Imperial College London

Department of Electrical and Electronic Engineering

Content Delivery over Multi-antenna Wireless Networks

Junlin Zhao

September 2019

Supervised by Dr. Deniz Gündüz

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Electrical and Electronic Engineering of Imperial College London
and the Diploma of Imperial College London

Abstract

The past few decades have witnessed unprecedented advances in information technology, which have significantly shaped the way we acquire and process information in our daily lives. Wireless communications has become the main means of access to data through mobile devices, resulting in a continuous exponential growth in wireless data traffic, mainly driven by the demand for high quality content. Various technologies have been proposed by researchers to tackle this growth in 5G and beyond, including the use of increasing number of antenna elements, integrated point-to-multipoint delivery and caching, which constitute the core of this thesis.

In particular, we study non-orthogonal content delivery in multiuser multiple-input-single-output (MISO) systems. First, a joint beamforming strategy for simultaneous delivery of broadcast and unicast services is investigated, based on layered division multiplexing (LDM) as a means of superposition coding. The system performance in terms of minimum required power under prescribed quality-of-service (QoS) requirements is examined in comparison with time division multiplexing (TDM). It is demonstrated through simulations that the non-orthogonal delivery strategy based on LDM significantly outperforms the orthogonal strategy based on TDM in terms of system throughput and reliability. To facilitate efficient implementation of the LDM-based beamforming design, we further propose a dual decomposition-based distributed approach.

Next, we study an efficient multicast beamforming design in cache-aided multiuser MISO systems, exploiting proactive content placement and coded delivery. It is observed that the complexity of this problem grows exponentially with the number of subfiles delivered to each user in each time slot, which itself grows exponentially with the number of users in the system. Therefore, we propose a

low-complexity alternative through time-sharing that limits the number of subfiles that can be received by a user in each time slot. Moreover, a joint design of content delivery and multicast beamforming is proposed to further enhance the system performance, under the constraint on maximum number of subfiles each user can decode in each time slot.

Finally, conclusions are drawn in Chapter 5, followed by an outlook for future works.

Declaration of Originality

I hereby certify that this thesis is the result of my own work under the guidance of my Ph.D. advisor, Dr. Deniz Gündüz. Any ideas or quotations from the work of other people are appropriately referenced.

Imperial College London
London, United Kingdom
18 September 2019

Junlin Zhao

Copyright Declaration

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Dr. Deniz Gündüz, for his continuous guidance and support for me during my PhD studies. It has been very enjoyable to work with him, and his vast knowledge in information theory and communications has always inspired me in our discussions. His passion and devotion to research have been a source of motivations and encouragements for me, and will continue helping me in moving forward. I would also like to thank China Scholarship Council for the financial support during my PhD studies.

I am grateful for the valuable opportunity to work with Professor Osvaldo Simeone, who has kindly provided a lot of precious suggestions and insightful comments on my work. I have also been very fortunate to spend 5 months at the Ohio State University, and I really appreciate the help from Professor Can Emre Koksal during my stay. I would also like to thank Dr. Bruno Clerckx and Professor Arumugam Nallanathan for kindly reviewing this thesis and the helpful comments during the viva.

Meanwhile, I would like to thank all the friends I have met at the Department of EEE, particularly those at the Information Processing and Communications Lab. They have been great company in office, and have offered tremendous help to me in both work and life. The campus life would not have been so great without them: Borzoo, Morteza, Giulio, Qianqian, Yonghua, Yuancheng, David, Samuel, Zaid, Sreejith, Mohammad, Joan, Emre, Nitish, Can, Ecenaz, Burak, Yuxuan, and Ganggang. Best wishes to all of my friends for successes in their future studies and careers.

I would like to give special thanks to my parents and my sister, who have been strongly supporting me in every aspect no matter where I am. This thesis would

not have been possible without their unconditional support and unchanged love.

London, United Kingdom
18 September 2019

Junlin

To my family

Contents

List of Figures

List of Tables

List of Acronyms

SISO	Single-input single-output
MISO	Multiple-input single-output
MIMO	Multiple-input multiple-output
NOMA	Non-orthogonal multiple access
OMA	Orthogonal multiple access
DPC	Dirty paper coding
SIC	Successive interference cancellation
BS	Base station
CSI	Channel state information
CSIT	Channel state information at the transmitter
SNR	Signal to noise ratio
SINR	Signal to noise plus interference ratio
QoS	Quality of service
WSM	Weighted sum-rate
MSE	Mean square error
WMMSE	Weighted minimum mean square error
LDM	Layered division multiplexing
TDM	Time division multiplexing
FDM	Frequency division multiplexing
SDP	Semidefinite programming
SDR	Semidefinite relaxation
SOCP	Second-order cone programming
SCA	Successive convex approximation
i.i.d.	Independent and identically distributed

Chapter 1

Introduction

1.1 Motivations

Wireless communications have been, and will continue to be, an essential component in our daily life. The past decade has seen tremendous breakthroughs as the wireless networks evolved to the third generation (3G) and the fourth generation (4G) of wireless networks, opening the door to high speed multimedia services on mobile devices. Various emerging applications, such as autonomous driving, remote robotic surgery, smart homes, virtual/augmented reality, immersive gaming, etc., are expected in the fifth generation (5G) of mobile networks, raising new challenges to both academia and industry [1]. Particularly, driven by the growing demand for the high data rate applications, the 5G networks are expected to offer a 1000-fold increase in network throughput [2].

Candidate approaches towards this goal include exploiting more spectrum resources, employing higher spectral efficiency and better interference management techniques, and densifying the networks. Moreover, analytics has revealed that a significant portion of the explosive data traffic growth is due to video content. It is anticipated that 79% of the mobile data traffic will be video by 2022 [3]. Therefore, highly efficient video content delivery is another key aspect to investigate to accommodate the 1000x data traffic increase in 5G, and is the focus of this thesis.

During the generation by generation evolutions of wireless networks, system throughput has always been one of the most important performance metrics in network design. As a promising technique to significantly increase the system throughput, multiple-input-multiple-output (MIMO) has been thoroughly investigated and standardized for highly efficient and reliable data transmission since 3G [4, 5]. In the most recent efforts to the development of 5G wireless networks, massive MIMO [6], as a variant of standard MIMO with extremely large number of antennas, is recognized as a key technique to provide massive connectivity and to further increase the network throughput, especially to enhance the transmission over sub-6GHz and millimeter wave frequency bands [7, 8].

Particularly, mobile broadband multimedia services (MBMS) has been introduced since 3G to support new point-to-multipoint radio bearers and multicast capability in the core network. The enhanced versions of MBMS, termed as evolved MBMS (eMBMS) and further evolved MBMS (FeMBMS), have been introduced in recent 3GPP releases. The current deployment of MBMS entails a reduction in system capacity for unicast services, since MBMS and unicast services are multiplexed in time in different sub-frames. Superposition coding, a form of non-orthogonal multiple access (NOMA), was proposed in [9] to improve unicast throughput and broadcast coverage with respect to traditional orthogonal frequency division multiplexing (FDM) or time division multiplexing (TDM), by simultaneously using the same frequency and time resources for multiple unicast or broadcast transmissions. In general, at the cost of an increased complexity at the receivers, which need to perform successive interference cancellation (SIC), NOMA can provide significant gains in spectral efficiency as compared to orthogonal multiple access (OMA) approaches, and is in fact optimal in achieving the capacity region of degraded Gaussian broadcast channels [10]. In addition to MBMS, various topics has been studied with NOMA, including cooperative relay networks [11], physical layer security [12], visible light communication [13], mmWave communication [14], etc., where performance gains are observed as compared to OMA.

Further exploration for spectral efficiency improvement is motivated by the fact that most of the data traffic occurs in daytime and leads to network congestion, whereas the resources are underutilized during off-peak periods. By proactively pushing contents to users during off-peak hours, caching in wireless networks can significantly improve the spectral efficiency, and reduce the bandwidth requirements and the transmission delay, particularly for video-on-demand (VoD) services [15]. While the application of caching has enjoyed great success for decades in computer networks, caching in wireless networks has just popularized recently, with studies recently carried out on downlink transmission with cache-aided base stations [16], D2D communications among cache-enabled users [17, 18], cooperative MIMO transmission with caching [19], caching in millimeter wave communications [20], and so on.

Following the pioneering work by Maddah-Ali and Niesen [21], coded caching has been known to outperform uncoded caching and achieve a global caching gain by creating and exploiting multicasting opportunities. With coded caching, uncoded contents can be proactively pushed at user devices without knowing users' demands, and a server can serve multiple users simultaneously by broadcasting specially designed XORs of files which are useful for these users to recover the desired contents. This feature is particularly favorable with wireless medium, which intrinsically has the broadcasting nature. Moreover, the global caching gain obtained via such broadcasting opportunities scales with the number of users in the network, which is particularly beneficial for large scale networks.

Motivated by the advantages of MIMO, NOMA, and coded content delivery in terms of spectral efficiency, we study efficient content delivery in multiple-input-single-output (MISO) networks, with transmit beamforming, successive decoding at users, and coded content delivery.

1.2 Outline and Contributions

First, in Chapter 2 we introduce the general field of transmit beamforming, SIC and NOMA techniques, and coded caching, which constitute the core of technical works presented in this thesis. Departing from the literature review, we propose novel beamforming strategies in the following technical chapters, considering various application scenarios in multiuser MISO networks that are of practical importance.

Chapter 3

We first overview the recent progress in standardization of terrestrial broadcasting systems, and technical advances in MBMS in cellular networks. Layered division multiplexing (LDM), as a form of superposition coding, has received considerable attention in the development of next generation terrestrial broadcasting systems. Having observed the advantages of applying LDM for incorporating TV broadcasting services in cellular networks, we study joint transmission of broadcast content and user-specific unicast contents in the downlink of multicell multiuser MISO networks, via transmit beamforming and LDM. Specifically, beamforming and power allocation between unicast and broadcast layers, termed as injection level in the LDM literature, are obtained with the aim of minimizing the total transmit power under constraints on the user-specific unicast rates and on the common broadcast rate. We also study the effects of imperfect channel state information (CSI) and imperfect channel coding to gain insights into robust implementation in practical systems. Performance of the beamforming strategy with LDM is compared to the strategy with TDM, and significant performance gains of LDM over TDM are demonstrated via numerical simulations. We further develop an efficient distributed implementation of the LDM-based beamforming strategy based on dual decomposition. The results presented in this chapter have been published in:

- **J. Zhao**, D. Gündüz, O. Simeone, and D. Gomez-Barquero, Non-orthogonal unicast and broadcast transmission via joint beamforming and LDM in cellular networks, to appear in IEEE Transactions on Broadcasting.

- **J. Zhao**, O. Simeone, D. Gündüz, and D. Gomez-Barquero, Non-orthogonal unicast and broadcast transmission via joint beamforming and LDM in cellular networks, 2016 IEEE Global Communications Conference (GLOBECOM), Washington, D.C., USA, 2016, pp. 1-6.

Chapter 4

We study efficient multicast beamforming designs in cache-aided multiuser MISO systems, exploiting proactive content placement and coded delivery. It is observed that the complexity of the optimal beamforming design problem grows exponentially with the number of subfiles delivered to each user in each time slot, which itself grows exponentially with the number of users in the system. Therefore, we propose a low-complexity alternative through time-sharing that limits the number of subfiles that can be received by a user in each time slot. Moreover, a joint design of content delivery and multicast beamforming is proposed to further enhance the system performance, with constraints on the maximum number of subfiles each user can decode in each time slot. It is shown via extensive numerical simulations that our proposed low-complexity schemes significantly outperform the state-of-the-art design in the literature. The content of this chapter is based on the following works that have been published or submitted:

- **J. Zhao**, M. Mohammadi Amiri, and D. Gündüz, Multi-antenna Content Delivery with Coded Caching: From the Complexity Perspective, submitted for possible journal publication.

- **J. Zhao**, M. Mohammadi Amiri, and D. Gündüz, A low-complexity cache-aided multi-antenna content delivery scheme, IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Cannes, France, 2019, pp. 1-5.

Chapter 5

Finally, in Chapter 5 we provide conclusions of the research presented in this dissertation, and discuss possible extensions that can be considered for future work.

The following paper has been published as a result of a work carried out during the PhD studies, but is not included in this dissertation:

- Y. Sun, **J. Zhao**, S. Zhou, and D. Gündüz, Heterogeneous Coded Computation across Heterogeneous Workers, 2019 IEEE Global Communications Conference (GlOBECOM), Waikoloa, HI, USA, Dec. 2019.

Chapter 2

Overview

2.1 Background

In this chapter, we first introduce multiuser MIMO systems, and briefly present the information-theoretic results on the capacity of multiuser broadcast channels. As is known in information theory, the capacity of Gaussian MIMO broadcast channels is achieved by dirty paper coding (DPC), which however is hard to implement due to the requirement of noncausal knowledge of interference to all the users. Therefore, we focus on linear beamforming schemes that can be more easily implemented. An overview is provided on relevant topics in the literature of multiuser MIMO, including beamforming for MISO broadcast channels, multi-cell coordinated beamforming, multigroup multicast beamforming, robust beamforming with imperfect CSIT, and physical layer techniques with coded caching.

2.1.1 Multiuser MIMO Systems

Fig. 2.1 depicts a single-cell multiuser MIMO network, which consists of an N_T-antenna base station (BS), and K users each with N_R antennas. Compared to the single-input-single-output (SISO) deployment, MIMO can provide both multiplexing and diversity gains by leveraging multiple antennas at the transmitter and the receivers. In addition to time and frequency resources, degrees of freedom (DoF) in the spatial domain can be also exploited to simultaneously transmit multiple data streams over the same time and frequency resource block.

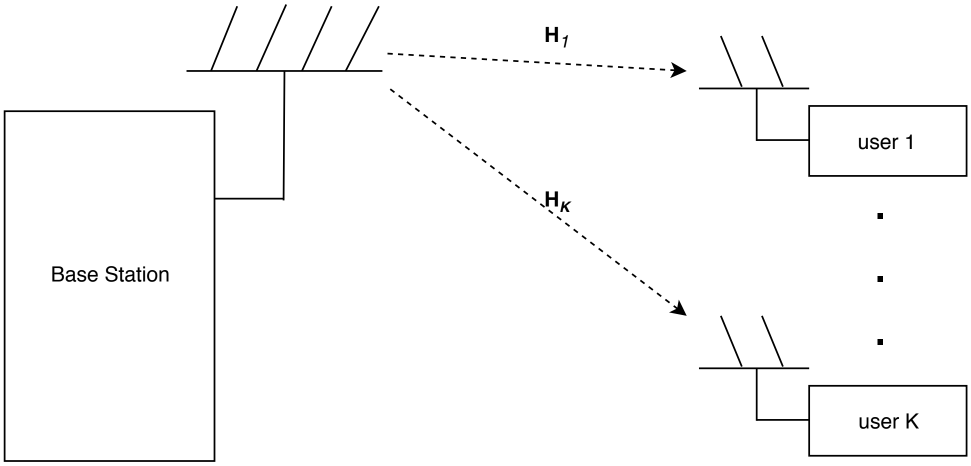


FIGURE 2.1: Illustration of a multiuser MIMO network with a base station equipped with $N_T = 4$ antennas, and $K = 4$ users each equipped with $N_R = 2$ antennas.

In the downlink of multiuser MIMO networks, the received signal at user k can be written as

$$\mathbf{y_k} = \mathbf{H}_k\mathbf{x} + \mathbf{n}_k, \tag{2.1}$$

where $\mathbf{H}_k$ is the channel matrix from the BS to user k, and $\mathbf{n}_k \sim \mathcal{CN}(0, \mathbf{I})$ denotes the circularly symmetric Gaussian noise. The covariance matrix of the input signal $\mathbf{x}$ is given by $\mathbf{\Sigma}_x \triangleq \mathrm{E}[\mathbf{x}\mathbf{x}^H]$, and satisfies $\mathrm{Tr}(\mathbf{\Sigma}_x) \leq P$, which indicates an average power constraint at the BS.

The capacity region of Gaussian MIMO broadcast channels, was unknown until the pioneering work by Caire and Shamai [22]. It has been shown that the sum rate capacity of the two-user case can be achieved by dirty paper coding (DPC) [23]. The result was extended to the general case with any number of users in subsequent works [24–26], also with the use of DPC. Finally, it has been shown in [27] that DPC can achieve the entire capacity region of Gaussian broadcast channels. Specifically, as depicted in Fig. 2.2, the transmitter first selects a codeword $\boldsymbol{x}_1$ for user 1, then chooses a codeword $\boldsymbol{x}_2$ for user 2 with the full knowledge $\boldsymbol{x}_1$. As such, the codeword $\boldsymbol{x}_1$ intended for user 1, which is

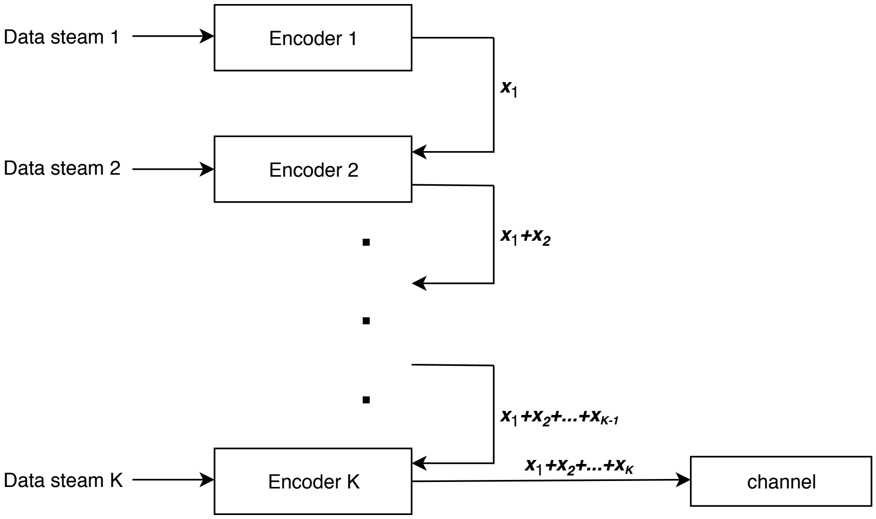


FIGURE 2.2: Diagram of dirty paper coding for K users.

interference to user 2, can be ”presubtracted” at the transmitter as if user 2 does not see $\boldsymbol{x}_1$. This process continues for all the K users. Suppose $\pi(i)$ denotes the user whose message is encoded in the i-th place, the achievable rate of user $\pi(i)$ is [28]:

$$R_{\pi(i)} = \frac{1}{2} \log \frac{\left| \mathbf{I} + \mathbf{H}_{\pi(i)} \left(\sum_{j \geq i} \boldsymbol{\Sigma}_{\pi(j)} \right) \mathbf{H}_{\pi(i)}^{H} \right|}{\left| \mathbf{I} + \mathbf{H}_{\pi(i)} \left(\sum_{j > i} \boldsymbol{\Sigma}_{\pi(j)} \right) \mathbf{H}_{\pi(i)}^{H} \right|}, \quad i = 1, \ldots, K, \tag{2.2}$$

where $\boldsymbol{\Sigma}_{\pi(i)}$ denotes the positive semidefinite transmit covariance matrix associated with user $\pi(i)$, and satisfies the average power constraint given by $\sum\limits_{k=1}^{K} \mathrm{Tr}(\boldsymbol{\Sigma}_{\pi(k)}) \leq P$.

While DPC is a powerful capacity-achieving technique, it requires noncausal knowledge of the interference for each user, and is thus hard to implement in practice. The prohibitive complexity in DPC implementations has motivated the research on low-complexity beamforming strategies in multiuser MIMO networks.

2.1.2 Transmit Beamforming

Throughout this thesis, we focus on the downlink transmission, and assume $N_R = 1$ for all the users for simplicity. This assumption can be justified by the fact that users are usually mobile devices with very stringent limitations on cost and power consumption. In general, the received signal at the user k can be written as

$$y_k = \boldsymbol{h}_k^H \boldsymbol{x} + n_k, \tag{2.3}$$

where $\boldsymbol{h}_k$ denotes the channel vector from the BS to user k, $\boldsymbol{x}$ is the transmitted signal at the BS, and $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the circularly symmetric complex Gaussian noise at user k. Specifically, with *linear precoding* at the transmitter, the transmitted signal at the BS can be represented by a linear superposition of user-specific messages, given by

$$\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{w}_k s_k, \tag{2.4}$$

where s_k is the message indented for user k, and $\boldsymbol{w}_k$ is the associated beamforming vector.

