf. Communications (ICC'17)*, pp. 1-7.

Sun, Y., Ng, D. W. K., Zhu, J. & Schober, R. (to appear). Robust and Secure Resource Allocation for Full-Duplex MISO Multicarrier NOMA Systems. *IEEE Trans. Commun.*

Tabassum, H., Hossain, E. & Hossain, J. (2017). Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes. *IEEE Trans. Commun.*, 65(8), 3555–3570.

Tan, C. W., Chiang, M. & Srikant, R. (2011). Maximizing Sum Rate and Minimizing MSE on Multiuser Downlink: Optimality, Fast Algorithms and Equivalence via Max-min SINR. *IEEE Trans. Signal Process.*, 59(12), 6127-6143.

Tao, M., Chen, E., Zhou, H. & Yu, W. (2016). Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN. *IEEE Trans. Wireless Commun.*, 15(9), 6118–6131.

Tervo, O., Tran, L.-N. & Juntti, M. (2015a). Decentralized Coordinated Beamforming for Weighted Sum Energy Efficiency Maximization in Multi-Cell MISO Downlink. *Proc. IEEE GlobalSIP 2015*.

Tervo, O., Tran, L.-N. & Juntti, M. (2015b). Optimal Energy-Efficient Transmit Beamforming for Multi-User MISO Downlink. *IEEE Trans. Signal Process.*, 63(20), 5574–5588.

Thomsen, H., de Carvalho, E. & Popovski, P. (2014). Using Wireless Network Coding to Replace a Wired With Wireless Backhaul. *IEEE Wireless Commun. Lett.*, 4(2), 141–144.

Tran, L.-N., Hanif, M.-F. & Juntti, M. (2014). A conic quadratic programming approach to physical layer multicasting for large-scale antenna arrays. *IEEE Signal Process. Lett.*, 21(1), 114-117.

Tran, L.-N., Hanif, M.-F., Tölli, A. & Juntti, M. (2012). Fast Converging Algorithm for Weighted Sum Rate Maximization in Multicell MISO Downlink. *IEEE Signal Process. Lett.*, 19(12), 872–875.

Tran, T. X., Hajisami, A. & Pompili, D. (2017). Cooperative Hierarchical Caching in 5G Cloud Radio Access Networks. *IEEE Network*, 31(4), 35–41.

Tse, D. & Viswanath, P. (2005). *Fundamental of Wireless Communication*. Cambridge, UK: Cambridge University Press.

Tuy, H., Al-Khayyal, F. & Thach, P. (2005). Monotonic optimization: Branch and cut methods. *Essays and Surveys in Global Optimization*, pp. 39 - 78.

Venkatraman, G., Tölli, A., Juntti, M. & Tran, L.-N. (2016a). Traffic Aware Resource Allocation Schemes for Multi-Cell MIMO-OFDM Systems. *IEEE Trans. Signal Process.*, 64(11), 2730–2745.

Venkatraman, G., Tölli, A., Juntti, M. & Tran, L.-N. (2016b). Traffic Aware Resource Allocation Schemes for Multi-Cell MIMO-OFDM Systems. *IEEE Trans. Signal Process.*, 64(11), 2730–2745.

Venkatraman, G., Tölli, A., Juntti, M. & Tran, L.-N. (2017). Multigroup Multicast Beamformer Design for MISO-OFDM With Antenna Selection. *IEEE Trans. Signal Process.*, 62(22), 5832–5847.

Venturino, L. & Buzzi, S. (2015). Energy-Aware and Rate-Aware Heuristic Beamforming in Downlink MIMO OFDMA Networks With Base-Station Coordination. *IEEE Trans. Veh. Technol.*, 64(7), 2897–2910.

Venturino, L., Zappone, A., Risi, C. & Buzzi, S. (2014). Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination. *IEEE Trans. Wireless Commun.*, PP(99), 1-14.