It is known that with perfect channel state information (CSI), a maximum of $\min(N_t, K)$ DoF can be achieved in MISO broadcast channels [29]. Moreover, linear precoding techniques, such as zero-forcing (ZF), can achieve this optimal DoF [30]. With ZF, the co-channel inter-user interference can be cancelled by exploiting the spatial degrees of freedom. Specifically, taking

$$\boldsymbol{h}_j^H \boldsymbol{w}_k = 0, \text{ for any } j \neq k \tag{2.5}$$

as the constraints to ensure cancellation of the inter-user interference, the ZF precoder can be obtained by solving a power minimization problem, given as

$$\min \sum_{k=1}^{K} \|\boldsymbol{w}_k\|^2 \tag{2.6}$$

$$\text{s.t. } \boldsymbol{h}_j^H \boldsymbol{w}_k = 0, \text{ for any } j \neq k. \tag{2.7}$$

This problem is feasible with $N_t \geq K$. It can be seen that ZF turns the MISO broadcast channel into K parallel scalar channels, and can be considered as an orthogonal transmission strategy in the spatial domain. However, the ZF design can be far from optimal in terms of spectral efficiency and user fairness. Therefore, max-min fair transmit beamforming has been introduced to maximize the minimum received signal-to-interference-plus-noise ratio (SINR) of the users under a total power budget. Other approaches include weighted sum rate maximization under a given power budget [31], and power minimization under prescribed quality-of-service requirements [32]. In these approaches, the inter-user interference is treated as noise when a user decodes the interested signal, with the SINR at user k given by

$$\text{SINR}_k \triangleq \frac{\left|\boldsymbol{h}_k^H \boldsymbol{w}_k\right|^2}{\sum\limits_{j\neq k} |\boldsymbol{h}_k^H \boldsymbol{w}_j| + \sigma_k^2}, \tag{2.8}$$

and the achievable rate for user k is given by

$$R_k = \log_2(1 + \text{SINR}_k). \tag{2.9}$$

Details of the design of linear transmit beamformers will be elaborated in Chapter 3 and Chapter 4.

2.1.3 SIC and NOMA

In the literature of transmit beamforming design, single-user detection is widely adopted, where each user decodes the desired message by treating the interference from other users' messages as noise [33], as indicated in (2.8). This interference management strategy generally yields satisfactory performance, but is dramatically suboptimal in theory compared to multiuser detection mechanism [34]. However, the co-channel interference, unless zero-forced, is usually non-negligible in the context of transmit beamforming, and becomes a key limiting factor on the system performance.

Initially proposed in [35] for scalar broadcast channels, successive interference cancellation (SIC) is a well-known multiuser detection technique that partially or fully decodes the interference before decoding the signal of interest, and thus yields higher achievable rates than treating interference as noise. Particularly, SIC is a key technique to achieve the capacity region of various channels, such as degraded Gaussian broadcast channels [36], Gaussian multiple access channels [37], and Gaussian interference channels with strong interference [38, 39]. SIC is also used in the Han-Kobayashi scheme [40], which yields the best-known achievable rate region of general discrete memoryless interference channels. It is noted that implementing SIC at the receivers is equivalent to DPC in terms of capacity in scalar broadcast channels. However, for the non-degraded MIMO broadcast channels discussed above, it is found that the achievable rate region obtained by SIC is contained within the DPC region [41].

Note that the performance gain comes at the cost of higher complexity of SIC for multiuser detection at the receivers. In addition to the high hardware complexity, SIC also suffers from error propagation, analog-to-digital quantization error, imperfect channel estimates [42], which has hindered the practical implementations of SIC. However, the thriving demands for network throughput has renewed the interest in SIC in recent years, which is the key component in

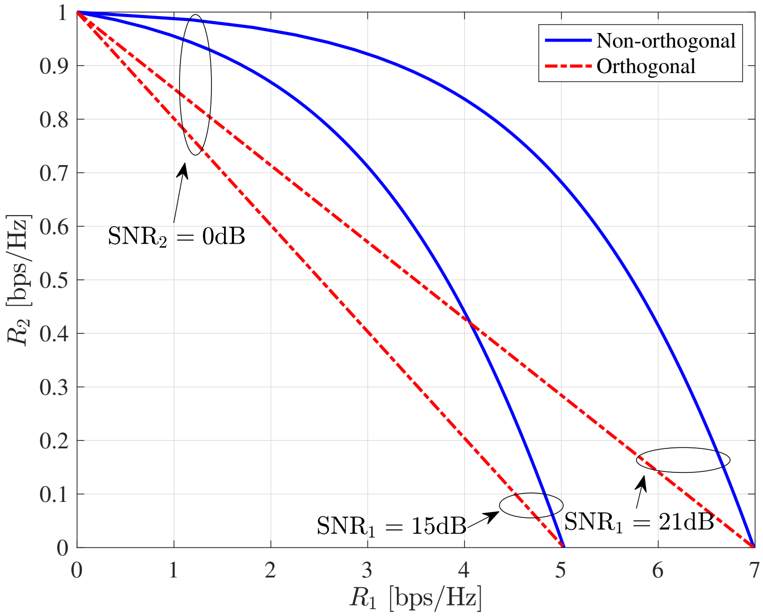


FIGURE 2.3: Achievable rate regions of NOMA and OMA for two-user scalar Gaussian broadcast channels.

state-of-the-art NOMA schemes.

Fig. 2.3 depicts the achievable rate regions of NOMA and OMA of scalar Gaussian broadcast channels with two users. The OMA regions are obtained using simple orthogonal schemes, such as time division and frequency division; the NOMA regions, which are in fact the exact capacity regions, are achieved with superposition coding and SIC. While this example considers the simple two-user setting, the sub-optimality of OMA compared to NOMA holds for the general case of arbitrary number of users.

2.1.4 Coded Caching

The application of caching to alleviate network congestion dates back to the 90s, when the Internet traffic experienced an explosive growth as World Wide Web (WWW) services intensified. By proactively pushing popular contents closer

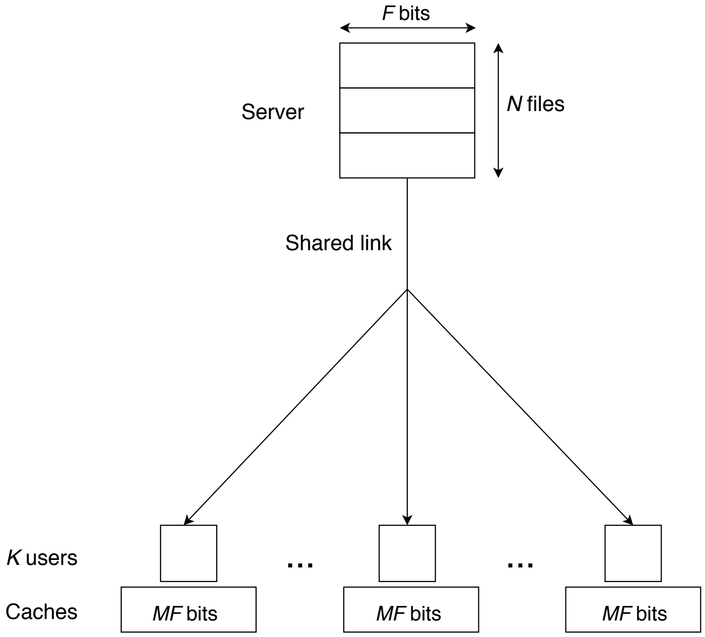


FIGURE 2.4: A single-server network with a library of N files, and K users each with a local memory which can store up to M files.

to users, e.g., at the gateways, caching can significantly increase the network throughput and reduce the latency in content delivery. Maddah-Ali and Niesen have recently proposed a novel proactive caching and content delivery scheme [21], which has greatly motivated the research on caching in wireless networks. In conventional uncoded caching schemes, a *local caching gain* is obtained when the requested content can be retrieved at local memories. In [21], by jointly designing the content placement and delivery, a *global caching gain* can be obtained such that multiple user demands can be satisfied by a single multicast transmission.

In particular, consider a single-server network as illustrated in Fig. 2.4. The server has a library of N files denoted by $\boldsymbol{V} \triangleq (V_1, \cdots, V_N)$, each uniformly distributed over the set $\{1, \cdots, 2^F\}$. K users are connected to the server via a shared link, and each user is equipped with a local cache memory of size MF bits. In the placement phase of conventional uncoded caching, the same M/N fraction of each file can be cached at all the users, which satisfies the cache capacity constraints. When a user requests a file in the library, the server sends

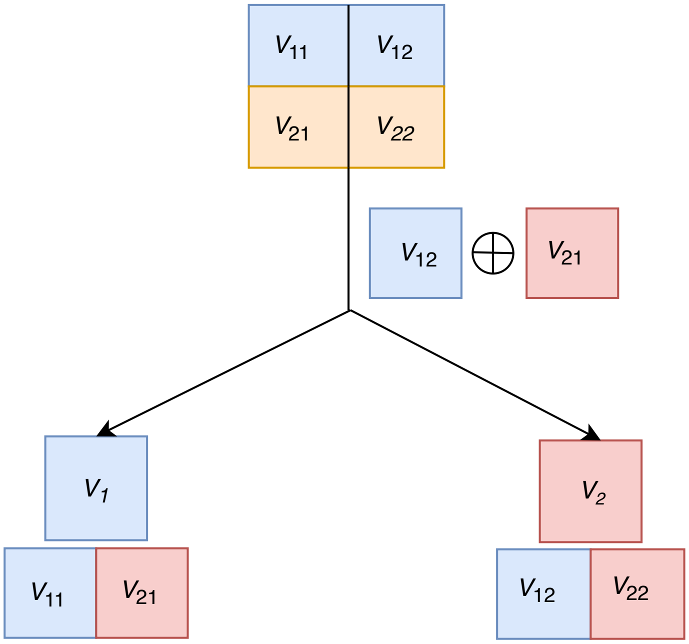


FIGURE 2.5: The coded caching scheme for the system with $N = 2$ files, $K = 2$ users, and $M = 1$, where user 1 requests file V_1, and user 2 requests file V_2.

the remaining $1 - M/N$ fraction of the requested file through the shared link. Note that the worst case of this setting is when all the users request distinct files. Therefore, the total number of bits to be delivered in the worst case is given by $KF(1 - \frac{M}{N})$, where $1 - \frac{M}{N}$ is referred to as the *local caching gain*. By jointly designing the content placement and delivery phases, the coded caching scheme in [21] further achieves a *global caching gain*. Next we present an example to describe the coded caching scheme in [21].

Example. Consider the case with $N = 2$, $K = 2$, and $M = 1$, as illustrated in Fig. 2.5. The files are represented by V_1 and V_2, and the demands of user 1 and user 2 are denoted by d_1 and d_2, respectively. In the coded caching scheme, each file is equally divided into 2 subfiles, i.e., V_1 is split into V_{11} and V_{12}, and V_2 is split into V_{21} and V_{22}. In the placement phase, uncoded subfiles are cached at users' local memories at negligible cost. Specifically, user 1 caches V_{11} and V_{21}, and user 2 caches V_{12} and V_{22}, which satisfies the cache capacity constraints at both users. Note that the placement is performed without any prior knowledge of user requests. When user requests are revealed, the server can transmit a single multicast message to simultaneously satisfy the requests, which is generated by bit-wise XOR between subfiles. Specifically, if $d_1 = 1$ and $d_2 = 2$, the message $V_{12} \oplus V_{21}$ can be multicast through the shared link, where $\frac{1}{2}F$ bits are transmitted. On receiving $V_{12} \oplus V_{21}$, both user 1 and user 2 can recover the remaining fraction of the requested file. While this example is for the case of $d_1 = 1$ and $d_1 = 2$, it is shown in [21] that a single coded message can always be designed for any combination of user requests. In general, the coded caching scheme in [21] achieves a delivery rate of $KF(1 - \frac{M}{N})\frac{1}{1+KM/N}$, which has a global caching gain of $\frac{1}{1+KM/N}$ as compared to the uncoded scheme. It is noted that the global caching gain scales with the aggregate cache size over all the users in the system. Particularly, the global caching gain is higher as the number of users increases, which is favorable in large scale networks.

2.2 Literature Review

In this section, relevant works in the literature of multiuser MIMO are overviewed, which can be categorized into the following five areas: *1*) beamforming for MISO broadcast channels, *2*) multi-cell coordinated beamforming, *3*) multigroup multicast beamforming, *4*) robust beamforming with imperfect CSIT, and *5*) physical layer techniques with coded caching.

2.2.1 Beamforming for MISO Broadcast Channels

Although DPC is a powerful capacity-achieving scheme in MIMO broadcast channels, its practical implementation is considered to be highly complicated. Beamforming, also commonly referred to as linear precoding, is a low-complexity alternative which generally yields satisfactory performance. The beamformer design involves determining both beamforming vectors and power weights for the users, and highly depends on the objective that is of interest to the system operator. While ZF beamforming is DoF optimal with perfect CSI, its performance is significantly degraded with limited power [43]. Typical design criteria include minimizing the required transmit power while satisfying a set of quality of service (QoS) constraints of each user, maximizing the minimum rate among users subject to a power constraint, and maximizing the total throughput under a constraint on the total transmit power.

It has been shown in [44, 45] that the downlink power minimization problem under SINR constraints can be efficiently solved based on the uplink-downlink duality. Particularly, [45] proved the existence of the optimal solution, and provided an iterative algorithm that converges to the optimal solution. The well-known uplink-downlink duality established that, under a sum-power constraint, the SINR region achievable by downlink beamforming in a MISO broadcast channel is shown to be identical to that of a dual uplink channel, which is computationally easier to handle. Based on the uplink-downlink duality, the problem of maximizing the minimum SINR of users, also known as max-min fairness problem, was globally solved in [32], also in an iterative manner. [46, 47] introduced semidefinite programming (SDP) to solve the downlink power minimization problem, and showed that the global optimal solution to the original problem may be obtained by solving the rank-relaxed problem. [48] showed that the power minimization problem and the max-min fairness problem can be turned into second-order cone program (SOCP) and generalized eigenvalue problem, respectively, which can then be tackled efficiently and globally by simple fixed-point iteration algorithms, without the need to resort to the virtual uplink problem. Moreover, it has been shown in [48] that the power minimization problem and the max-min fairness problem are inverse problems. It is noted that the works above all considered a sum-power constraint at the BS. For the sum-power minimization problem under additional per-antenna power constraints, [49] extended the aforementioned uplink-downlink duality, and provided an interior-point algorithm that converges to the optimal solution.

Different from power optimization and max-min fairness problems, which can be solved to optimality, sum-rate and weighted sum-rate (WSR) maximization problems are generally NP-hard, and local optimal solutions were often found. In [50] and [51], sum-rate and WSR optimization problems were studied, where local optimal solutions were obtained, respectively. Especially, [51] leveraged the uplink-downlink duality with respect to mean square error (MSE) to solve the problem. As a new approach to dealing with WSR problems, the equivalence between the WSR problem and the weighted minimum mean square error (WMMSE) problem was established in [31].

Building upon the results above, downlink beamforming in different settings and emerging application scenarios has been investigated, such as cognitive radio [52], secure communication [53], simultaneous information and power transfer [54], etc.

2.2.2 Multi-cell Coordinated Beamforming

Interference has always been a severe performance limiting factor in wireless systems, and becomes more problematic in multi-cell networks where inter-cell interference severely degrades the system performance. However, by allowing a certain level of information exchange between different cells, inter-cell interference can be properly managed in the network. Particularly, with full BS cooperation, where the BSs are linked by high-capacity delay-free links, and share CSI of both direct and interfering channels and full data signals of the associated users, the multi-cell downlink channel is essentially a MIMO broadcast channel, with distributed transmit antennas and individual power constraints. As shown in [55], cooperative downlink transmission among BSs can significantly enhance the system performance, at the price of significantly increased overhead.

Interference coordination, where only CSI but not data sharing is allowed within the network, has emerged as a more affordable technique in multi-cell systems. In this case, the multi-cell downlink channel can be essentially modeled as an interference channel. Coordinated beamforming was firstly studied in [44], where a suboptimal solution was obtained with an iterative algorithm developed based on uplink-downlink duality. Although optimal beamforming design problems have the nonconvex nature similar to multi-cell power control problems, in [56], the problem of minimizing total transmit power under individual SINR constraints for single-antenna users was efficiently solved to global optimality. Assuming that each cell only serves one active user, another multi-cell beamforming implementation was proposed in [57] by characterizing the capacity region of MIMO interference channels. Different from the works without cooperation between BSs, [58–62] considered clustered BS cooperation, where each user can be served by a set of cooperative, if not all, BSs, thereby striking a balance between full cooperation and interference coordination. Particularly, [60] studied the network utility maximization problem subject to power constraints, and proved its NP-hardness.

2.2.3 Multigroup Multicast Beamforming

Multicast beamforming, where the multi-antenna BS multicasts distinct data streams to multiple users or user groups, is an efficient physical layer technique for content delivery. In [63], the authors first investigated the single-group multicasting problem, and have proved its NP-hardness. [64] studied a cognitive radio system, where single-group multicasting for secondary users co-exists with the primary point-to-point transmission pair. Multicast beamforming for multiple user groups has been considered in [65], and two beamforming strategies are designed, namely minimizing total transmission power while guaranteeing a prescribed SINR at each receiver, and maximizing the overall minimum SINR under a total power budget. Instead of imposing a sum-power constraint at the BS, [66] considered multicast beamforming with more practical per-antenna power constraints. Multigroup multicasting with per-antenna power has been considered in [67, 68] for the massive MIMO setting. In case of so-called overloaded systems, [69] adopted a rate-splitting approach to handle the inter-group interference, and achieved significant performance gains over classical schemes. Symbol-level precoding has attracted much attention recently, and multicast beamforming on a symbol-per-symbol basis has been discussed in [70].

While there are a number of works concentrated on the single-cell case as mentioned above, multicast beamforming in multi-cell systems has also been extensively studied. [71, 72] considered the multi-cell multicasting problem aiming at maximizing the minimum SINR but subject to the total power among all the BSs, which is thus mathematically similar to the single-cell multigroup system studied in [65]. In [73], the same problem is revisited with power constraints on individual BSs. However, data sharing is allowed among BSs, which requires a large amount of information exchange overhead. To alleviate the significant overhead due to data sharing, [74] studied a simplified multi-cell multicast problem where there exists only one user group in each cell, and the same problem without BS cooperation in massive MIMO systems was studied in [75] by the same authors. The more general multi-cell multigroup multicast beamforming design was investigated in [76] for both power minimization and max-min fairness problems. Multi-cell multicast beamforming has also been investigated cloud radio access networks [77, 78].

2.2.4 Robust Beamforming with Imperfect CSIT

In practice, channel state information at the transmitter (CSIT) suffers from inaccuracies introduced by channel estimation errors in time division duplex (TDD) systems, or quantization errors in frequency division duplex (FDD) systems. In addition, CSIT can be outdated if the user mobility speed is faster than the CSIT update speed. Since the performance of beamforming relies on the use of CSIT and was found sensitive to such CSIT errors, numerous studies have been carried out for robust beamforming against imperfect CSIT, which generally aim at guaranteeing a certain level of performance under CSIT imperfections. Overall, the resultant robust precoder depends on the design criterion and the model of CSIT imperfections.

In case where CSIT imperfection is due to channel estimation errors, the CSIT errors are usually modeled as random variables with Gaussian distribution, and the average or outage performance of the system is therefore considered, as in [79–82]. In a different approach, CSIT errors due to quantization errors are assumed to be bounded, and a common approach is worst-case optimization that satisfies certain QoS requirements. The authors of [83] studied this problem by minimizing the transmit power subject to the worst-case SINR requirements with bounded CSIT errors, and provided solutions based on conservative approximations. For the same problem, [84] obtained a suboptimal solution using S-procedure and SDP. An improved result with reduced computational complexity was presented in [85], also using S-procedure. It is remarked that the use of S-procedure [86] is prevalent in the literature, e.g., see [77, 87–90], to combat bounded CSI errors, which turns the problem with infinitely many constraints into a tractable formulation with finite number of linear matrix inequalities. The problem of minimizing worst-case MSE of all users was considered and shown to be NP-hard in [82, 91], and the authors offered similar iterative algorithms convergent to suboptimal solutions.

Robust beamforming in a multi-cell network was studied with convex approximations in [92,93], where distributed solutions with limited information exchange between cells were proposed. [94] obtained the robust multi-cell beamforming design with global optimality with a branch-and-bound algorithm, which can serve as a benchmark to evaluate suboptimal algorithms that have reduced computational complexity. Recently, robust designs based on rate-splitting were presented in [95, 96], which have been shown to outperform conventional robust designs thanks to its advanced interference management capabilities.

2.2.5 Physical Layer Techniques with Coded Caching

Following the seminal work of Maddah-Ali and Niesen [21], many studies have been carried out for coded delivery over noisy broadcast channels in recent years. When users may have different channel capacities, the user with the worst channel condition becomes the bottleneck limiting the performance of multicasting. The global caching gain promised in [21] is hence not straightforward in practice. Coded caching in erasure broadcast channels is studied in [97] and [98] by allocating cache memories at weak receivers to overcome this bottleneck. A simple binary Gaussian broadcast channel is considered in [99], and an interference enhancement scheme is used to overcome the limitation of weak users. A cache-aided multicasting strategy over a Gaussian broadcast channel is presented in [100], with superposition coding and power allocation. The authors in [101] consider fading channels, and show that a linear increase in the sum delivery rate with the number of users can be achieved with user selection.

Another important line of research has focused on evaluating the performance of coded caching and delivery in the presence of multiple transmit antennas at the server. Multicast beamforming, where the multiple-antenna BS multicasts distinct data streams to multiple user groups, is an efficient physical layer technique [63,65,74]. In [102], the authors extend the results in [101] to MISO fading channels, where the same linear increase in content delivery rate with respect to the number of users is achieved without CSIT, and an improvement is obtained with spatial multiplexing when CSIT is available. In [103], coded delivery is employed along with zero-forcing to simultaneously exploit spatial multiplexing and caching gains. With multiple antennas at the BS, coded messages can be nulled at unintended user groups, which increases the number of users simultaneously served as compared to the single antenna setting. Particularly, this approach was found to achieve the near-optimal DoF in [104]. In addition to the gain in content delivery rate, employing multiple transmit antennas also allows reducing the subpacketization level required in coded caching [105].

By treating the transmission of coded subfiles as a coordinated beamforming problem, improved spectral efficiency is achieved in [106] by optimizing the beamforming vectors, which is also shown to achieve the same DoF as in [103] in special cases. Observing that the complexity of the beamforming problem grows exponentially with the number of subfiles delivered to each user, a low-complexity design was proposed in [107] by limiting the number of subfiles decoded by each user, which was shown to outperform the scheme in [106] in the high SNR regime with insufficient transmit antennas, and at all SNR and rate values with sufficient transmit antennas. Memory-sharing is proposed in [108] to apply the content placement scheme of [21] for a fraction of the library, which exploits both the spatial multiplexing gain and the global caching gain by sending a common message together with user-dependent messages. The impact of imperfect CSIT on achievable DoF is considered for MISO broadcast channels in [109]. Similarly to [108], [110] adopts memory-sharing, and proposes a joint unicast and multicast beamforming approach.

Chapter 3

Joint Beamforming for Unicast and Broadcast Transmission

3.1 Introduction

With the growing demand for multimedia streaming applications, research efforts to incorporate multicast and broadcast transmission into the cellular network architecture have intensified in recent years. In 3G networks, MBMS was introduced to support new point-to-multipoint radio bearers and multicast capability in the core network [111]. However, due to its reduced capacity, which did not meet the requirement of mass media services, MBMS has never been deployed commercially.

Following many field trials worldwide, the first commercial deployment of eMBMS, commercially known as LTE Broadcast, was launched in South Korea in 2014 [112]. eMBMS provides full integration and seamless transition between broadcast and unicast modes [113], and significant performance improvement with respect to MBMS, thanks to the higher and more flexible data rates provided by the LTE architecture. Furthermore, it also allows single frequency network (SFN) operation across different cells as in digital television broadcasting, since the LTE waveform is OFDM-based. While it is commonly accepted that eMBMS, in its current form, needs further enhancements to be adopted as a successful commercial platform for TV broadcasting [114], it has been proposed as a converged platform in the UHF band for TV and mobile broadband [115], [116].

For eMBMS TV services, a study has been carried out within 3GPP in 2015 for application scenarios and use cases, as well as for potential requirements and improvements [117]. In 2017, advances have been published in the 3GPP Release 14, including standardization of radio interfaces between mobile network operators and broadcasters and the possibility for free-to-air reception, which is an essential feature for broadcasting TV programs over mobile networks [118]. While the standardization and evolvement of point-to-multipoint transmission are primarily led by multimedia broadcasting services, point-to-multipoint transmission techniques have also been adopted in LTE-Advanced Pro for emerging use cases including vehicular to everything (V2X), Internet of things (IoT) and machine-type communication (MTC) [119]. In 2019, the Study Item on potential enhancements on the existing 5G architecture for 5G multicast-broadcast services has been approved by 3GPP, which opens the door to the standardization of MBMS in the 3GPP Release 17 for 5G [120].

While the existing standards of MBMS are based on orthogonal multiplexing, where MBMS and unicast services are scheduled in different time frames, superposition coding, as a NOMA technique, has been adopted in the next-generation TV broadcasting US standard ATSC 3.0 [121] under the name layer division multiplexing (LDM) [122]. At the cost of an increased complexity at the receivers, which need to perform interference cancellation by decoding the generic broadcast content prior to decoding the unicast content, LDM may provide significant gains especially when the superposed signals exhibit large disparities in terms of signal-to-noise-plus-interference ratio (SINR). This is expected to be the case for multiplexing broadcast and unicast services. In fact, the unicast throughput is limited by intercell interference; and hence, increasing the transmit unicast power across the network does not necessarily improve the unicast SINR. In contrast, broadcast does not suffer from intercell interference in an SFN, and increasing the broadcast power results in an increased SINR. This not only helps improve the reliability of the broadcast layer, but it also reduces the interference on the unicast messages as the broadcast layer can be decoded and cancelled more reliably. A performance comparison of LDM with TDM/FDM for unequal error protection in broadcast systems in the absence of multicell interference from an information theoretic perspective can be found in [123].

In this chapter, we study the performance of non-orthogonal unicast and broadcast transmission in a cellular network via LDM, in order to demonstrate and quantify its benefits compared to orthogonal transmission methods, i.e., TDM and FDM. We assume an SFN operation for the broadcast layer, while allowing arbitrarily clustered cooperation for the unicast data streams. Cooperative transmission for broadcast traffic, and potentially also for unicast data streams, takes place by means of distributed beamforming at multi-antenna base stations. To better account for potential practical impairments, and to evaluate the robustness of LDM in real systems, we also consider imperfections in channel state information (CSI) through an additive error model. Beamforming and power allocation between unicast and broadcast layers, and the so-called injection level in the LDM literature (see, e.g., [123]), are optimized with the aim of minimizing the sum-power under constraints on the user-specific unicast rates and the common broadcast rate. The optimization of orthogonal transmission via TDM/FDM is also studied for comparison, and the corresponding nonconvex optimization problems are tackled by means of successive convex approximation (SCA) techniques [124], as well as through the calculation of performance upper bounds by means of the S-procedure followed by semidefinite relaxation (SDR) [125].

Finally, we also present an efficient distributed implementation of the proposed LDM system based on the dual decomposition method. The dual decomposition based-algorithm allows each cluster of BSs cooperating to transmit a unicast message to obtain their beamforming vector locally with limited information exchange. A completely distributed implementation is not viable due to the presence of the broadcast layer, whose beamforming vector needs to be determined centrally at one of the BSs or in the cloud; however, local computation of the unicast beamforming vectors allows exploiting the computation resources distributed across the network, which can help parallelize these computations.

Orthogonal multiplexing of unicast and multicast services based on block diagonalization was presented in [126]. [127–129] studied the capacity region of Gaussian broadcast channels with a common message. In contrast to the literature, this work focuses on performance comparison between LDM and TDM for joint unicast and broadcast transmission, and have demonstrated the superiority of LDM over TDM when using optimization-based beamforming. After the publication of our conference paper, more recent studies have employed rate-splitting for joint unicast and broadcast transmission [130]. Specifically, the rate-splitting approach has been shown to outperform the conventional linear precoding scheme adopted here, with increased complexity at the transmitter and the receiver. Moreover, a better DoF can be achieved with rate-splitting in case of imperfect CSIT.

3.2 System Model

We investigate downlink transmission in a cellular network that serves both unicast and broadcast traffic. Specifically, we focus on a scenario in which a dedicated unicast data stream is to be delivered to each user, while there is a common broadcast data stream intended for all the users. A more general broadcast traffic model, in which distinct data streams are sent to different subsets of users, could be included in the analysis at the cost of a more cumbersome notation, but will not be further pursued in this chapter.

As illustrated in Fig. 3.1, the network is comprised of N cells, each consisting of a base station (BS) with N_T antennas and K single-antenna mobile users. The notation (n, k) identifies the k-th user in cell n. All BSs cooperate via joint beamforming for the broadcast stream to all the users, while an arbitrary cluster

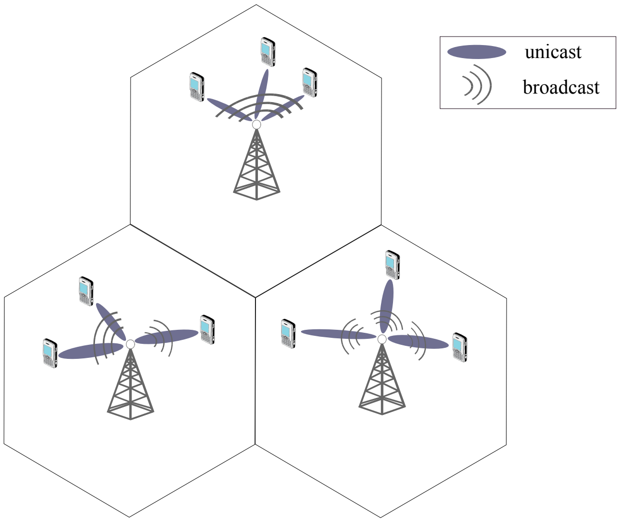


FIGURE 3.1: Illustration of a multicell network with N=3 cells and $K = 3$ users in each cell with simultaneous unicast and broadcast transmission.

$\mathcal{C}_{n,k}$ of BSs cooperate for the unicast transmission to user (n, k). Accordingly, all the BSs have access to the broadcast data stream, while only the BSs in cluster $\mathcal{C}_{n,k}$ are informed about the unicast data stream to be delivered to user (n, k). Note that, non-cooperative unicast transmission, whereby each BS serves only the users in its own cell, can be obtained as a special case when $\mathcal{C}_{n,k} = \{n\}$, for all users (n, k). Similarly, fully cooperative unicast transmission is obtained when $\mathcal{C}_{n,k} = \{1, \ldots, N\}$, for all users (n, k). We denote the set of users whose unicast messages are available at BS i as

$$\mathcal{U}_i = \{(n, k) \mid i \in \mathcal{C}_{n,k}\}. \tag{3.1}$$

We assume frequency-flat quasi-static complex channels, and define $\boldsymbol{h}_{i,n,k} \in \mathbb{C}^{N_T \times 1}$ as the channel vector from the BS in cell i to user (n, k). We use the notation $s^U_{n,k}$ to denote an encoded unicast symbol intended for user (n, k), and s^B to represent an encoded broadcast symbol. The signal received by user (n, k)

at any given channel use can then be written as

$$y_{n,k} = \sum_{i=1}^{N} \boldsymbol{h}_{i,n,k}^{H} \boldsymbol{x}_i + n_{n,k}, \tag{3.2}$$

where $\boldsymbol{x}_i \in \mathbb{C}^{N_T \times 1}$ is the symbol transmitted by BS i, and $n_{n,k} \sim \mathcal{CN}(0, \sigma_{n,k}^2)$ is the additive white Gaussian noise. We assume that both the intended and the interference signals at each user are in perfect synchronization without inter-symbol interference.

In practice, BSs have to operate with imperfect CSI. In FDD systems, it may arise from errors in downlink training-based CSI estimation, limited resolution in CSI feedback links, or from delays in CSI acquisition over fading channels, while in TDD systems, CSI errors are caused by impairments in channel estimation or imperfect channel reciprocity (see [131] and references therein). In this chapter, we consider only the imperfection on CSIT, and assume that the CSI is perfectly known at the receivers. As common in the literature, we model the CSI uncertainty with an additive error by setting

$$\boldsymbol{h}_{i,n,k} = \hat{\boldsymbol{h}}_{i,n,k} + \boldsymbol{e}_{i,n,k}, \tag{3.3}$$

where $\hat{\boldsymbol{h}}_{i,n,k} \in \mathbb{C}^{M \times 1}$ is the estimated complex channel vector from cell i to user (n, k) available at the BSs, and $\boldsymbol{e}_{i,n,k} \in \mathbb{C}^{N_T \times 1}$ is the additive channel error. For analytic convenience, we consider a bounded uncertainty set for CSI errors, which is typically used to model CSI imperfection resulting from quantization error due to feedback links of limited capacity. Hence, the set of channel vectors from BS i to user (n, k) can be defined as

$$\mathcal{H}_{i,n,k} = \{\boldsymbol{h}_{i,n,k} : \boldsymbol{h}_{i,n,k} = \hat{\boldsymbol{h}}_{i,n,k} + \boldsymbol{e}_{i,n,k},\ \boldsymbol{e}_{i,n,k}^{H} \boldsymbol{Q}_{i,n,k} \boldsymbol{e}_{i,n,k} \leq 1\},\ \forall i, n, k, \tag{3.4}$$

where $\boldsymbol{Q}_{i,n,k}$ is a known positive definite matrix. Accordingly, the structure of the uncertainty set of the quantization error vectors is known at the transmitters.

In what follows, we will consider two modes of transmission, namely orthogonal transmission via TDM and non-orthogonal transmission via LDM, where the former will serve as a benchmark to evaluate the potential performance gains from the LDM scheme.

3.2.1 TDM

We first consider the standard TDM approach based on the orthogonal transmission of unicast and broadcast signals. Note that orthogonalization can also be realized by means of other multiplexing schemes such as FDM, yielding the same mathematical formulation. With TDM, each transmission slot of duration T channel uses is divided into two subslots: a subslot of duration T_0 channel uses for unicast transmission, and a subslot of duration $T - T_0$ for broadcast transmission. Therefore, the signal $\boldsymbol{x}_i$ transmitted by cell i can be written as

$$\boldsymbol{x}_i = \begin{cases} \sum\limits_{(n,k)\in\mathcal{U}_i} \boldsymbol{w}^U_{i,n,k} s^U_{n,k} & \text{for } \; 0 \le t < T_0 \\ \boldsymbol{w}^B_i s^B & \text{for } \; T_0 \le t < T \end{cases}, \tag{3.5}$$

where $\boldsymbol{w}^U_{i,n,k} \in \mathbb{C}^{N_T \times 1}$ represents the unicast beamforming vector applied at the BS in cell i towards user (n,k), and $\boldsymbol{w}^B_i \in \mathbb{C}^{N_T \times 1}$ is the broadcast beamforming vector applied at the same BS.

The received signal $y_{n,k}$ at user (n,k) can be expressed as

$$y_{n,k} = \begin{cases} \Big(\sum\limits_{i\in\mathcal{C}_{n,k}} \boldsymbol{h}^H_{i,n,k} \boldsymbol{w}^U_{i,n,k} \Big) s^U_{n,k} + z_{n,k} + n_{n,k} & \text{for } 0 \le t < T_0 \\ \Big(\sum\limits_{i=1}^{N} \boldsymbol{h}^H_{i,n,k} \boldsymbol{w}^B_i \Big) s^B + n_{n,k} & \text{for } T_0 \le t < T \end{cases}, \tag{3.6}$$

where

$$z_{n,k} = \sum_{(p,q)\neq(n,k)} \left(\sum_{i\in\mathcal{C}_{p,q}} \boldsymbol{h}^H_{i,n,k} \boldsymbol{w}^U_{i,p,q} \right) s^U_{p,q} \tag{3.7}$$

denotes the interference at user (n, k).

3.2.2 LDM

In LDM, the transmitted signal $\boldsymbol{x}_i$ from the BS in cell i is the superposition of the broadcast and unicast signals for the entire time slot, which can be written as

$$\boldsymbol{x}_i = \boldsymbol{w}_i^B s^B + \sum_{(n,k)\in\mathcal{U}_i} \boldsymbol{w}_{i,n,k}^U s_{n,k}^U \quad \text{for } 0 \leq t \leq T, \tag{3.8}$$

for all channel uses in an entire time slot, i.e., for $0 \leq t \leq T$. We note that the power ratio between broadcast and unicast, which is referred to as the *injection level* (IL) in the literature (see, e.g., [123]), can be obtained as

$$\text{IL} = 10 \log_{10} \frac{P^B}{P^U}, \tag{3.9}$$

where $P^B = \sum_{i=1}^N ||\boldsymbol{w}_i^B||^2$ is the total broadcast power, and $P^U = \sum_{i=1}^N \sum_{(n,k)\in\mathcal{U}_i} ||\boldsymbol{w}_{i,n,k}^U||^2$ is the total unicast power. The received signal at user (n, k) is given by

$$\begin{aligned} y_{n,k} = & \Big(\sum_{i=1}^N \boldsymbol{h}_{i,n,k}^H \boldsymbol{w}_i^B\Big) s^B + \Big(\sum_{i\in\mathcal{C}_{n,k}} \boldsymbol{h}_{i,n,k}^H \boldsymbol{w}_{i,n,k}^U\Big) s_{n,k}^U \\ & + z_{n,k} + n_{n,k}, \quad \text{for } 0 \leq t \leq T, \end{aligned} \tag{3.10}$$

where $z_{n,k}$ is the interference as defined in (3.7).

It is remarked that, this chapter is based on our result using the common linear precoding technique [132], while a rate-splitting approach has recently been shown to outperform the linear precoding scheme adopted here, with increased transceiver complexity [130].

3.3 Problem Formulation

As common in the literature [131], we consider robust design that optimizes the system performance subject to worst-case QoS constraints. Specifically, the power minimization problem for the above systems can be expressed in the following form:

$$\min_{\{\boldsymbol{w}_i^B\},\{\boldsymbol{w}_{i,n,k}^U\}} \quad \sum_{i=1}^{N} \Big(||\boldsymbol{w}_i^B||^2 + \sum_{(n,k)\in\mathcal{U}_i} ||\boldsymbol{w}_{i,n,k}^U||^2\Big) \tag{3.11a}$$

$$\text{s.t.} \quad \min_{\mathcal{H}} \ \text{SINR}_{n,k}^B \geq \gamma^B, \ \forall n, k, \tag{3.11b}$$

$$\min_{\mathcal{H}} \ \text{SINR}_{n,k}^U \geq \gamma_{n,k}^U, \ \forall n, k, \tag{3.11c}$$

where the explicit expressions for the SINRs at user (n, k) for broadcast and unicast transmissions, namely $\text{SINR}_{n,k}^B$ and $\text{SINR}_{n,k}^U$ will be given below for TDM and LDM separately. The constraints in (3.11b) and (3.11c) are imposed on the worst-case SINRs for all possible channel realizations in the set $\mathcal{H} = \prod_{i,n,k} \mathcal{H}_{i,n,k}$. Note that, since all the users receive the same broadcast signal, we have enforced a common broadcast QoS requirement. In contrast, the unicast SINR requirements are allowed to be user-dependent.

It is remarked that, a common approach in the literature, as we adopt here, is to evaluate the performance of beamforming designs with respect to the power consumption for guaranteeing certain operational goals in practice. Alternatively, one may maximize a system utility function, e.g., weighted sum-rate or minimum SINR among users, subject to total or per-BS power constraints, which yields the best achievable performance when the system operates at full capacity.

3.3.1 TDM

From the expression of the received signal in (3.6), we derive the SINR for the broadcast layer in TDM for user (n,k) as

$$\mathrm{SINR}_{n,k}^{B\text{-TDM}} = \frac{|\boldsymbol{h}_{n,k}^{H}\boldsymbol{w}^{B}|^{2}}{\sigma_{n,k}^{2}}, \tag{3.12}$$

where $\boldsymbol{h}_{n,k} = [\boldsymbol{h}_{1,n,k}^{T}, \ldots, \boldsymbol{h}_{N,n,k}^{T}]^{T} \in \mathbb{C}^{NN_{T}\times 1}$ is the aggregated channel vector from all the BSs to user (n,k). All broadcast beamforming vectors are similarly aggregated into the vector $\boldsymbol{w}^{B} = [\boldsymbol{w}_{1}^{B^{T}}, \ldots, \boldsymbol{w}_{N}^{B^{T}}]^{T} \in \mathbb{C}^{NN_{T}\times 1}$. The SINR for the unicast layer is instead given as

$$\mathrm{SINR}_{n,k}^{U\text{-TDM}} = \frac{|\boldsymbol{h}_{n,k}^{(n,k)^{H}}\boldsymbol{w}_{n,k}^{U}|^{2}}{\sum\limits_{(p,q)\neq(n,k)} |\boldsymbol{h}_{n,k}^{(p,q)^{H}}\boldsymbol{w}_{p,q}^{U}|^{2} + \sigma_{n,k}^{2}}, \tag{3.13}$$

where $\boldsymbol{h}_{n,k}^{(p,q)} = [\boldsymbol{h}_{i,n,k}^{T}]_{i\in\mathcal{C}_{p,q}}^{T}$ is the aggregated channel vector to user (n,k) from all the BSs in cluster $\mathcal{C}_{p,q}$ of BSs that serve user (p,q), and $\boldsymbol{w}_{n,k}^{U} = [\boldsymbol{w}_{i,n,k}^{U^{T}}]_{i\in\mathcal{C}_{n,k}}^{T}$ is similarly defined as the aggregate unicast beamforming vector for user (n,k) from all the BSs in cluster $\mathcal{C}_{n,k}$.