Viswanathan, H. & Mukherjee, S. (2006). Throughput-range tradeoff of wireless mesh backhaul networks. *IEEE J. Sel. Areas Commun.*, 24(3), 593–602.

Vu, Q.-D., Nguyen, K.-G. & Juntti, M. (2016a). Max-min Fairness for Multicast Multigroup Multicell Transmission under Backhaul Constraints. *Proc. IEEE Global Telecom. Conf. (GLOBECOM 2016)*.

Vu, Q.-D., Nguyen, K.-G. & Juntti, M. (2016b). Max-min Fairness for Multicast Multigroup Multicell Transmission under Backhaul Constraints. *GLOBECOM 2016*, pp. 1-6.

Wang, N., Hossain, E. & Bhargava, V. K. (2015a). Downlink cell association for large-scale MIMO HetNets employing small cell wireless backhaul. *Proc. CCECE '15*, pp. 1042–1047.

Wang, N., Hossain, E. & Bhargava, V. K. (2015b). Backhauling 5G small cells: A radio resource management perspective. 22(5), 41-49.

Wang, N., Hossain, E. & Bhargava, V. K. (2016). Joint Downlink Cell Association and Bandwidth Allocation for Wireless Backhauling in Two-Tier HetNets with Large-Scale Antenna Arrays. *IEEE Trans. Wireless Commun.*, 15(5), 3251–3268.

Wang, T. & Vandendorpe, L. (2010). Iterative Resource Allocation for Maximizing Weighted Sum Min-Rate in Downlink Cellular OFDMA Systems. *IEEE Trans. Signal Process.*, 59(1), 223-234.

Wang, T. & Vandendorpe, L. (2012). On the SCALE Algorithm for Multiuser Multicarrier Power Spectrum Management. *IEEE Trans. Signal Process.*, 60(9), 4992-4998.

Wang, X., Chen, M., Taleb, T., Ksentini, A. & Leung, V. C. Cache in the air: Exploiting content caching and delivery techniques for 5G systems.

Wei, Z., Yuan, J., Ng, D. W. K., Elkashlan, M. & Ding, Z. (2016). A Survey of Downlink Non-orthogonal Multiple Access for 5G Wireless Communication Networks. *ZTE Commun.*, 14(4), 17-25.

Wei, Z., Dai, L., Ng, D. W. K. & Yuan, J. (2017a). Performance analysis of a hybrid downlink-uplink cooperative NOMA scheme. *Proc. the 86th IEEE Veh. Technol. Conf. (VTC'17 Fall)*, pp. 1–7.

Wei, Z., Ng, D. W. K., Yuan, J. & Wang, H. M. (2017b). Optimal Resource Allocation for Power-Efficient MC-NOMA With Imperfect Channel State Information. *IEEE Trans. Commun.*, 65(9), 3944–3961.

Wong, V. W., Schober, R., Ng, D. W. K. & Wang, L. C. (2017). *Key Technologies for 5G Wireless Systems*. England, UK: Cambridge University Press.

Wu, Q., Zeng, Y. & Zhang, R. (2018). Joint Trajectory and Communication Design for Multi-UAV Enabled Wireless Networks. *IEEE Trans. Wireless Commun.*, 17(3), 2109–2121.

Xia, B., Fan, Y., Thompson, J. & Poor, H. V. (2008). Buffering in a Three-node Relay Network. *IEEE Trans. Wireless Commun.*, 7(11), 4492–4496.

Xia, B., Fan, Y., Thompson, J. & Poor, H. V. Buffering in a Three-Node Relay Network.

Xiang, L., Ng, D. W. K., Islam, T., Schober, R., Wong, V. W. S. & Wang, J. Cross-Layer Optimization of Fast Video Delivery in Cache- and Buffer-Enabled Relaying Networks.

Xiang, L., Ng, D. W. K., Islam, T., Schober, R., Wong, V. W. S. & Wang, J. (2017). Cross-Layer Optimization of Fast Video Delivery in Cache- and Buffer-Enabled Relaying Networks. *IEEE Trans. Veh. Technol.*, 66(12), 11366 - 11382.