We observe that the SINR targets $\gamma_{n,k}^{U\text{-TDM}}$ and $\gamma^{B\text{-TDM}}$ for unicast and broadcast traffic can be obtained from the corresponding transmission rates $R_{n,k}^{U}$ and R^{B}, respectively, as

$$\frac{T_0}{T}\log_2(1+\gamma_{n,k}^{U\text{-TDM}}) = R_{n,k}^{U}, \tag{3.14}$$

and

$$\frac{T-T_0}{T}\log_2(1+\gamma^{B\text{-TDM}}) = R^{B}. \tag{3.15}$$

3.3.2 LDM

With LDM, the broadcast layer, which is intended for all the users and usually has a higher SINR, is decoded first by treating unicast signals as noise, as in [9]. The users decode their unicast data streams after canceling the decoded broadcast message. The broadcast SINR in LDM for user (n, k) is hence obtained from the received signal (3.10) as follows

$$\mathrm{SINR}_{n,k}^{B\text{-LDM}} = \frac{|\boldsymbol{h}_{n,k}^{H} \boldsymbol{w}^{B}|^2}{\sum\limits_{(p,q)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^{U}|^2 + \sigma_{n,k}^2}, \tag{3.16}$$

while the unicast SINR is the same as TDM given in (3.13), i.e.,

$$\mathrm{SINR}_{n,k}^{U\text{-LDM}} = \mathrm{SINR}_{n,k}^{U\text{-TDM}}. \tag{3.17}$$

Similarly to TDM, SINR thresholds for unicast and broadcast can be obtained from the transmission rates $R_{n,k}^{U}$ and R^{B}, respectively, as

$$\log_2(1 + \gamma_{n,k}^{U\text{-LDM}}) = R_{n,k}^{U}, \tag{3.18}$$

and

$$\log_2(1 + \gamma^{B\text{-LDM}}) = R^{B}. \tag{3.19}$$

In [133], a performance lower bound on the power minimization problem is obtained by standard SDR, assuming that perfect CSI is available at all the BSs. In this chapter, the problem formulation incorporates CSI uncertainty in (3.11b) and (3.11c) by imposing constraints on the worst-case performance over all possible channel realizations on the optimization problem. The formulated worst-case quadratically-constrained quadratic program (QCQP) is intractable due to the induced additional constraints on the CSI error vectors. Nevertheless, the uncertainty due to CSI errors can be tackled by applying the S-procedure as in [131], as a result of which SDR can be employed as in the perfect CSI case to obtain a lower bound on the optimal solution. Furthermore, an achievable beamformer design under the worst-case SINR constraints will be obtained based on SCA, and its performance will be compared with the obtained lower bound.

3.4 Lower Bound via S-Procedure

The optimization problem in (3.11) contains an infinite number of constraints in (3.11b) and (3.11c), thus it is intractable. To address this issue, S-procedure [125] will be adopted to derive an equivalent but tractable problem formulation. Specifically, the constraints in (3.11b) and (3.11c) can be equivalently turned into a finite number of linear matrix inequalities, thereby allowing use of efficient optimization tools. Following the CSI error model in (3.4) we can form the aggregated CSI error vector $\boldsymbol{e}_{n,k}$ for user (n,k) consistent with the aggregated channel vector $\boldsymbol{h}_{n,k}$, and define the relaxed set of possible channel vectors to user (n,k) as:

$$\mathcal{H}_{n,k} \triangleq \{\boldsymbol{h}_{n,k} : \boldsymbol{h}_{n,k} = \hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k},\ \boldsymbol{e}_{n,k}^H \boldsymbol{Q}_{n,k} \boldsymbol{e}_{n,k} \leq 1\}, \tag{3.20}$$

where

$$\boldsymbol{Q}_{n,k} \triangleq \frac{1}{N} \begin{bmatrix} \boldsymbol{Q}_{1,n,k} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{Q}_{N,n,k} \end{bmatrix}. \tag{3.21}$$

It is noted that the set of possible channel vectors in (3.20) is a relaxed version of the original set given in (3.4). For reference, we present the S-procedure in the following lemma for completeness.

Lemma 3.1 (S-procedure)**.** *Let* $f_i(\boldsymbol{x}) \triangleq \boldsymbol{x}^H \boldsymbol{F}_i \boldsymbol{x} + \boldsymbol{g}_i^H \boldsymbol{x} + \boldsymbol{x}^H \boldsymbol{g}_i + c_i$*, for* $i = 0, 1$*, where* $\boldsymbol{F}_i \in \mathbb{C}^{NM \times NM}$ *is Hermitian semidefinite,* $\boldsymbol{g} \in \mathbb{C}^{NM \times 1}$*, and* $c_i \in \mathbb{R}$*, then* $f_1(\boldsymbol{x}) \leq 0$ *for all* $\boldsymbol{x}$ *satisfying* $f_0(\boldsymbol{x}) \leq 0$ *holds if and only if there exists a* $\lambda \geq 0$ *such that*

$$\begin{bmatrix} \boldsymbol{F}_1 & \boldsymbol{g}_1 \\ \boldsymbol{g}_1^H & c_1 \end{bmatrix} \preceq \lambda \begin{bmatrix} \boldsymbol{F}_0 & \boldsymbol{g}_0 \\ \boldsymbol{g}_0^H & c_0 \end{bmatrix}. \tag{3.22}$$

3.4.1 TDM

The constraint for the broadcast layer in (3.11b) can be rewritten as

$$(\hat{\boldsymbol{h}}_{n,k}^H + \boldsymbol{e}_{n,k}^H) \boldsymbol{W}^B (\hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k}) \geq \sigma_{n,k}^2 \gamma_{n,k}^B, \text{ for } \forall \boldsymbol{e}_{n,k}^H \boldsymbol{Q}_{n,k} \boldsymbol{e}_{n,k} \leq 1,$$

where $\boldsymbol{W}^B \triangleq \boldsymbol{w}^B \boldsymbol{w}^{B^H}$. By applying the S-procedure, the worst-case SINR constraint in (3.11b) can be recast as

$$\begin{bmatrix} \boldsymbol{W}^B & \boldsymbol{W}^B \hat{\boldsymbol{h}}_{n,k} \\ \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{W}^B & \frac{1}{\gamma_{n,k}^B} \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{W}^B \hat{\boldsymbol{h}}_{n,k} - \sigma_{n,k}^2 \end{bmatrix} + \lambda_{n,k}^B \begin{bmatrix} \boldsymbol{Q}_{n,k} & \boldsymbol{0} \\ \boldsymbol{0}^T & -1 \end{bmatrix} \succeq 0, \tag{3.23}$$

for some $\lambda_{n,k}^B \geq 0, \ \forall n, k$. Accordingly to Lemma 3.1, the constraints on the unicast transmissions in (3.11c) can be written as

$$\begin{aligned} &(\hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k})^H \Big(\frac{1}{\gamma_{n,k}^U} \boldsymbol{T}_{n,k}^T \boldsymbol{W}_{n,k}^U \boldsymbol{T}_{n,k} - \sum_{(p,q) \neq (n,k)} \boldsymbol{T}_{p,q}^T \boldsymbol{W}_{p,q}^U \boldsymbol{T}_{p,q} \Big) \cdot \\ &\quad \cdot (\hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k}) \geq \sigma_{n,k}^2, \ \text{ for } \forall \boldsymbol{e}_{n,k}^H \boldsymbol{Q}_{n,k} \boldsymbol{e}_{n,k} \leq 1, \end{aligned} \tag{3.24}$$

where $\boldsymbol{W}_{n,k}^U \triangleq \boldsymbol{w}_{n,k}^U \boldsymbol{w}_{n,k}^{U^H}$, and $\boldsymbol{T}_{p,q}$ is a constructed block matrix of dimension $|\mathcal{C}_{p,q}| \times N$ such that $\boldsymbol{h}_{n,k}^{(p,q)} = \boldsymbol{T}_{p,q} \boldsymbol{h}_{n,k}$. Following the S-procedure, the worst-case

SINR constraint for the unicast layer can be recast as

$$\begin{bmatrix} \boldsymbol{V}_{n,k} & \boldsymbol{V}_{n,k}\hat{\boldsymbol{h}}_{n,k} \\ \hat{\boldsymbol{h}}_{n,k}^H\boldsymbol{V}_{n,k} & \hat{\boldsymbol{h}}_{n,k}^H\boldsymbol{V}_{n,k}\hat{\boldsymbol{h}}_{n,k} - \sigma_{n,k}^2 \end{bmatrix} + \lambda_{n,k}^U \begin{bmatrix} \boldsymbol{Q}_{n,k} & \boldsymbol{0} \\ \boldsymbol{0}^T & -1 \end{bmatrix} \succeq 0, \forall n, k, \tag{3.25}$$

for some $\lambda_{n,k}^U \geq 0,\ \forall n, k$, where $\boldsymbol{V}_{n,k}$ is defined as

$$\boldsymbol{V}_{n,k} \triangleq \frac{1}{\gamma_{n,k}^U}\boldsymbol{T}_{n,k}^T\boldsymbol{W}_{n,k}^U\boldsymbol{T}_{n,k} - \sum_{(p,q)\neq(n,k)} \boldsymbol{T}_{p,q}^T\boldsymbol{W}_{p,q}^U\boldsymbol{T}_{p,q}. \tag{3.26}$$

Following these transforms and definitions, the problem in (3.11) can be relaxed to a tractable semidefinite program by dropping the rank constraints on matrices $\boldsymbol{W}^B$ and $\boldsymbol{W}_{n,k}^U$. Specifically, for TDM, the relaxed problem after SDR is given by

$$\min_{\boldsymbol{W}^B,\{\boldsymbol{W}_{n,k}^U\},\{\lambda_{n,k}^B\},\{\lambda_{n,k}^U\}} \quad \mathrm{tr}(\boldsymbol{W}^B) + \sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{tr}(\boldsymbol{W}_{n,k}^U) \tag{3.27a}$$

$$\text{s.t.} \quad (3.23) \text{ and } (3.25), \tag{3.27b}$$

$$\lambda_{n,k}^B \geq 0, \lambda_{n,k}^U \geq 0,\ \forall n, k. \tag{3.27c}$$

3.4.2 LDM

Similar to the analysis in TDM, the constraint on the broadcast transmission in LDM can be equivalently written as

$$\begin{bmatrix} \boldsymbol{U} & \boldsymbol{U}\hat{\boldsymbol{h}}_{n,k} \\ \hat{\boldsymbol{h}}_{n,k}^H\boldsymbol{U} & \hat{\boldsymbol{h}}_{n,k}^H\boldsymbol{U}\hat{\boldsymbol{h}}_{n,k} - \sigma_{n,k}^2 \end{bmatrix} + \lambda_{n,k}^B \begin{bmatrix} \boldsymbol{Q}_{n,k} & \boldsymbol{0} \\ \boldsymbol{0}^T & -1 \end{bmatrix} \succeq 0, \tag{3.28}$$

where $\lambda_{n,k}^U \geq 0,\ \forall n, k$, and $\boldsymbol{U}$ is defined as

$$\boldsymbol{U} \triangleq \frac{1}{\gamma_{n,k}^B}\boldsymbol{W}^B - \sum_{(p,q)} \boldsymbol{T}_{p,q}^T\boldsymbol{W}_{p,q}^U\boldsymbol{T}_{p,q}. \tag{3.29}$$

The unicast constraint in LDM can be reformulated as in (3.25), hence the relaxed problem after dropping the rank-1 constraints on matrices $\boldsymbol{W}^B$ and $\boldsymbol{W}_{n,k}^U$ is obtained as follows:

$$\min_{\boldsymbol{W}^B,\{\boldsymbol{W}_{n,k}^U\},\{\lambda_{n,k}^B\},\{\lambda_{n,k}^U\}} \quad \mathrm{tr}(\boldsymbol{W}^B) + \sum_{n=1}^{N}\sum_{k=1}^{K} \mathrm{tr}(\boldsymbol{W}_{n,k}^U) \tag{3.30a}$$

$$\text{s.t.} \quad (3.25) \text{ and } (3.28), \tag{3.30b}$$

$$\lambda_{n,k}^B \geq 0, \lambda_{n,k}^U \geq 0, \ \forall n, k. \tag{3.30c}$$

As the rank-1 constraint has been dropped in (3.27) and (3.30), the corresponding optimal solutions provide lower bounds on the optimal solutions of the original problems in (3.11). Note that, under perfect CSI, i.e., $\boldsymbol{e}_{i,n,k} = \boldsymbol{0}$, the problem formulation in (11) boils down to the one presented in [133], and the solution obtained by first applying the S-procedure is equal to that obtained directly by SDR.

3.5 Upper Bound via SCA

Instead of adopting Gaussian randomization [134] to obtain a feasible (achievable) beamforming scheme, we leverage the SCA method [124] to obtain an achievable beamformer, which yields an upper bound on the minimum required power. In particular, by rewriting the nonconvex QoS constraints as the difference of convex (DC) functions, the SCA algorithm reduces to the conventional convex-concave procedure [135]. We remark that the SCA scheme is known to converge to a stationary point of the original problem [124].

In order to apply the SCA approach, each nonconvex constraint in (3.11) will be expressed as

$$g(\boldsymbol{w}) = g^{+}(\boldsymbol{w}) - g^{-}(\boldsymbol{w}) \leq 0, \tag{3.31}$$

where $g^{+}(\boldsymbol{w})$ and $g^{-}(\boldsymbol{w})$ are both convex functions on the set of all beamforming vectors $\boldsymbol{w}$. Then a convex upper bound is obtained by linearizing the nonconvex part around any given vector $\boldsymbol{u}$, yielding the stricter constraint on the solution $\boldsymbol{w}$ as

$$\tilde{g}(\boldsymbol{w};\boldsymbol{u}) \triangleq g^{+}(\boldsymbol{w}) - g^{-}(\boldsymbol{u}) - \triangledown_{\boldsymbol{w}} g^{-}(\boldsymbol{u})^{T}(\boldsymbol{w}-\boldsymbol{u}) \leq 0. \tag{3.32}$$

3.5.1 TDM

The constraint in (3.11b) on the broadcast layer can be approximated and replaced by the following tighter constraint:

$$|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}| - |\boldsymbol{e}_{n,k}^{H}\boldsymbol{w}^{B}| \geq \sqrt{\gamma^{B}}\sigma_{n,k} \text{ for } \forall\ \boldsymbol{e}_{n,k}^{H}\boldsymbol{Q}_{n,k}\boldsymbol{e}_{n,k} \leq 1, \tag{3.33}$$

which can be further tightened as:

$$|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}| - \|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\| \geq \sqrt{\gamma^{B}}\sigma_{n,k}, \tag{3.34}$$

since $|\boldsymbol{e}_{n,k}^{H}\boldsymbol{w}^{B}| \leq \|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\|$ holds for the CSI error vectors $\boldsymbol{e}_{n,k}$ as we have $\boldsymbol{e}_{n,k} \in \{\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{u} \mid \|\boldsymbol{u}\| \leq 1\}$.

The constraint in (3.34) is in the DC form, for which SCA can be adopted to obtain an iterative algorithm which converges to a stationary point of the original problem. The constraint at the ν-th iteration of the SCA algorithm is given by

$$\sqrt{\gamma^{B}}\sigma_{n,k} + \|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\| + |\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}(\nu)| - 2\frac{\hat{\boldsymbol{h}}_{n,k}^{H}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^{H}}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}(\nu)|}\boldsymbol{w}^{B} \leq 0,\ \forall n,k. \tag{3.35}$$

Also, the constraint in (3.11c) for the unicast transmission can be tightened by considering the worst-case SINR, i.e.,

$$\frac{\min\limits_{\mathcal{H}} \ |\boldsymbol{h}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U|^2}{\max\limits_{\mathcal{H}} \sum\limits_{(p,q)\neq(n,k)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U|^2 + \sigma_{n,k}^2} \geq \gamma_{n,k}^U, \ \text{for } \forall n, k, \tag{3.36}$$

which can then be replaced equivalently by the following set of constraints:

$$\max_{\mathcal{H}} \ |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U| \leq \beta_{n,k}^{(p,q)}, \ \forall n, k, \forall (p,q) \neq (n,k), \tag{3.37a}$$

$$\min_{\mathcal{H}} \ |\boldsymbol{h}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U| \geq t_{n,k}^U, \tag{3.37b}$$

$$\gamma_{n,k}^U \Big(\sum_{(p,q)\neq(n,k)} \beta_{n,k}^{(p,q)^2} + \sigma_{n,k}^2 \Big) - t_{n,k}^{U2} \leq 0, \tag{3.37c}$$

where $\{t_{n,k}^U\}$ and $\{\beta_{n,k}^{(p,q)}\}$ are auxiliary variables. Note that $\beta_{n,k}^{(p,q)}$ indicates the interference power from BSs in the cluster $\mathcal{C}_{p,q}$ to user (n,k), and $t_{n,k}^U$ indicates the received unicast power at user (n,k). The constraint in (3.37a) and (3.37b) can be further relaxed by

$$|\hat{\boldsymbol{h}}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U| + |\boldsymbol{Q}_{n,k}^{(p,q)^{-1/2}} \boldsymbol{w}_{p,q}^U| \leq \beta_{n,k}^{(p,q)}, \forall n, k, \forall (p,q) \neq (n,k), \tag{3.38}$$

and

$$t_{n,k}^U + \|\boldsymbol{Q}_{n,k}^{(n,k)^{-1/2}} \boldsymbol{w}_{n,k}^U\| - |\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U| \leq 0, \tag{3.39}$$

respectively, where $\boldsymbol{Q}_{n,k}^{(p,q)^{-1/2}} = \boldsymbol{Q}_{n,k}^{-1/2} \boldsymbol{T}_{p,q}$. According to (3.31) and (3.32), in the SCA algorithm, the corresponding constraints in the ν-th iteration for (3.37c) and (3.39) can be written as

$$\gamma_{n,k}^U \Big(\sum_{(p,q)\neq(n,k)} \beta_{n,k}^{(p,q)^2} + \sigma_{n,k}^2 \Big) + t_{n,k}^{U2}(\nu) - 2t_{n,k}^U(\nu) t_{n,k}^U \leq 0, \forall n, k, \tag{3.40}$$

Table 3.1: SCA Algorithm for the Joint Beamforming Problem with TDM and LDM

```
STEP 0: Set $\nu = 1$. Set a step size $\mu$.
        Initialize $\boldsymbol{w}^B(1)$ and $\boldsymbol{w}_{n,k}^U(1)$ with feasible values
STEP 1: If a stopping criterion is satisfied, then STOP
STEP 2: Set $\boldsymbol{w}^B(\nu+1) = \boldsymbol{w}^B(\nu) + \mu(\boldsymbol{w}^B - \boldsymbol{w}^B(\nu))$,
        $\boldsymbol{w}_{n,k}^U(\nu+1) = \boldsymbol{w}_{n,k}^U(\nu) + \mu(\boldsymbol{w}_{n,k}^U - \boldsymbol{w}_{n,k}^U(\nu))$,
        where $\{\boldsymbol{w}^B\}$ and $\{\boldsymbol{w}_{n,k}^U\}$ are obtained as solutions
        of problems (3.42) for TDM and (3.48) for LDM
STEP 3: Set $\nu = \nu + 1$, and go to STEP 1
```

and

$$t_{n,k}^U + \|\boldsymbol{Q}_{n,k}^{(n,k)^{-\frac{1}{2}}} \boldsymbol{w}_{n,k}^U\| + |\boldsymbol{h}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U(\nu)| - 2\frac{\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \hat{\boldsymbol{h}}_{n,k}^{(n,k)} \boldsymbol{w}_{n,k}^{U^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U(\nu)|} \boldsymbol{w}_{n,k}^U \leq 0, \ \forall n, k, \tag{3.41}$$

respectively.

Due to the fact that the feasible convexified constraints in (3.35), (3.38), (3.40) and (3.41) are stricter than the original constraints in (3.11), the solution obtained at each iteration is feasible for the original problem (3.11) as long as a feasible initial point is available. When the stopping criterion is satisfied, we take the last iteration as the solution of the SCA algorithm. Please refer to Table 3.1 for an algorithmic description of the SCA approach.

When obtaining the numerical results in the next section, initialization of the SCA algorithm is carried out based on the solution $\{\boldsymbol{W}^B\}$ and $\{\boldsymbol{W}_{n,k}^U\}$ obtained from the S-procedure. Specifically, we perform a rank-1 reduction of matrices $\{\boldsymbol{W}^B\}$ and $\{\boldsymbol{W}_{n,k}^U\}$, obtaining vectors $\{\boldsymbol{w}^B\}$ and $\{\boldsymbol{w}_{n,k}^U\}$, respectively, as the largest principal component. These vectors are then scaled with the smallest common factor t, which is evaluated through line search, to satisfy constraints (3.11b) and (3.11c), yielding the initial points $\{\boldsymbol{w}^B(1)\}$ and $\{\boldsymbol{w}_{n,k}^U(1)\}$ for SCA. If a feasible value for t is not found through a line search, then the SCA method is considered to be infeasible. Further discussion on this point can be found in Section V.

As a summary, the relaxed version of the problem for (3.11) in TDM in the SCA form is given as

$$\min_{\boldsymbol{w}^B,\{\boldsymbol{w}^U_{n,k}\},\{\beta^{(p,q)}_{n,k}\},\{t^U_{n,k}\}} \quad \|\boldsymbol{w}^B\|^2 + \sum_{(n,k)} \|\boldsymbol{w}^U_{n,k}\|^2 \tag{3.42a}$$

$$\text{s.t.} \quad (3.35), \ (3.38), \ (3.40), \ \text{and } (3.41). \tag{3.42b}$$

3.5.2 LDM

Similarly to the TDM approach, the constraint in (3.11b) can be relaxed as the worst-case SINR constraint, i.e.,

$$\frac{\min\limits_{\mathcal{H}} \ |\boldsymbol{h}^H_{n,k}\boldsymbol{w}^B|^2}{\max\limits_{\mathcal{H}} \sum\limits_{(p,q)} |\boldsymbol{h}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}|^2 + \sigma^2_{n,k}} \geq \gamma^B, \tag{3.43}$$

which is then replaced by the following equivalent constraints:

$$\max_{\mathcal{H}} \ |\boldsymbol{h}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}| \leq \beta^{(p,q)}_{n,k}, \tag{3.44a}$$

$$\min_{\mathcal{H}} \ |\boldsymbol{h}^H_{n,k}\boldsymbol{w}^B| \geq t^B_{n,k}, \tag{3.44b}$$

$$\gamma^U_{n,k}\big(\sum_{(p,q)} \beta^{(p,q)^2}_{n,k} + \sigma^2_{n,k}\Big) - t^{B^2}_{n,k} \leq 0 \tag{3.44c}$$

for all n, k, where $\{t^B_{n,k}\}$ are auxiliary variables indicating the received broadcast power at user (n, k). Similarly to the relaxation we adopt for the TDM case, the constraint in (3.44a) can be relaxed as

$$|\hat{\boldsymbol{h}}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}| + |\boldsymbol{Q}^{(p,q)^{-1/2}}_{n,k}\boldsymbol{w}^U_{p,q}| \leq \beta^{(p,q)}_{n,k}, \ \forall n, k, p, q \tag{3.45}$$

for all n, k. The constraint in (3.44b) can be relaxed as

$$t^B_{n,k} + \|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\boldsymbol{w}^B\| - |\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B| \leq 0, \tag{3.46}$$

which is in the convex-concave form. According to (3.31) and (3.32), in the SCA algorithm, the corresponding constraints in the ν-th iteration for (3.44c) and (3.46) can be written as

$$\gamma^B_{n,k}\Big(\sum_{(p,q)}\beta^{(p,q)^2}_{n,k} + \sigma^2_{n,k}\Big) + t^{B^2}_{n,k}(\nu) - 2t^B_{n,k}(\nu)t^B_{n,k} \leq 0, \forall n, k, \tag{3.47}$$

and

$$t^B_{n,k} + \|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\boldsymbol{w}^B\| + |\boldsymbol{h}^H_{n,k}\boldsymbol{w}^B(\nu)| - 2\frac{\hat{\boldsymbol{h}}^H_{n,k}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B(\nu)|}\boldsymbol{w}^B \leq 0,\ \forall n, k,$$

respectively. As a summary, the relaxed version of the (3.11) for LDM in the SCA form is given as

$$\min_{\boldsymbol{w}^B,\{\boldsymbol{w}^U_{n,k}\},\{\beta^{(p,q)}_{n,k}\},\{t^B_{n,k}\},\{t^U_{n,k}\}} \|\boldsymbol{w}^B\|^2 + \sum_{(n,k)}\|\boldsymbol{w}^U_{n,k}\|^2 \tag{3.48a}$$

$$\text{s.t. } (3.40),\ (3.41),\ (3.45), (3.47),\ \text{and } (3.48a). \tag{3.48b}$$

3.6 Distributed Approach

In this section, we propose a distributed algorithm to solve the SCA problem in (3.48) using dual decomposition as in [124]. In particular, while the broadcast beamforming vector $\boldsymbol{w}^B$ is designed at a central node that gathers full CSI between all the BSs and the users, the optimization of unicast beamforming vectors $\{\boldsymbol{w}^U_{n,k}\}$ is offloaded to the processing unit of the corresponding cluster $\mathcal{C}_{n,k}$, which can be located at one of the BSs within the cluster. This distributed implementation is made possible by the fact that the optimization of $\{\boldsymbol{w}^U_{n,k}\}$ can be decomposed into NK independent subproblems, and the processing unit of each cluster $\mathcal{C}_{n,k}$ can calculate $\boldsymbol{w}_{n,k}^{U}$ locally, but still optimally, based only on local CSI, in addition to certain limited information exchange with other clusters.

The benefits of this distributed implementation are as follows. First, it reduces the computational requirements on the central processing unit as compared to the centralized approach. This is done by parallelizing the computation by distributing it across many nodes in the network. Second, transmitting all the CSI back to a central unit may lead to increased CSI uncertainty, as the CSI could need to be further compressed to be communicated to a single node. In our formulation here, for simplicity, we consider the same CSI error variance for both the broadcast and unicast beamforming optimization problems. Finally, in the absence of a broadcast message destined for the whole network, all computations can be carried out locally at the cluster heads.

For clarity, to start, we reproduce the problem in (3.48):

$$\min \|\boldsymbol{w}^{B}\|^2 + \sum_{(n,k)} \|\boldsymbol{w}_{n,k}^{U}\|^2 \tag{3.49a}$$

$$\text{s.t. } |\hat{\boldsymbol{h}}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^{U}| + \|\boldsymbol{Q}_{n,k}^{(p,q)^{-1/2}} \boldsymbol{w}_{p,q}^{U}\| \leq \beta_{n,k}^{(p,q)}, \forall n, k, p, q, \tag{3.49b}$$

$$\gamma_{n,k}^{U} \left(\sum_{(p,q)\neq(n,k)} \beta_{n,k}^{(p,q)^2} + \sigma_{n,k}^{2} \right) + t_{n,k}^{U^2}(\nu) - 2\, t_{n,k}^{U}(\nu) t_{n,k}^{U} \leq 0,\ \forall n, k, \tag{3.49c}$$

$$t_{n,k}^{U} + \|\boldsymbol{Q}_{n,k}^{(n,k)^{-\frac{1}{2}}} \boldsymbol{w}_{n,k}^{U}\| + |\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^{U}(\nu)| - 2 \frac{\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \hat{\boldsymbol{h}}_{n,k}^{(n,k)} \boldsymbol{w}_{n,k}^{U^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^{U}(\nu)|} \boldsymbol{w}_{n,k}^{U} \leq 0,\ \forall n, k, \tag{3.49d}$$

$$\gamma_{n,k}^{B} \left(\sum_{(p,q)} \beta_{n,k}^{(p,q)^2} + \sigma_{n,k}^{2} \right) + t_{n,k}^{B^2}(\nu) - 2 t_{n,k}^{B}(\nu) t_{n,k}^{B} \leq 0,\ \forall n, k, \tag{3.49e}$$

$$t_{n,k}^{B} + \|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}} \boldsymbol{w}^{B}\| + |\hat{\boldsymbol{h}}_{n,k}^{H} \boldsymbol{w}^{B}(\nu)|$$

$$- 2\frac{\hat{\boldsymbol{h}}_{n,k}^{H}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}(\nu)|}\boldsymbol{w}^{B} \leq 0,\ \forall n,k. \tag{3.49f}$$

We now introduce Lagrangian multipliers $\boldsymbol{\lambda} \triangleq \{\lambda_{n,k}^{(p,q)}\}, \boldsymbol{\mu} \triangleq \{\mu_{n,k}\}, \boldsymbol{\kappa} \triangleq \{\kappa_{n,k}\}, \boldsymbol{\xi} \triangleq \{\xi_{n,k}\}, \boldsymbol{\rho} \triangleq \{\rho_{n,k}\}$ for the constraints in (3.49b)-(3.49f), respectively, and define $\boldsymbol{z} \triangleq \left(\boldsymbol{w}^{B}, \{\boldsymbol{w}_{n,k}^{U}\}, \{\beta_{n,k}^{(p,q)}\}, \{t_{n,k}^{B}\}, \{t_{n,k}^{U}\}\right)$. Then the Lagrangian of (3.49) can then be obtained as

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\xi}, \boldsymbol{\rho}, \boldsymbol{z}; \boldsymbol{z}(\nu)\right) &= \mathcal{L}_{\boldsymbol{w}^{B}}\left(\boldsymbol{\rho}, \boldsymbol{w}^{B}; \boldsymbol{w}^{B}(\nu)\right) \\
&+ \sum_{n,k} \mathcal{L}_{\boldsymbol{w}_{n,k}^{U}}\left(\boldsymbol{\lambda}_{n,k}, \kappa_{n,k}, \boldsymbol{w}_{n,k}^{U}; \boldsymbol{w}_{n,k}^{U}(\nu)\right) \\
&+ \sum_{n,k} \mathcal{L}_{\boldsymbol{\beta}_{n,k}}\left(\boldsymbol{\lambda}_{n,k}, \mu_{n,k}, \xi_{n,k}, \boldsymbol{\beta}_{n,k}\right) \\
&+ \sum_{n,k} \mathcal{L}_{t_{n,k}^{U}}\left(\mu_{n,k}, \kappa_{n,k}, t_{n,k}^{U}; t_{n,k}^{U}(\nu)\right) \\
&+ \sum_{n,k} \mathcal{L}_{t_{n,k}^{B}}\left(\xi_{n,k}, \rho_{n,k}, t_{n,k}^{B}; t_{n,k}^{B}(\nu)\right),
\end{aligned} \tag{3.50}$$

where

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{w}^{B}}\left(\boldsymbol{\rho}, \boldsymbol{w}^{B}; \boldsymbol{w}^{B}(\nu)\right) &\triangleq \|\boldsymbol{w}^{B}\|^{2} + \sum_{n,k} \rho_{n,k}\|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\| \\
&-2\sum_{n,k} \rho_{n,k}\frac{\hat{\boldsymbol{h}}_{n,k}^{H}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}(\nu)|}\boldsymbol{w}^{B},
\end{aligned} \tag{3.51a}$$

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{w}_{n,k}^{U}}\left(\boldsymbol{\lambda}^{(n,k)}, \kappa_{n,k}, \boldsymbol{w}_{n,k}^{U}; \boldsymbol{w}_{n,k}^{U}(\nu)\right) &\triangleq \|\boldsymbol{w}_{n,k}^{U}\|^{2} \\
&+ \sum_{p,q} \lambda_{p,q}^{(n,k)}\left(|\hat{\boldsymbol{h}}_{p,q}^{(n,k)^H}\boldsymbol{w}_{n,k}^{U}| + |\boldsymbol{Q}_{p,q}^{(n,k)^{-1/2}}\boldsymbol{w}_{n,k}^{U}|\right) \\
&+\kappa_{n,k}\|\boldsymbol{Q}_{n,k}^{(n,k)^{-\frac{1}{2}}}\boldsymbol{w}_{n,k}^{U}\| - 2\kappa_{n,k}\frac{\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H}\hat{\boldsymbol{h}}_{n,k}^{(n,k)}\boldsymbol{w}_{n,k}^{U^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H}\boldsymbol{w}_{n,k}^{U}(\nu)|}\boldsymbol{w}_{n,k}^{U},
\end{aligned} \tag{3.51b}$$

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\beta}_{n,k}}\left(\boldsymbol{\lambda}_{n,k}, \mu_{n,k}, \xi_{n,k}, \boldsymbol{\beta}_{n,k}\right) &\triangleq -\sum_{p,q} \lambda_{n,k}^{(p,q)}\beta_{n,k}^{(p,q)} \\
&+\mu_{n,k}\gamma_{n,k}^{U}\sum_{(p,q)\neq(n,k)} \beta_{n,k}^{(p,q)^2} + \xi_{n,k}\gamma_{n,k}^{B}\sum_{(p,q)}\beta_{n,k}^{(p,q)^2},
\end{aligned} \tag{3.51c}$$

$$\mathcal{L}_{t_{n,k}^{U}}\left(\mu_{n,k}, \kappa_{n,k}, t_{n,k}^{U}; t_{n,k}^{U}(\nu)\right) \triangleq -2\mu_{n,k}t_{n,k}^{U}(\nu)t_{n,k}^{U} + \kappa_{n,k}t_{n,k}^{U}, \tag{3.51d}$$

$$\mathcal{L}_{t^B_{n,k}}\big(\xi_{n,k},\rho_{n,k},t^B_{n,k};t^B_{n,k}(\nu)\big)\triangleq -2\xi_{n,k}t^B_{n,k}(\nu)t^B_{n,k}+\rho_{n,k}t^B_{n,k}. \tag{3.51e}$$

The optimization problem in (3.49) is strongly convex and satisfies Slater's condition, thus strong duality holds. Therefore, the optimal solution can be obtained by solving its dual problem, which is given by

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho}} \quad D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right) \tag{3.52a}$$

$$\text{s.t.} \quad \boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\mu}\geq\mathbf{0},\boldsymbol{\kappa}\geq\mathbf{0},\boldsymbol{\xi}\geq\mathbf{0},\boldsymbol{\rho}\geq\mathbf{0}, \tag{3.52b}$$

where the dual function $D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right)$ is obtained by minimizing the Lagrangian over the primal variables as

$$D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right)=\min_{\boldsymbol{w}^B}\mathcal{L}_{\boldsymbol{w}^B}\left(\boldsymbol{\rho},\boldsymbol{w}^B;\boldsymbol{w}^B(\nu)\right) \tag{3.53a}$$

$$+\sum_{n,k}\min_{\boldsymbol{w}^U_{n,k}}\mathcal{L}_{\boldsymbol{w}^U_{n,k}}\left(\boldsymbol{\lambda}_{n,k},\kappa_{n,k},\boldsymbol{w}^U_{n,k};\boldsymbol{w}^U_{n,k}(\nu)\right) \tag{3.53b}$$

$$+\sum_{n,k}\min_{\beta^{(p,q)}_{n,k}}\mathcal{L}_{\{\beta^{(p,q)}_{n,k}\}}\left(\boldsymbol{\lambda}_{n,k},\mu_{n,k},\xi_{n,k},\beta^{(p,q)}_{n,k}\right) \tag{3.53c}$$

$$+\sum_{n,k}\min_{t^U_{n,k}}\mathcal{L}_{t^U_{n,k}}\left(\mu_{n,k},\kappa_{n,k},t^U_{n,k};t^U_{n,k}(\nu)\right) \tag{3.53d}$$

$$+\sum_{n,k}\min_{t^B_{n,k}}\mathcal{L}_{t^B_{n,k}}\left(\xi_{n,k},\rho_{n,k},t^B_{n,k};t^B_{n,k}(\nu)\right), \tag{3.53e}$$

yielding the optimal solutions $\hat{\boldsymbol{z}}\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho}\right)=\left(\hat{\boldsymbol{w}}^B_{n,k},\{\hat{\boldsymbol{w}}^U_{n,k}\},\{\hat{\beta}^{(p,q)}_{n,k}\},\{\hat{t}^U_{n,k}\},\{\hat{t}^B_{n,k}\}\right)$. The optimization over $\boldsymbol{w}^U_{n,k},\beta^{(p,q)}_{n,k},t^U_{n,k},t^B_{n,k}$ in (3.53) can be decomposed into NK separable subproblems. The dual function $D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right)$ is differentiable with its gradient given by

$$\nabla_{\lambda^{n,k}_{p,q}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right)=|\hat{\boldsymbol{h}}^{(n,k)^H}_{p,q}\hat{\boldsymbol{w}}^U_{n,k}|$$

$$+\|\boldsymbol{Q}^{(n,k)^{-1/2}}_{p,q}\hat{\boldsymbol{w}}^U_{n,k}\|-\hat{\beta}^{(n,k)}_{p,q},\ \forall p,q, \tag{3.54a}$$

$$\nabla_{\mu_{n,k}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right)=\gamma^U_{n,k}\left(\sum_{(p,q)\neq(n,k)}\hat{\beta}^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\right.$$

$$
\begin{aligned}
&+ \hat{t}_{n,k}^{U^2}(\nu) - 2t_{n,k}^{U}(\nu)\hat{t}_{n,k}^{U}, && (3.54\text{b}) \\
\nabla_{\kappa_{n,k}} D\left(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\xi}, \boldsymbol{\rho}; \hat{\boldsymbol{z}}(\nu)\right) &= \hat{t}_{n,k}^{U} + \|\boldsymbol{Q}_{n,k}^{(n,k)^{-\frac{1}{2}}} \hat{\boldsymbol{w}}_{n,k}^{U}\| \\
&+ |\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^{U}(\nu)| - 2\frac{\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \hat{\boldsymbol{h}}_{n,k}^{(n,k)} \boldsymbol{w}_{n,k}^{U^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^{U}(\nu)|} \hat{\boldsymbol{w}}_{n,k}^{U}, && (3.54\text{c}) \\
\nabla_{\xi_{n,k}} D\left(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\xi}, \boldsymbol{\rho}; \hat{\boldsymbol{z}}(\nu)\right) &= \gamma_{n,k}^{B} \left(\sum_{(p,q)} \hat{\beta}_{n,k}^{(p,q)^2} + \sigma_{n,k}^2 \right) \\
&+ t_{n,k}^{B^2}(\nu) - 2t_{n,k}^{B}(\nu)\hat{t}_{n,k}^{B}, && (3.54\text{d}) \\
\nabla_{\rho_{n,k}} D\left(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\xi}, \boldsymbol{\rho}; \hat{\boldsymbol{z}}(\nu)\right) &= \hat{t}_{n,k}^{B} + \|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}} \hat{\boldsymbol{w}}^{B}\| \\
&+ |\hat{\boldsymbol{h}}_{n,k}^{H} \hat{\boldsymbol{w}}^{B}(\nu)| - 2\frac{\hat{\boldsymbol{h}}_{n,k}^{H} \hat{\boldsymbol{h}}_{n,k} \boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{H} \boldsymbol{w}^{B}(\nu)|} \hat{\boldsymbol{w}}^{B}, && (3.54\text{e})
\end{aligned}
$$

all of which can be computed efficiently in a distributed manner.

Overall, the obtained algorithm is a double-loop scheme. The outer loop consists of the SCA iterations as described in Table 3.1. In each of the SCA iteration, gradient descent based dual ascent algorithm is adopted. First, the primal variable $\boldsymbol{z}^j$ is updated by solving the optimization problems outlined in (3.53a)-(3.53e), each of which is solved by solving NK subproblems. Specifically, the update of $\boldsymbol{w}_{n,k}^{U^j}$ only requires local CSI, i.e., $\hat{\boldsymbol{h}}_{p,q}^{(n,k)}$ for $\forall p, q$, and other local information such as $\boldsymbol{\lambda}^{(n,k)}$ and $\kappa_{n,k}$. Similarly, the updates of $t_{n,k}^{U}$ and $t_{n,k}^{B}$ only require local information. On the other hand, the update of the networkwide beamforming vector $\boldsymbol{w}^{B^j}$ needs full CSI across the network, as well as gathered information $\rho_{n,k}$ from all the clusters. The update of $\boldsymbol{\beta}_{n,k}$, which measures the received interference powers at user (n, k) from BSs outside the cluster $\mathcal{C}_{n,k}$, involves the exchange of $\{\lambda_{n,k}^{(p,q)}\}$ from all p, q. Once the primal variable is updated, dual variable updates can be executed with the gradient descent method, with gradient given in (3.54a)-(3.54e), respectively. Note that the update of dual variables can be performed locally with the message $\boldsymbol{w}^{B^j}$ from the central processing unit. The detailed algorithm description can be found in Table 3.2. We finally

Table 3.2: Distributed Algorithm within the v-th SCA iteration in LDM

```
STEP 0: Set $j = 1$. Initialize dual variables $\boldsymbol{\lambda}^0, \boldsymbol{\mu}^0, \boldsymbol{\kappa}^0, \boldsymbol{\xi}^0, \boldsymbol{\rho}^0$.
STEP 1: If the stopping criterion is satisfied, then STOP
STEP 2: At the central node:
            solve (3.53a) to obtain $\boldsymbol{w}^{B^j}$
        At each cluster $\mathcal{C}_{n,k}$:
            update $\boldsymbol{w}^{U^j}_{n,k}, t^{U^j}_{n,k}, t^{B^j}_{n,k}$ with only local information
            update $\boldsymbol{\beta}^j_{n,k}$ with $\lambda^{(p,q)^{j-1}}_{n,k}$ from $\mathcal{C}_{p,q}$ where $(p,q) \neq (n,k)$
STEP3: The central node broadcasts $\boldsymbol{w}^{B^j}$ to all the clusters
        Each cluster $\mathcal{C}_{n,k}$ sends $\beta^{(p,q)^j}_{n,k}$ to $\mathcal{C}_{p,q}$
STEP 4: At each cluster $\mathcal{C}_{n,k}$:
            update $\boldsymbol{\lambda}^{n,k^j}, \mu^j_{n,k}, \kappa^j_{n,k}, \xi^j_{n,k}, \rho^j_{n,k}$ according to (3.54a)-(3.54e)
            Each cluster $\mathcal{C}_{n,k}$ sends $\lambda^{(n,k)^j}_{p,q}$ to $\mathcal{C}_{p,q}$
STEP 4: Set $j = j + 1$, and go to STEP 1
```

remark that, while the computation of the broadcast beamforming vector is performed at a processing unit with full CSI, the proposed implementation is more efficient when compared to the centralized approach thanks to the distributed optimization of unicast transmissions. Specifically, the optimization problems in (3.53b)-(3.53e) can be solved in parallel using distributed computing resources, and each of the problems is for a single scalar variable or for a vector of dimension M or NK.