Xiong, C., Lu, L. & Li, G. Y. (2014). Energy-Efficient Spectrum Access in Cognitive Radios. *IEEE J. Sel. Areas Commun.*, 32(3), 550–562.

Xu *et al.* (2014). Energy-Efficient CoMP Precoding in Heterogeneous Networks. *IEEE Trans. Signal Process.*, 62(4), 1005–1017.

Yadav, A., Nguyen, T. M. & Ajib, W. (2016a). Optimal Energy Management in Hybrid Energy Small Cell Access Points. *IEEE Trans. Commun.*, 64(12), 5334–5348.

Yadav, A., Nguyen, T. M. & Ajib, W. (2016b). Joint grid energy-throughput optimization for hybrid energy small cell access points. *Proc. IEEE 12th Int. Conf. on Wireless and Mobile Comp., Netw. and Comm. (WiMob) (2016)*, pp. 1-6.

Yang, H., Geraci, G. & Quek, T. Q. S. (2016a). Energy Efficient Design of MIMO Heterogeneous Networks with Wireless Backhaul. *IEEE Trans. Wireless Commun.*, 15(7), 4914–4927.

Yang, H. H., Geraci, G. & Quek, T. Q. S. (2016b). Energy-Efficient Design of MIMO Heterogeneous Networks with Wireless Backhaul. *IEEE Trans. Wireless Commun.*, 15(7), 4914–4927.

Yang *et al.*, C. (2013). How do we design CoMP to achieve its promised potential? 20(1), 67–74.

Zeng, Y. & Zhang, R. (2017). Energy-efficient UAV communication with trajectory optimization. *IEEE Trans. Wireless Commun.*, 16(6), 3747–3760.

Zeng, Y., Zhang, R. & Lim, T. J. (2016a). Wireless communications with unmanned aerial vehicles: Opportunities and challenges. *IEEE Commun. Mag.*, 54(5), 36–42.

Zeng, Y., Zhang, R. & Lim, T. J. (2016b). Throughput maximization for UAV-enabled mobile relaying systems. *IEEE Trans. Commun.*, 64(12), 4983–4996.

Zhang, H., Jiang, C., Mao, X. & Chen, H.-H. (2015). Interference-limited resource optimization in cognitive femtocells with fairness and imperfect spectrum sensing. *IEEE Trans. Veh. Technol.*, 65(3), 1761–1771.

Zhang, Q., Li, Q. & Qin, J. (2016). Robust Beamforming for Nonorthogonal Multiple-Access Systems in MISO Channels. *IEEE Trans. Veh. Technol.*, 65(12), 10231-10236.

Zhang *et al.*, H. (2014). Resource Allocation in Spectrum-Sharing OFDMA Femtocells With Heterogeneous Services. *IEEE Trans. Commun.*, 62(7), 2366–2377.

Zhang *et al.*, H. (2015). Resource Allocation for Cognitive Small Cell Networks: A Cooperative Bargaining Game Theoretic Approach. *IEEE Trans. Wireless Commun.*, 14(6), 3481–3493.

Zhao, J., Quek, T. Q. S. & Lei, Z. (2015). Heterogeneous Cellular Networks Using Wireless Backhaul: Fast Admission Control and Large System Analysis. *IEEE J. Sel. Areas Commun.*, 33(10), 2128–2143.

Zhou *et al.*, S. Green mobile access network with dynamic base station energy saving. *In Proc. ACM MobiCom*, 9(262).

Zlatanov, N. & Schober, R. (2013). Buffer-Aided Relaying With Adaptive Link Selection—Fixed and Mixed Rate Transmission. *IEEE Trans. Inf. Theory*, 59(5), 2816–2840.

Zlatanov, N., Ikhlef, A., Islam, T. & Schober, R. Buffer-aided cooperative communications: Opportunities and challenges.