3.7 Simulation Results

In this section, simulation results are presented to obtain insights into the performance comparison between LDM and TDM for the purpose of transmission of unicast and broadcast services in cellular systems. Unless stated otherwise, we consider a network comprised of macro-cells, each with $K = 3$ single-antenna active users. The radius of each cell is 500 m, and the users are located uniformly around the BS at a distance of 400 m. Each BS is equipped with $N_T = 3$ antennas. All channel vectors $\boldsymbol{h}_{i,n,k}$ are written as $\boldsymbol{h}_{i,n,k} = \left(10^{-\mathrm{PL}/10}\right)^{1/2} \tilde{\boldsymbol{h}}_{i,n,k}$, where the path loss exponent is modeled as $\mathrm{PL} = 148.1 + 37.6\log_{10}(d_{i,n,k})$, with $d_{i,n,k}$ denoting the distance (in kilometers) between the i-th BS and user (n, k), and $\tilde{\boldsymbol{h}}_{i,n,k}$ denoting an i.i.d. vector accounting for Rayleigh fading of unitary power. The noise variance is set to $\sigma^2_{n,k} = -134$ dBW for all users (n, k). Unless stated otherwise, we assume non-cooperative unicast transmission, i.e., each BS is informed only about the unicast data streams of its own users.

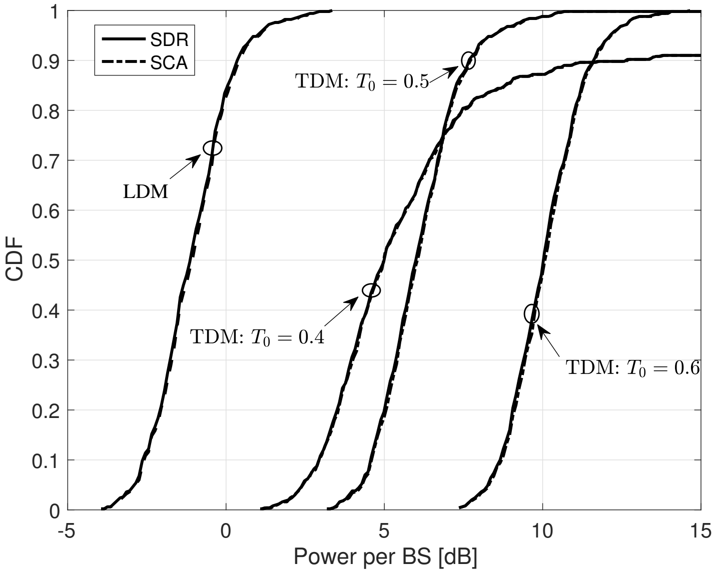


FIGURE 3.2: The CDF of power consumption per BS with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

3.7.1 Perfect CSI

Initially, we assume perfect CSI at all the BSs in the network. We plot the cumulative distribution function (CDF) of the transmission power per BS for LDM and TDM with $N = 3$ cells in Fig. 3.2. For the latter, we consider different values for the fraction of time T_0/T devoted to unicast traffic. Other values of T_0/T were seen not to improve the performance. The transmission power per BS is defined as the sum-power divided by the number of BSs. We observe that the curves may represent improper CDFs in the sense that their asymptotic values may be below 1. This gap accounts for the probability of the set of channel realizations in which the problem is found to be infeasible. We refer to the previous section for the assumed definition of infeasibility for SCA, whereas the standard definition is used for the convex problems in (3.27) and (3.30) solved using the S-procedure. Henceforth, we refer to the probability of an infeasible channel realization as the *outage probability*.

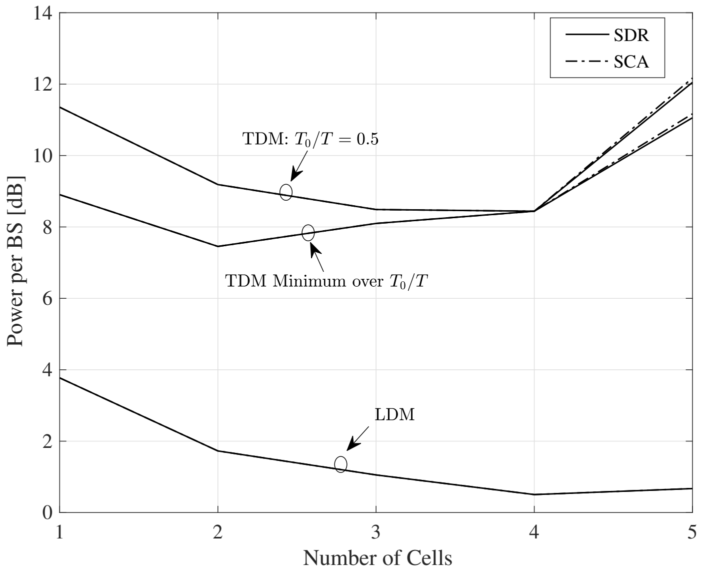


FIGURE 3.3: Power consumption per BS as a function of the number of cells with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

We can observe from Fig. 3.2 that LDM enables a significant reduction in the transmission power per BS as compared with TDM. In fact, even with an optimized choice of T_0/T, LDM can improve the 95th percentile of the transmitted power per BS by around 7 dB. Another observation is that SCA operates close to the lower bound set by SDR. Note also that LDM has a significantly lower outage probability than TDM. Finally, we remark that a large value of T_0/T is beneficial to obtain a lower outage probability in TDM, suggesting that the unicast constraints have more significant impact on the feasibility of the problem due to the need to cope with the mutual interference among unicast data streams. For the rest of this section, the displayed power values correspond to the 95th percentile of the corresponding CDF.

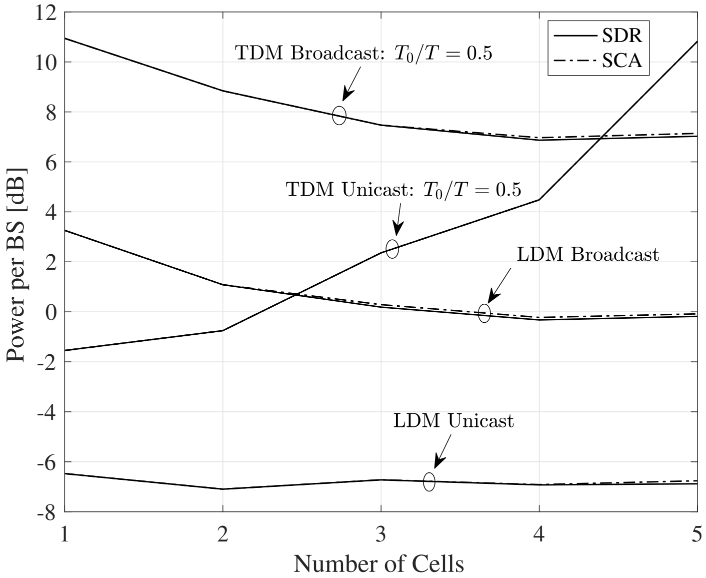


FIGURE 3.4: Power consumption per BS as a function of the number of cells with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

Next we study the impact of the number of cells on the performance of the system. To this end, Fig. 3.3 and Fig. 3.4 show the power per BS as a function of the number of cells. Specifically, Fig. 3.3 shows the overall power per BS, while Fig. 3.4 illustrates separately the power per BS used for the broadcast and unicast layers. Note that in Fig. 3.4 we fixed $T_0/T = 0.5$, while in Fig. 3.3 we also show the power obtained by selecting, for any number of cells, the value of T_0/T that minimizes the overall sum-power consumption (obtained by a line search). A key observation from Fig. 3.3 is that the power saving afforded by LDM increases with the number of cells. This gain can be attributed to the following two facts: (i) the optimal injection level is high (see Fig. 3.4), and hence the broadcast layer requires more power than unicast; and (ii) the performance of LDM is enhanced by the presence of more cells broadcasting the same message in the SFN, which increases the broadcast SINR and the broadcast layer can be more easily canceled by the users. The latter fact can be seen from Fig. 3.4, in which the required unicast power decreases with the number of cells when using LDM, unlike in TDM. Furthermore, the optimal IL of TDM decreases significantly, also suggesting that TDM is more sensitive to the mutual interference introduced by unicast data streams.

Fig. 3.5 compares the required power per BS for non-cooperative unicast transmission and for fully cooperative unicast transmission, i.e., clusters $\mathcal{C}_{n,k} = \{1, \ldots, N\}$ for all users (n, k). Here we consider a network comprised of $N = 3$ cells, and set $T_0/T = 0.8$ for TDM. From Fig. 3.5, it can be concluded that a higher unicast rate entails larger power savings by means of cooperative unicast transmission, especially for TDM. It is also worth mentioning that the LDM approach without BS cooperation in unicast transmission can even outperform the fully cooperative TDM approach in certain scenarios, e.g., when the rate for unicast messages is considerably lower than the broadcast rate.

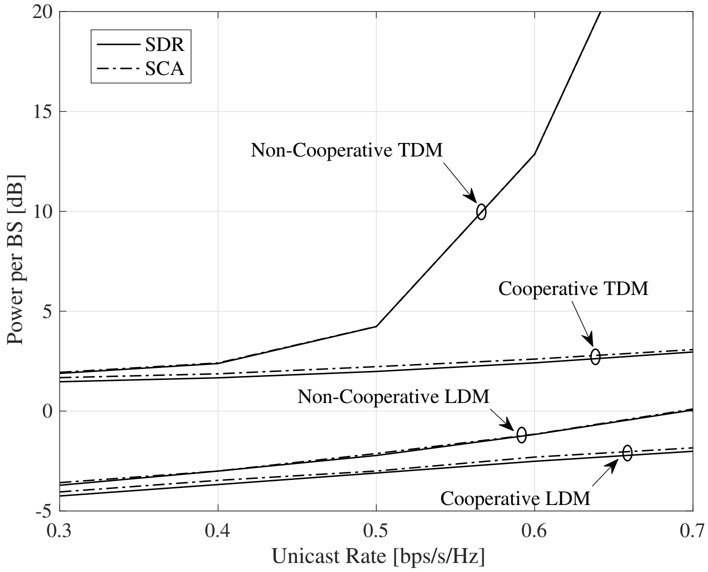


FIGURE 3.5: Power consumption per BS, separately for the unicast and broadcast signals, for values of unicast rate with R^B=2 bps/Hz for non-cooperative and fully cooperative schemes.

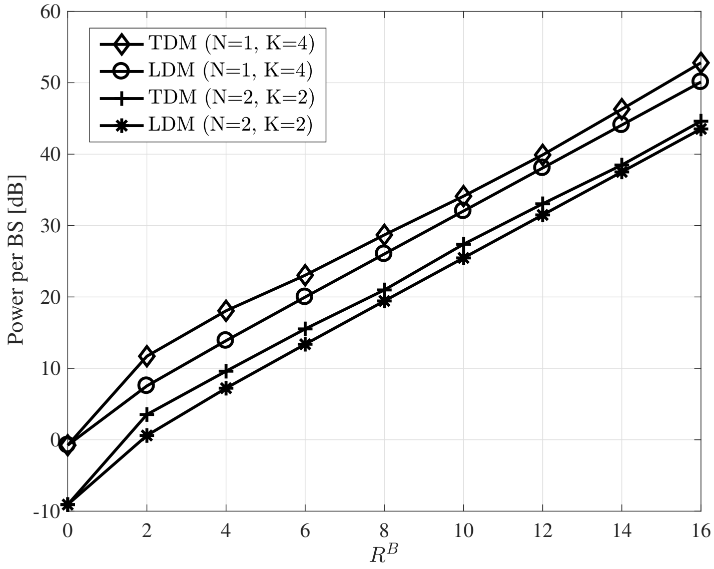


FIGURE 3.6: Power consumption per BS as a function of the broadcast rate with R^U=0.5 bps/Hz

We present the required power per BS of LDM and TDM as a function of the broadcast rate in Fig. 3.6. The unicast rate is set to $R^U = 0.5$ bps/Hz for all the users. The optimal time allocation T_0 in TDM is found by a line search with step size 0.05. When only unicast transmissions exist, i.e., $R^B = 0$, both LDM and TDM problems boil down to the multigroup multicast beamforming problem, and have the same performance in terms of power consumption. When the broadcast message and unicast messages are jointly transmitted, LDM always outperforms TDM in the considered range of broadcast rates. It is also concluded that the performance gain of LDM is larger with a higher user density.

Finally, we show the impact of the distance between users and the BS on the performance of TDM and LDM in Fig. 3.7. Here we consider the network consisting of $N = 5$ cells, each with a BS of $M = 5$ antennas. The scenarios with $K = 1$ and $K = 5$ users in each cell are simulated to observe the impact of user density on the performance of the system. It can be seen that LDM always outperforms TDM and has a power gain of around 5 dB in the considered range of distances. It is also observed that LDM can provide the same level of performance for cell-edge users as cell-center users in TDM.

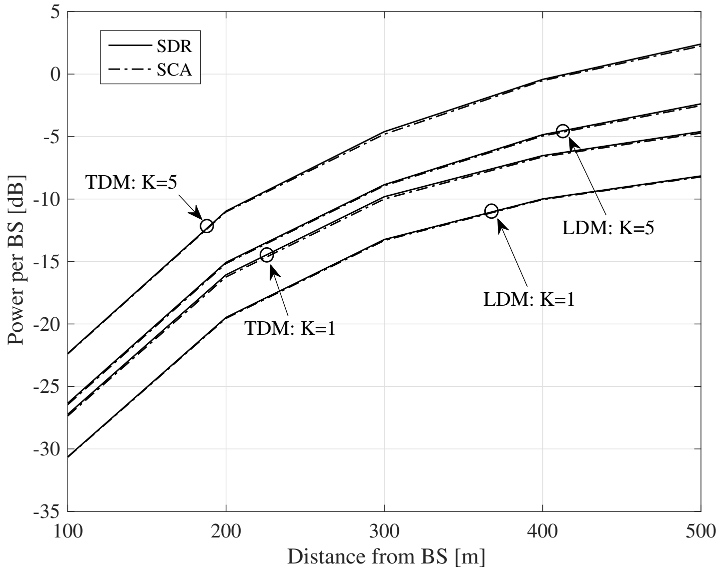


FIGURE 3.7: Power consumption per BS as a function of the distance between users and BSs with R^B=1 bps/Hz, R^U=0.5 bps/Hz, $N = 5$, and $M = 5$.

Next, we present the performance comparison between TDM and LDM considering two practical impairments, namely, imperfect channel coding, and imperfect CSI.

3.7.2 Imperfect Channel Coding

To account for the channel coding suboptimality, the SNR gap to capacity for broadcast and unicast layers is introduced as in [123]. Then, the SINR expressions of the broadcast signal are modified as follows:

$$\mathrm{SINR}_{n,k}^{B\text{-TDM}} = \lambda^B \frac{|\boldsymbol{h}_{n,k}^H \boldsymbol{w}^B|^2}{\sigma_{n,k}^2} \tag{3.55}$$

and

$$\mathrm{SINR}_{n,k}^{B\text{-LDM}} = \lambda^{B} \frac{|\boldsymbol{h}_{n,k}^{H} \boldsymbol{w}^{B}|^2}{\sum\limits_{(p,q)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^{U}|^2 + \sigma_{n,k}^2}, \tag{3.56}$$

as opposed to (3.12) and (3.16) for TDM and LDM, respectively, where λ^B is the SNR gap to capacity of the broadcast layer. Similarly, the SINR expressions for the unicast transmission in (3.13) and (3.17) are modified to

$$\begin{aligned} \mathrm{SINR}_{n,k}^{U\text{-LDM}} &= \mathrm{SINR}_{n,k}^{U\text{-TDM}} \\ &= \lambda^{U} \frac{|\boldsymbol{h}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^{U}|^2}{\sum\limits_{(p,q)\neq(n,k)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^{U}|^2 + \sigma_{n,k}^2}, \end{aligned} \tag{3.57}$$

where λ^U is the SNR gap to capacity for the unicast layer.

The outage probability versus the SNR gap, measured in dB, is presented in Fig. 3.8(a), while the corresponding transmission power per BS for LDM and TDM are depicted in Fig. 3.8(b). It can be observed that the outage probability of TDM significantly increases with the increased SNR gap from perfect channel coding, while the outage probability of LDM remains zero in our setting. In the state-of-the-art terrestrial broadcasting system where $\lambda^U = \lambda^B = -1$ dB are considered as the realistic values for the SNR gaps of the two layers [123], although TDM provides acceptable system service availability, the power consumption is found to be about 10 dB higher than LDM, as shown in Fig. 3.8(b). It can be further noticed that even when the SNR gap is 3 dB in LDM, the power consumption is still lower than TDM with ideal channel coding.

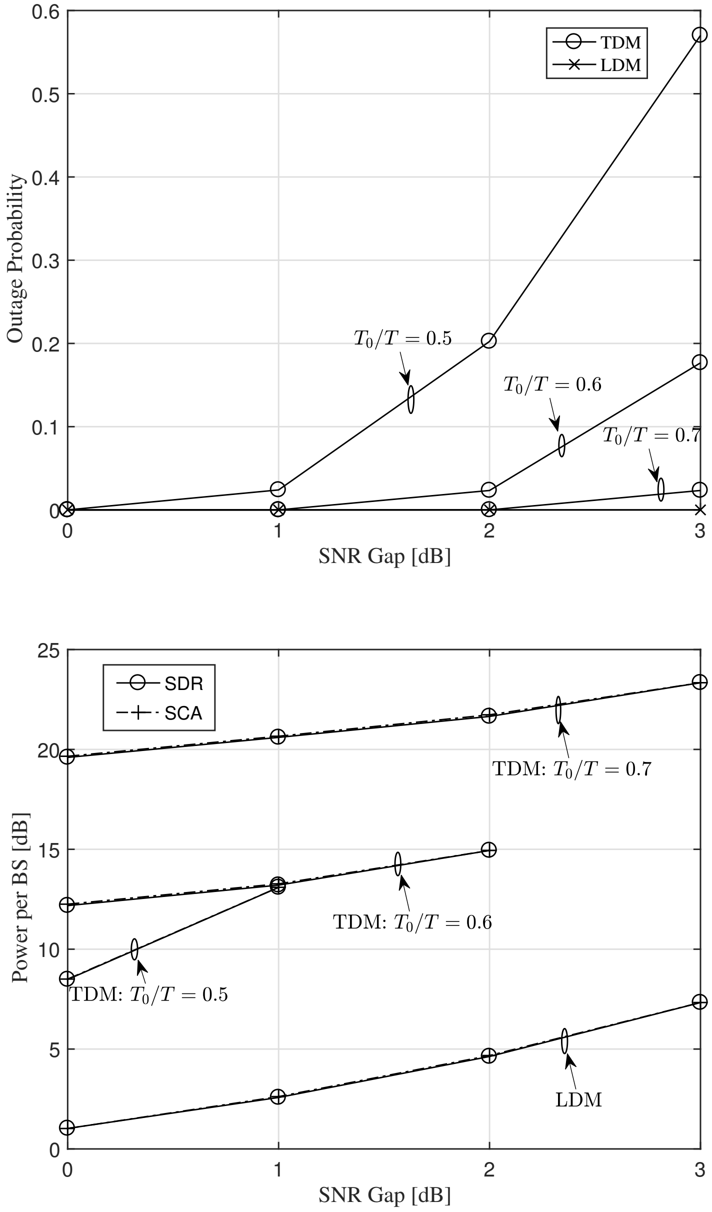


FIGURE 3.8: Outage probability and power consumption per BS for various values of SNR gap from ideal channel coding with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

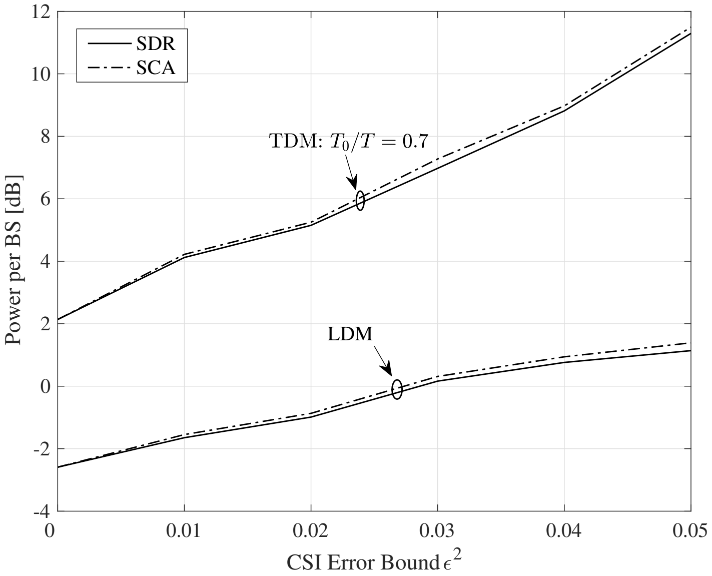


FIGURE 3.9: Power consumption per BS as a function of CSI error bound ϵ^2 with target rates R^B=1 bps/Hz and R^U=1 bps/Hz.

3.7.3 Imperfect CSI

We then demonstrate the effect of imperfect CSI on the performance. The channel error covariance matrix is set as $\boldsymbol{Q}_{i,n,k} = 1/\epsilon^2 \boldsymbol{I}_M$, where ϵ^2 is the common CSI error variance for all $\boldsymbol{e}_{i,n,k}$'s. It is observed in Fig. 3.9 that the power consumption per BS increases for both TDM and LDM systems, with the increase in CSI error variance ϵ^2. It is interesting to note that the minimum required power of TDM increases faster than that of LDM, indicating that TDM is more sensitive to CSI errors compared to LDM. This effect resembles the results encountered with higher unicast rate requirement and more users. In general, LDM outperforms TDM in terms not only of power consumption, but also of robustness against flexible system QoS targets and CSI imperfections.

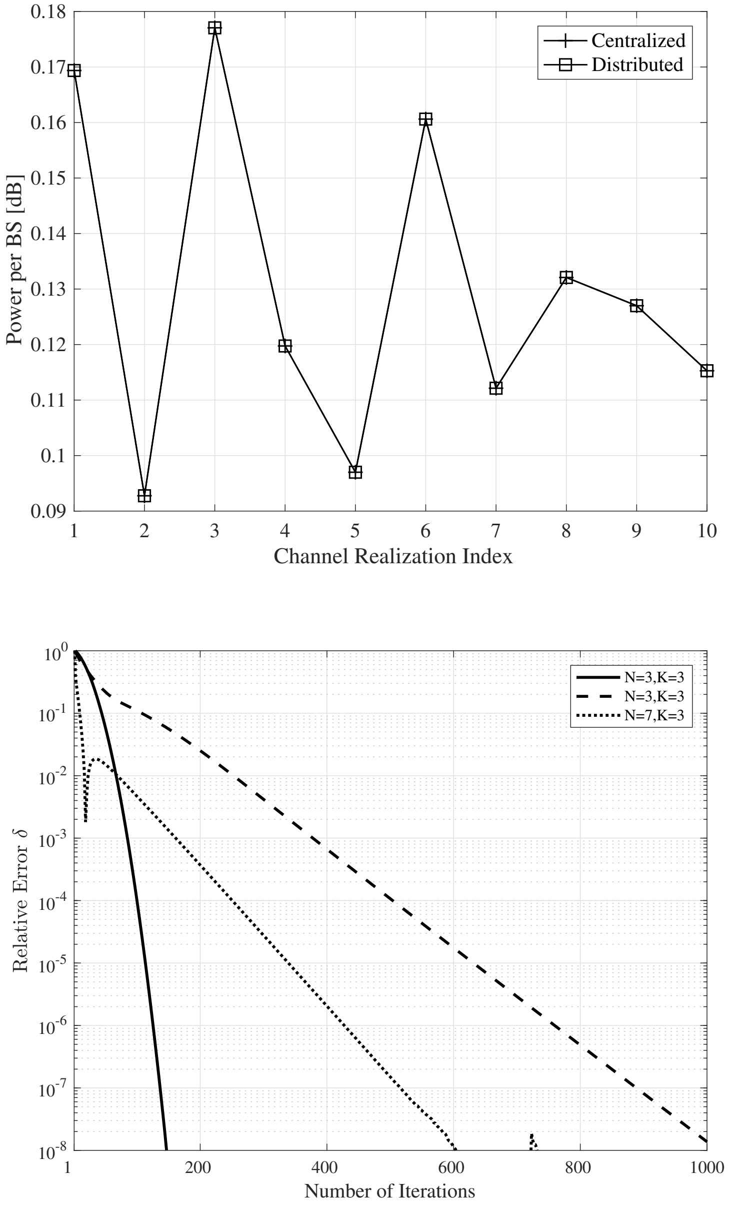


Figure 3.10: Convergence of the dual decomposition-based algorithm and relative error within dual ascent iterations for a given SCA subproblem.

3.7.4 Distributed Implementation

We first demonstrate that the distributed algorithm can converge to the same optimal solution as the centralized scheme, as shown in Fig. 3.10(a). The centralized solution was obtained by solving the optimization problem in (3.48) by CVX. The performance of the proposed dual decomposition-based distributed algorithm is studied in Fig. 3.10(b). The relative error at the j-th iteration of the algorithm in the ν-th SCA loop is computed by $\delta = |p^j - p^*|/p^*$, where p^j denotes the dual ascent solution at iteration j, and p^* denotes the optimal solution obtained by CVX in the best precision mode. The appropriate penalty parameters ρ are found empirically to observe fast convergence. Accordingly, Fig. 3.10(b) shows the convergence behavior of the distributed solution as a function of the number of iterations. It can be seen that for LDM, the algorithm converges fast to achieve an acceptable relative value, say $\delta = 10^{-4}$, within 500 iterations for a $N = 7$ cell network.

3.8 Conclusions

In this chapter, we have analyzed the performance gain of LDM over TDM/FDM as a potential NOMA approach for simultaneous transmission of broadcast and unicast messages over cellular networks. Joint beamforming design and power allocation was formulated as a sum-power minimization problem under distinct QoS constraints for the individual unicast messages and the common broadcast message. The resulting non-convex problem has been tackled by means of SCA and S-procedure, which provide upper and lower bounds on the optimal solution, respectively. Our numerical results have shown that the upper and lower bounds are tight, which indicates the near-optimality of the proposed solutions. We have also observed that LDM significantly improves the performance as compared to orthogonal transmission, and that it provides power savings for both the unicast and broadcast transmissions thanks to more efficient resource allocation. We have seen that the benefit of the increased bandwidth available for the broadcast layer outweighs the interference caused by unicast transmissions. In the case of imperfect CSI, we have noted that, while increased CSI error adversely affects both LDM and TDM, the increase in minimum required power as a function of the CSI error variance is much faster with TDM compared to LDM, indicating that LDM also provides better robustness against CSI uncertainties commonly experienced in real systems. A dual decomposition-based distributed solution has also been presented, which facilitates efficient distributed implementation for the LDM technique.

Chapter 4

Beamforming for Coded Content Delivery

4.1 Introduction

In this chapter, motivated by the results in [103] and [106], we consider a cache-aided MISO broadcast channel. Firstly, a general framework for cache-aided downlink beamforming is formulated, focusing on the minimum required transmit power for delivering the contents at a prescribed common rate. The resultant nonconvex optimization problem is tackled by successive convex approximation (SCA), which is guaranteed to converge to a stationary solution of the original nonconvex problem. As noted in [106], the beamforming design involves solving an optimization problem with exponentially increasing number of constraints with the number of coded messages each user decodes in each time slot. To limit the complexity, we propose a novel content delivery scheme, in which the coded subfiles, each targeted at a different subset of receivers, are delivered over multiple orthogonal time slots, while the number of coded messages each user decodes in each time slot can be flexibly adjusted. Unlike the simplified scheme proposed in [106], the proposed scheme does not limit the number of users served in each time slot, but directly limits the number of messages each user decodes, and hence, the complexity of the decoder. We propose a greedy algorithm that decides the multicast messages to be delivered at each time slot, and the number of time slots. We then consider the joint design of the beamforming vectors together with the content delivery scheme with a constraint on the maximum

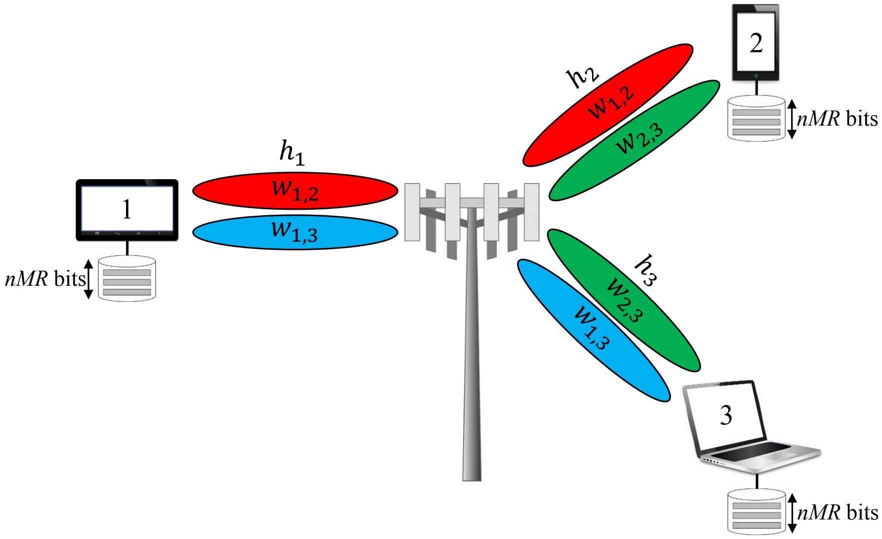


FIGURE 4.1: Illustration of a cache-aided MISO channel with $K = 3$ users. Multi-antenna BS employs multicast beamforming to deliver the missing parts of users' requests.

number of messages each user can decode in any time slot. We formulate this joint optimization as a power minimization problem with sparsity constraints, and solve it via SCA to obtain a stationary solution. Our numerical results show that the proposed greedy scheme has a minimal performance gap to that of the optimization-based delivery scheme, and provide significant gains over the one proposed in [106] in terms of transmit power, particularly in the high rate/high signal-to-noise ratio (SNR) regime.

4.2 System Model

We consider downlink transmission within a single cell, where a BS equipped with N_T antennas serves K single-antenna cache-equipped users, as illustrated in Fig. 4.1. We consider a library of N files, denoted by $\boldsymbol{V} \triangleq (V_1, \cdots, V_N)$, each distributed uniformly over the set $\left[2^{nR}\right]$[1], available at the BS, where R and n

[1] For any positive real X, we define $[X]$ as the set of positive integers less than or equal to X.

represent the rate of each file and the blocklength, respectively. Each user is equipped with a local cache that can store up to M files, and the corresponding *caching factor*, t, is defined as the ratio of the total cache capacity across all the receivers to the library size, $t \triangleq MK/N$.

Contents are placed at users' caches during off-peak periods without any prior information on user requests. Caching function for user k is denoted by $\phi_k^{(n)}$: $\left[2^{nR}\right]^N \rightarrow \left[2^{nMR}\right]$, which maps the library to the cache contents Z_k at user k, i.e., $Z_k = \phi_k(\boldsymbol{V})$, $k \in [K]$, without the knowledge of channel state information (CSI). Once the users reveal their demands $\boldsymbol{d} \triangleq (d_1, \ldots, d_K)$, where $d_k \in [N], \forall k \in [K]$, signal $\boldsymbol{x} \in \mathbb{C}^{N_T \times n}$ is transmitted, where $\boldsymbol{x} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n]$, and $\boldsymbol{x}_i \in \mathbb{C}^{N_T \times 1}$ is the channel input vector at time i, $i = 1, \cdots, n$. An average power constraint P is imposed on each channel input $\boldsymbol{x}$. User k receives

$$\boldsymbol{y}_k = \boldsymbol{h}_k^H \boldsymbol{x} + \boldsymbol{n}_k, \tag{4.1}$$

where $\boldsymbol{h}_k \in \mathbb{C}^{N_T \times 1}$ is the channel vector from the BS to the k-th user, and $\boldsymbol{n}_k \in \mathbb{C}^{1 \times n}$ is the additive white Gaussian noise at user k with each entry independent and identically (i.i.d.) distributed according to $\mathcal{CN}(0, \sigma_k^2)$, $k \in [K]$. We assume that the CSI is perfectly known to the BS and the receivers in the delivery phase. Hence, the encoding function at the BS, $\psi^{(n)} : \left[2^{nR}\right]^N \times [N]^K \times \mathbb{C}^{N_T \times K} \rightarrow \mathbb{C}^{N_T \times n}$, maps the library, the demand vector, and the CSI to the channel input vector. We note here that, while the channel encoding function $\psi^{(n)}$ depends on the demand vector and the CSI, caching functions $\phi_k^{(n)}$ depend only on the library. After receiving $\boldsymbol{y}_k$, user k reconstructs $\hat{V}_{d_k}$ using its local cache content Z_k, channel vector $\boldsymbol{h}_k$, and demand vector $\boldsymbol{d}$ through function $\mu_k^{(n)} : \mathbb{C}^n \times \left[2^{nMR}\right] \times \mathbb{C}^{N_T \times 1} \times [N]^K \rightarrow \left[2^{nR}\right]$, i.e., $\hat{V}_k = \mu_k^{(n)} (\boldsymbol{y}_k, Z_k, \boldsymbol{h}_k, \boldsymbol{d})$, $k \in [K]$. The probability of error is defined as $\mathrm{P_e} \triangleq \max_{\boldsymbol{d}} \max_{k \in [K]} \Pr\{V_{d_k} \neq \hat{V}_k\}$. An (R, M, P) tuple is *achievable* if there exist a sequence of caching functions $\phi_1^{(n)}, \ldots, \phi_K^{(n)}$, encoding function $\psi^{(n)}$, and decoding functions $\mu_1^{(n)}, \ldots, \mu_K^{(n)}$, such that $\mathrm{P_e} \rightarrow 0$ as $n \rightarrow \infty$. For file rate

R and cache size M, our goal is to characterize

$$P^* (R, M) \triangleq \inf \{P : (R, M, P) \, \text{is achievable}\} , \tag{4.2}$$

which characterizes the minimum required transmit power that guarantees the reliable delivery of any demand vector.

4.3 An Achievable Delivery Scheme

In this section, we present a multi-antenna transmission scheme with coded caching, where the cache placement and coded content generation follows [21], while beamforming is employed at the BS to multicast coded subfiles to receivers.

4.3.1 Placement and Delivery Schemes

For a caching factor $t \in \{1, \dots, K-1\}$, we represent t-element subsets of $[K]$ by $\mathcal{G}_1^t, \dots, \mathcal{G}_{\binom{K}{t}}^t$. File V_i, $i \in [N]$, is divided equally into $\binom{K}{t}$ disjoint subfiles $V_{i,\mathcal{G}_1^t}, \dots, V_{i,\mathcal{G}_{\binom{K}{t}}^t}$, each consisting of $n\frac{R}{\binom{K}{t}}$ bits. User k, $k \in [K]$, caches subfile $V_{i,\mathcal{G}_j^t}$, if $k \in \mathcal{G}_j^t$, $\forall j \in [\binom{K}{t}]$. The cache content of user k is then given by $\bigcup_{i \in [N]} \bigcup_{j \in [\binom{K}{t}] : k \in \mathcal{G}_j^t} V_{i,\mathcal{G}_j^t}$.

During the *delivery phase*, for any demand combination $\boldsymbol{d}$, we aim to deliver the coded message

$$s_{\mathcal{G}_j^{t+1}} \triangleq \bigoplus\nolimits_{k \in \mathcal{G}_j^{t+1}} V_{d_k, \mathcal{G}_j^{t+1} \backslash \{k\}} \tag{4.3}$$

to all the users in set $\mathcal{G}_j^{t+1}$, for $j \in [\binom{K}{t+1}]$. Observe that, after receiving $s_{\mathcal{G}_j^{t+1}}$, each user $k \in \mathcal{G}_j^{t+1}$ can recover subfile $V_{d_k, \mathcal{G}_j^{t+1} \backslash \{k\}}$ having access to $V_{d_l, \mathcal{G}_j^{t+1} \backslash \{l\}}$, $\forall l \in \mathcal{G}_j^{t+1} \backslash \{k\}$.

We define $\mathcal{S} \triangleq \{\mathcal{G}_1^{t+1}, \dots, \mathcal{G}_{\binom{K}{t+1}}^{t+1}\}$ as the set of all the multicast messages, with each message $\mathcal{T} \in \mathcal{S}$ represented by the set of users it is targeting, and let

$\mathcal{S}_k \subset \mathcal{S}$ denote the subset of messages targeting user k. We have $|\mathcal{S}| = \binom{K}{t+1}$ and $|\mathcal{S}_k| = \binom{K-1}{t}$.

The following settings will be used to explain the proposed scheme:

Setting 1: Let $N = 5,\ K = 5,\ M = 1$. We have $t = \frac{MK}{N} = 1$. Each file is split into $\binom{K}{t} = 5$ disjoint subfiles of the same size, where we represent file i, $i \in [N]$, as

$$V_i = \left\{V_{i,\{1\}}, V_{i,\{2\}}, V_{i,\{3\}} V_{i,\{4\}}, V_{i,\{5\}}\right\}. \tag{4.4}$$

The cache content of user k is $Z_k = \cup_{n\in[N]} V_{n,\{k\}},\ k \in [K]$, which satisfies the cache capacity constraint. For a demand combination $\boldsymbol{d}$, all user demands can be fulfilled by delivering the following $\binom{K}{t+1} = 10$ subfiles:

$$\begin{aligned}
&s_{\{1,2\}} = V_{d_1,\{2\}} \oplus V_{d_2,\{1\}},\ s_{\{1,3\}} = V_{d_1,\{3\}} \oplus V_{d_3,\{1\}},\\
&s_{\{1,4\}} = V_{d_1,\{4\}} \oplus V_{d_4,\{1\}},\ s_{\{1,5\}} = V_{d_1,\{5\}} \oplus V_{d_5,\{1\}},\\
&s_{\{2,3\}} = V_{d_2,\{3\}} \oplus V_{d_3,\{2\}},\ s_{\{2,4\}} = V_{d_2,\{4\}} \oplus V_{d_4,\{2\}},\\
&s_{\{2,5\}} = V_{d_2,\{5\}} \oplus V_{d_5,\{2\}},\ s_{\{3,4\}} = V_{d_3,\{4\}} \oplus V_{d_4,\{3\}},\\
&s_{\{3,5\}} = V_{d_3,\{5\}} \oplus V_{d_5,\{3\}},\ s_{\{4,5\}} = V_{d_4,\{5\}} \oplus V_{d_5,\{4\}}.
\end{aligned}$$

Setting 2: Let $N = 4,\ K = 4,\ M = 1$. We have $t = \frac{MK}{N} = 1$. Each file is split into $\binom{K}{t} = 4$ disjoint subfiles of the same size. For a demand combination $\boldsymbol{d}$, all user demands can be fulfilled by delivering the following $\binom{K}{t+1} = 6$ subfiles:

$$s_{\{1,2\}},\ s_{\{1,3\}},\ s_{\{1,4\}},\ s_{\{2,3\}},\ s_{\{2,4\}},\ s_{\{3,4\}}. \tag{4.5}$$

Note that the message $s_{\mathcal{T}}$ is intended for users in set $\mathcal{T}$, but interferes with users in set $[K]\backslash\mathcal{T}$. Moreover, for any demand combination $\boldsymbol{d}$, all the users are required to decode the same number of messages, which is $\binom{K-1}{t}$.

4.3.2 Multi-Antenna Transmission Scheme

The delivery of the coded messages in set $\mathcal{S}$ to their respective receivers is a multi-antenna multi-message multicasting problem. Before introducing our low-complexity scheme in the next section, we present here a general transmission strategy based on message-splitting and time-division transmission. The messages in $\mathcal{S}$ can be transmitted over B orthogonal time slots, the i-th of which is of blocklength $n_i, i \in [B]$, where $\sum_{i=1}^{B} n_i = n$. The transmitted signal $\boldsymbol{x}(i) \triangleq [\boldsymbol{x}_{\sum_{j=1}^{i-1} n_j+1} \cdots \boldsymbol{x}_{\sum_{j=1}^{i} n_j}]$ at time slot $i \in [B]$ is given by

$$\boldsymbol{x}(i) = \sum\nolimits_{\mathcal{T}\in\mathcal{S}} \boldsymbol{w}_{\mathcal{T}}(i)\boldsymbol{s}_{\mathcal{T}}(i), \tag{4.6}$$

where $\boldsymbol{s}_{\mathcal{T}}(i) \in \mathbb{C}^{1\times n_i}$ is the unit power complex Gaussian signal of block length n_i, modulated from the corresponding message $s_{\mathcal{T}}$ in (4.3), intended for the users in set $\mathcal{T}$, transmitted in time slot i, encoded by the beamforming vector $\boldsymbol{w}_{\mathcal{T}}(i) \in \mathbb{C}^{N_T\times 1}$.

The received signal at user k in time slot i is

$$\boldsymbol{y}_k(i) = \underbrace{\boldsymbol{h}_k^H \sum_{\mathcal{T}\in\mathcal{S}_k} \boldsymbol{w}_{\mathcal{T}}(i)\boldsymbol{s}_{\mathcal{T}}(i)}_{\text{desired messages}} + \underbrace{\boldsymbol{h}_k^H \sum_{\mathcal{I}\in\mathcal{S}_k^C} \boldsymbol{w}_{\mathcal{I}}(i)\boldsymbol{s}_{\mathcal{I}}(i)}_{\text{interference}} + \boldsymbol{n}_k(i), \tag{4.7}$$

where $\mathcal{S}_k^C$ is the complement of set $\mathcal{S}_k$ in $\mathcal{S}$. Let $\Pi_{\mathcal{S}_k}$ denote the collection of all non-empty subsets of $\mathcal{S}_k$, with each element of $\Pi_{\mathcal{S}_k}$ denoted by $\pi_{\mathcal{S}_k}^{j}$, $j \in [2^{\binom{K-1}{t}} - 1]$. We denote $\mathcal{S}(i) \subset \mathcal{S}$ as the subset of messages transmitted in time slot i, i.e., $\mathcal{T} \in \mathcal{S}(i)$ if $\boldsymbol{w}_{\mathcal{T}}(i) \neq \mathbf{0}$.

Note that each user may receive more than one message in each transmission slot. From the capacity region of the associated Gaussian multiple access channel, following conditions must be satisfied for successful decoding of all the intended messages at user k, $k \in [K]$, at time slot i:

$$\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} R^{\mathcal{T}}(i) \leq \frac{n_i}{n}\log_2\left(1+\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} \gamma^{\mathcal{T}}_k(i)\right),\ \forall \pi^j_{\mathcal{S}_k} \in \Pi_{\mathcal{S}_k}, \tag{4.8}$$

where $R^{\mathcal{T}}(i)$ is the rate of message $\boldsymbol{s}_{\mathcal{T}}(i)$, and $\gamma^{\mathcal{T}}_k(i)$ is the received signal-to-interference-plus-noise ratio (SINR) of message $s_{\mathcal{T}}(i)$ at user k at time slot i, given by

$$\gamma^{\mathcal{T}}_k(i) \triangleq \frac{|\boldsymbol{h}^H_k \boldsymbol{w}_{\mathcal{T}}(i)|^2}{\sum_{\mathcal{I}\in\mathcal{S}^C_k} |\boldsymbol{h}^H_k \boldsymbol{w}_{\mathcal{I}}(i)|^2 + \sigma^2_k}, \tag{4.9}$$

for any $\mathcal{T} \ni k$, or equivalently, any $\mathcal{T} \in \mathcal{S}_k$. The rate of message $\mathcal{T}$ is the sum of the rate of submessages $s_{\mathcal{T}}(i)$, and must satisfy

$$\textstyle\sum_{i=1}^{B} R^{\mathcal{T}}(i) \geq \frac{R}{\binom{K}{t}},\ \forall \mathcal{T}. \tag{4.10}$$

Note that this scheme is quite flexible; each multicast message can be split into B messages and transmitted over B time slots. It can be specialized to different content delivery schemes by specifying the subset of transmitted subfiles in each time slot and the blocklength of each time slot, i.e., $\{\mathcal{S}(i)\}_{i=1}^B$ and $\{n_i\}$. Let

$$v_{\mathcal{T}}(i) = \begin{cases} 1 & \text{if } \ \mathcal{T} \in \mathcal{S}(i) \\ 0 & \text{if } \ \mathcal{T} \notin \mathcal{S}(i) \end{cases} \tag{4.11}$$

be the indicator function specifying whether message $\mathcal{T}$ is transmitted at time slot i or not. Note that $\|\boldsymbol{v}_{\mathcal{T}}\|_1 \geq 1$ is required to fulfill users' demands, where $\boldsymbol{v}_{\mathcal{T}} \triangleq [v_{\mathcal{T}}(1), \cdots, v_{\mathcal{T}}(B)]$. It is readily seen that $v_{\mathcal{T}}(i)$ can be inferred by the corresponding beamforming vector $w_{\mathcal{T}}(i)$, or equivalently, by the message rate $R_{\mathcal{T}}(i)$.

4.3.3 Transmit Power Minimization

For any given delivery scheme specified by $v_{\mathcal{T}}(i)$ and n_i, $\forall i \in [B], \forall \mathcal{T} \in \mathcal{S}$, the associated minimum required transmit power problem is obtained as follows:

$$P \triangleq \min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\},\{R^{\mathcal{T}}(i)\}} \sum_{\mathcal{T}\in\mathcal{S}} \sum_{i=1}^{B} \frac{n_i}{n} \|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.12a}$$

$$\text{s.t.} \quad \sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} R^{\mathcal{T}}(i) \leq \frac{n_i}{n} \log_2\left(1 + \sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} \gamma_k^{\mathcal{T}}(i)\right),\ \forall \pi^j_{\mathcal{S}_k} \in \Pi_{\mathcal{S}_k},\ i \in [B], \tag{4.12b}$$

$$\textstyle\sum_{i=1}^{B} R^{\mathcal{T}}(i) \geq \frac{R}{\binom{K}{t}},\ \forall \mathcal{T}, \tag{4.12c}$$

$$R^{\mathcal{T}}(i) = 0,\ \forall v_{\mathcal{T}}(i) = 0,\ \forall i, \tag{4.12d}$$

where $\gamma_k^{\mathcal{T}}(i)$ is defined in (4.9). Here, constraints in (4.12b) guarantee that the rates of the messages targeting each user in each time slot are within the capacity region, constraints in (4.12c) ensure that sufficient information is delivered for each coded subfile over B time slots, while (4.12d) represents the specific content delivery scheme.

Note that the problem in (4.12) is a generalization of various well-known NP-hard problems depending on the specific content delivery scheme. For $B = |\mathcal{S}|$ with $|\mathcal{S}(i)| = 1,\ \forall i$, the problem boils down to a series of standard multicast beamforming problems, where a common message is broadcast to a different subset of $t+1$ users in each time slot [63]. When $|\mathcal{S}(i)| > 1$, $\mathcal{T} \bigcap \mathcal{T}' = \emptyset$ if $\mathcal{T} \neq \mathcal{T}' \in \mathcal{S}(i)$, $\forall i$, and $\mathcal{S}(i) \bigcap \mathcal{S}(j) = \emptyset$ if $i \neq j$, we need to solve the conventional multigroup multicast beamforming problem at each time slot [65]. It can be seen from the problem formulation in (4.12) that the content delivery scheme specified by $v_{\mathcal{T}}(i)$ and n_i affects the minimum required power. As described in [108] and [136], a straightforward incorporation of the coded delivery scheme to the multi-antenna setting can be by transmitting a single coded message in each time slot. However, this does not fully exploit the spatial multiplexing gain provided by the multiple antennas, and results in poor DoF performance in the high SNR regime. Another approach studied in [137] is to select the coded messages targeting non-overlapping user groups in parallel. Obviously, when considering efficient implementations of coded caching within wireless networks, the content delivery scheme is an important factor on the system performance and needs to be carefully designed.

We remark here that, even when the delivery scheme is specified, the problem in (4.12) is computationally intractable due to the non-convex constraints in (4.12b). However, it is noted that the constrains are in the form of difference of convex functions, which can be approximated by linearizing the concave functions, resulting in a convex problem that can be solved via SCA techniques. To see this, we first rewrite the problem in (4.12) as

$$\min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\},\{R^{\mathcal{T}}(i),\{\eta_{\pi^j_{\mathcal{S}_k}}(i)\}} \sum_{i=1}^{B}\sum_{\mathcal{T}\in\mathcal{S}} \frac{n_i}{n}\|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.13a}$$

$$\text{s.t.} \sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} R^{\mathcal{T}}(i) \leq \frac{n_i}{n}\log_2(1+\eta_{\pi^j_{\mathcal{S}_k}}(i)),\ \forall\pi^j_{\mathcal{S}_k}\in\Pi_{\mathcal{S}_k},\forall k,i, \tag{4.13b}$$

$$\textstyle\sum_{\mathcal{I}\in\mathcal{S}_k^C}|\boldsymbol{h}_k^H\boldsymbol{w}_{\mathcal{I}}(i)|^2 - \frac{\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}}|\boldsymbol{h}_k^H\boldsymbol{w}_{\mathcal{T}}(i)|^2}{\eta_{\pi^j_{\mathcal{S}_k}}(i)} + \sigma_k^2 \leq 0,\ \forall\pi^j_{\mathcal{S}_k}\in\Pi_{\mathcal{S}_k},\forall k,i, \tag{4.13c}$$

$$\text{(4.12c) and (4.12d)}, \tag{4.13d}$$

where $\eta_{\pi^j_{\mathcal{S}_k}}(i)$ are auxiliary variables. The constraint in (4.13c) is the difference of convex function, since $\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}}|\boldsymbol{h}_k^H\boldsymbol{w}_{\mathcal{T}}(i)|^2/\eta_{\pi^j_{\mathcal{S}_k}}(i)$ is the sum of quadratic-over-linear functions of $\boldsymbol{w}_{\mathcal{T}}(i)$ and $\eta_{\pi^j_{\mathcal{S}_k}}(i)$. Therefore, a sequence of convex subproblems can be solved iteratively to approximately tackle this convex-concave problem

[135], with the subproblem in the $(\nu+1)$-th iteration given by

$$\min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\},\{R^{\mathcal{T}}(i),\{\eta_{\pi^j_{\mathcal{S}_k}}(i)\}} \quad \sum_{i=1}^{B}\sum_{\mathcal{T}\in\mathcal{S}} \frac{n_i}{n}\|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.14a}$$

$$\text{s.t.} \quad \sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} R^{\mathcal{T}}(i) \leq \frac{n_i}{n}\log_2(1+\eta_{\pi^j_{\mathcal{S}_k}}(i)),\ \forall\pi^j_{\mathcal{S}_k}\in\Pi_{\mathcal{S}_k},\forall k,i, \tag{4.14b}$$

$$\sum\nolimits_{\mathcal{I}\in\mathcal{S}_k^C}|\boldsymbol{h}_k^H\boldsymbol{w}_{\mathcal{I}}(i)|^2 + \frac{\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}}|\boldsymbol{h}_k^H\boldsymbol{w}_{\mathcal{T}}^{\nu}(i)|^2}{\eta^{\nu^2}_{\pi^j_{\mathcal{S}_k}}(i)}\eta_{\pi^j_{\mathcal{S}_k}}(i)$$
$$-\frac{2\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}}\boldsymbol{w}_{\mathcal{T}}^{\nu^H}(i)\boldsymbol{h}_k\boldsymbol{h}_k^H\boldsymbol{w}_{\mathcal{T}}(i)}{\eta^{\nu}_{\pi^j_{\mathcal{S}_k}}(i)} + \sigma_k^2 \leq 0,\ \forall\pi^j_{\mathcal{S}_k}\in\Pi_{\mathcal{S}_k},\forall k,i, \tag{4.14c}$$

$$\text{(4.12c) and (4.12d)}, \tag{4.14d}$$

given the solution of $\boldsymbol{w}^{\nu}_{\mathcal{T}}(i)$, $R^{\mathcal{T}^{\nu}}(i)$, and $\eta^{\nu}_{\pi^j_{\mathcal{S}_k}}(i)$ obtained in the ν-th SCA iteration. Each of the convex subproblems can be efficiently solved with standard interior-point algorithms or off-the-shelf solvers, and the SCA approach is guaranteed to converge to a stationary solution of the original problem in (4.12) [124]. Details of the SCA algorithm are outlined in Table. 4.1.

An initial point in the feasible set of problem (4.12) is required to initialize the SCA algorithm. We first observe that for any feasible target rates $\{R^{\mathcal{T}}(i)|\forall\mathcal{T}\in\mathcal{S}\}_{i=1}^{B}$ that satisfy the constraints in (4.12c) and (4.12d), the problem in (4.12) can be decoupled and decomposed into B parallel subproblems, each for a distinct time slot $i\in[B]$, given by

$$\{\boldsymbol{w}^*_{\mathcal{T}}(i)\}_{\mathcal{T}\in\mathcal{S}(i)} = \arg\min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\}} \sum_{\mathcal{T}\in\mathcal{S}(i)} \|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.15a}$$

$$\text{s.t.} \quad \sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} R^{\mathcal{T}}(i) \leq \frac{n_i}{n}\log_2\left(1+\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}}\gamma_k^{\mathcal{T}}(i)\right),\ \forall\pi^j_{\mathcal{S}_k}\in\Pi_{\mathcal{S}_k}, \tag{4.15b}$$

$$\gamma_k^{\mathcal{T}}(i) = \frac{|\boldsymbol{h}_k^H \boldsymbol{w}_{\mathcal{T}}(i)|^2}{\sum_{\mathcal{I}\in\mathcal{S}_k^C} |\boldsymbol{h}_k^H \boldsymbol{w}_{\mathcal{I}}(i)|^2 + \sigma_k^2}, \tag{4.15c}$$

$$\|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 = 0 \text{ for } \forall v_{\mathcal{T}}(i) = 0, \tag{4.15d}$$

which is nonconvex. Nevertheless, it can be transformed into a semidefinite programming problem by introducing $\boldsymbol{W}_{\mathcal{T}}(i) \triangleq \boldsymbol{w}_{\mathcal{T}}(i)\boldsymbol{w}_{\mathcal{T}}^H(i)$ and dropping the rank-1 constraints on $\boldsymbol{W}_{\mathcal{T}}(i)$, which is given by

$$\{\boldsymbol{W}_{\mathcal{T}}^*(i)\}_{\mathcal{T}\in\mathcal{S}(i)} = \underset{\{\boldsymbol{W}_{\mathcal{T}}(i)\}}{\arg\min} \sum_{\mathcal{T}\in\mathcal{S}} \mathrm{Tr}\{\boldsymbol{W}_{\mathcal{T}}(i)\} \tag{4.16a}$$

$$\text{s.t.} \left(2^{\frac{n}{n_i}\sum_{\mathcal{T}\in\pi_{\mathcal{S}_k}^j} R^{\mathcal{T}}(i)} - 1\right)\left(\sum_{\mathcal{I}\in\mathcal{S}_k^C} \mathrm{Tr}\{\boldsymbol{H}_k \boldsymbol{W}_{\mathcal{I}}(i)\} + \sigma_k^2\right) - \sum_{\mathcal{T}\in\pi_{\mathcal{S}_k}^j} \mathrm{Tr}\{\boldsymbol{H}_k \boldsymbol{W}_{\mathcal{T}}(i)\} \le 0,\ \forall \pi_{\mathcal{S}_k}^j \in \Pi_{\mathcal{S}_k}, \forall k, \tag{4.16b}$$

$$\boldsymbol{W}_{\mathcal{T}}(i) \succeq 0, \forall v_{\mathcal{T}}(i) \neq 0, \tag{4.16c}$$

$$\boldsymbol{W}_{\mathcal{T}}(i) = 0, \forall v_{\mathcal{T}}(i) = 0, \tag{4.16d}$$

and can be efficiently solved with standard interior-point algorithms. However, the solution obtained with semidefinite relaxation is not necessarily rank-1. If the obtained $\boldsymbol{W}_{\mathcal{T}}(i)$'s are all rank-1, then the optimal solution of (4.15) can be readily recovered from $\boldsymbol{W}_{\mathcal{T}}(i)$. Otherwise, Gaussian randomization can be adopted to obtain a feasible approximation to the optimal solution of (4.15). Note that the solution given by (4.15) is an upper bound on the minimum required power in (4.12) as the rates $\{R^{\mathcal{T}}(i)|\forall\mathcal{T}\in\mathcal{S}\}_{i=1}^B$ are not optimized, which hence can serve as an initial point in the successive convex approximation algorithm to obtain a tighter upper bound on the problem in (4.12).

Table 4.1: SCA Algorithm for the Multicast Beamforming Problem with a Given Coded Delivery Scheme

```
STEP 0: Set $\nu = 1$. Set a step size $\mu$.
        Initialize $\boldsymbol{w}^{\nu}_{\mathcal{T}}(i)$, $R^{\nu}_{\mathcal{T}}(i)$, and $\eta^{\nu}_{\pi^{j}_{\mathcal{S}_k}}(i)$ with feasible values
STEP 1: If a stopping criterion is satisfied, then STOP
STEP 2: Solve the optimization problem in (4.14)
STEP 3: Update $\boldsymbol{w}^{\nu+1}_{\mathcal{T}}(i) = \boldsymbol{w}^{\nu}_{\mathcal{T}}(i) + \mu\,(\boldsymbol{w}_{\mathcal{T}}(i) - \boldsymbol{w}^{\nu}_{\mathcal{T}}(i))$,
        $R^{\nu+1}_{\mathcal{T}}(i) = R^{\nu}_{\mathcal{T}}(i) + \mu\,(R_{\mathcal{T}}(i) - R^{\nu}_{\mathcal{T}}(i))$,
        $\eta^{\nu+1}_{\pi^{j}_{\mathcal{S}_k}}(i) = \eta_{\pi^{j}_{\mathcal{S}_k}}(i) + \mu\left(\eta_{\pi^{j}_{\mathcal{S}_k}}(i) - \eta^{\nu}_{\pi^{j}_{\mathcal{S}_k}}(i)\right)$,
STEP 4: Set $\nu = \nu + 1$, and go to STEP 1
```

4.4 A Low-Complexity Design

In this section, we propose a low-complexity content delivery scheme with the flexibility to adjust the number of coded messages intended for each user at each time slot. Observing that if a set $\mathcal{S}(i) = \{\mathcal{T} | v_{\mathcal{T}}(i) = 1\}$ of messages are transmitted in time slot i, $c_k(i) \triangleq |\mathcal{S}(i) \bigcap \mathcal{S}_k|$ messages are transmitted to user k, which results in $2^{c_k(i)} - 1$ constraints only for user k in time slot i in problem (4.12). Computational complexity of problem (4.12) increases drastically with the number of constraints, rendering the numerical optimization problem practically infeasible. In addition, a multi-user detection scheme needs to be employed at the users, whose complexity also increases with $c_k(i)$.

A low complexity scheme is proposed in [106] by limiting the number of users to be served in each time slot, thereby indirectly reducing the number of coded messages to be decoded by each user. Specifically, an integer parameter $\alpha \in [\min\{N_T, K - t\}]$ is leveraged in [106] to control the number of active users in each time slot, which is set to $t + \alpha$, and leads to a content delivery scheme with $B = \binom{K}{t+\alpha}$ time slots. In each time slot, a fraction of the desired coded messages for all the active users are transmitted. In addition to α, another integer parameter β determines the possible set partitions of the user subset in each time slot. When $t + \alpha$ is divisible by $t + \beta$, the user subset can be partitioned into $\frac{t+\alpha}{t+\beta}$

non-overlapping subsets, and a fraction of the desired coded messages for each partition can be transmitted simultaneously. It is shown in [106] that the system performance can be improved if multiple groups of messages can be transmitted in parallel, i.e., $\frac{t+\alpha}{t+\beta} \geq 2$, as compared to the case $\beta = \alpha$. Moreover, the number of messages for each user to decode in each time slot is $\binom{t+\beta-1}{t}$, which is an exponential function of β. Therefore, by adjusting the value of β, the number of coded messages for each user in each time slot is indirectly adjusted.

Instead of limiting the subsets of users to be served in each time slot, we propose to directly adjust the number of coded messages targeted to each user. We will show that this results in a more efficient delivery scheme than the one in [106]. In Setting 1, if we transmit all the messages in one time slot, i.e., $B = 1$, a total of $|\mathcal{S}| = \binom{K}{t+1} = 10$ coded subfiles are transmitted simultaneously, with each user decoding $\binom{K-1}{t} = 4$ messages. Accordingly, in the optimization problem in (4.12) we will have $K \times (2^{|\mathcal{S}_k|} - 1) = 75$ constraints. To alleviate the computational complexity, the low complexity scheme in [106] splits each subfile into 3 minifiles, and the coded messages are grouped to serve a subset of $t + \alpha = 3$ users in each of the $B = \binom{K}{t+\alpha} = 10$ time slots. Within each time slot, each user needs to decode 2 messages. Note that the power minimization problem for each time slot can be solved independently; therefore, we would need to solve 10 smaller optimization problems, each with $3 \times 3 = 9$ constraints.

In contrast, we propose to serve as many users as needed at each time slot while keeping $c_k(i)$ under a given threshold s for each user k. In our Setting 1, we can satisfy all the user requests in only 2 time slots, by setting nonzero rate targets for the messages in

$$\begin{aligned}
&\mathcal{S}(1) = \{\{1,2\},\{2,3\},\{3,4\},\{4,5\},\{1,5\}\}, \text{ and} \\
&\mathcal{S}(2) = \{\{1,3\},\{2,4\},\{3,5\},\{1,4\},\{2,5\}\}
\end{aligned}$$

in time slots 1 and 2, respectively. Note that each user k decodes only $c_k(i) = s =$

2 messages in each time slot, the same as the delivery scheme in [106], requiring the same implementation complexity at each user; however, 5 users are served in each time slot, which results in a significantly smaller number of time slots. Thus, we need to solve only two optimization problems at the BS, each with $5 \times 3 = 15$ constraints.

In general, the number of constraints in the optimization problem in (4.12) increases exponentially with s, which results in exponentially increasing number of constraints in the problem in each SCA iteration. Thus the computational complexity of the delivery scheme can be largely alleviated by choosing a small s value, which also simplifies the multi-user detection algorithm.

The key idea of our proposed low-complexity scheme is to divide set $\mathcal{S}$ into disjoint subsets $\mathcal{S}(1), \cdots, \mathcal{S}(B)$, with $c_k(i) \leq s$, $\forall k, i$, while keeping B as small as possible. Since the total number of subfiles to transmit is fixed, choosing a small value of B, i.e., completing the delivery phase within a small number of time slots, requires multiplexing more messages in each time slot, without increasing the complexity of the receivers. To obtain this low-complexity scheme, the following optimization problem can be formulated:

$$\min_{\{\boldsymbol{v}_{\mathcal{T}}\}, B} \quad B \tag{4.17a}$$

$$\text{s.t.} \sum_{\mathcal{T} \ni k} v_{\mathcal{T}}(i) \leq s, \forall i \in [B], \; k \in [K], \tag{4.17b}$$

$$\sum_{i=1}^{B} v_{\mathcal{T}}(i) = 1, \forall \mathcal{T}, \tag{4.17c}$$

$$v_{\mathcal{T}}(i) \in \{0, 1\}, \forall \mathcal{T}, i \in [B], \tag{4.17d}$$

where constraint (4.17b) imposes that each user decodes no more than s messages in each time slot, while (4.17c) requires that each message will be transmitted in only one time slot. However, since the problem itself varies with variable B, the problem is not in a tractable form. By introducing $L \geq B$ as a prescribed parameter that determines the dimension of the problem, and an auxiliary variable $\boldsymbol{q} \in \{0,1\}^L$, problem (13) can be equivalently written as

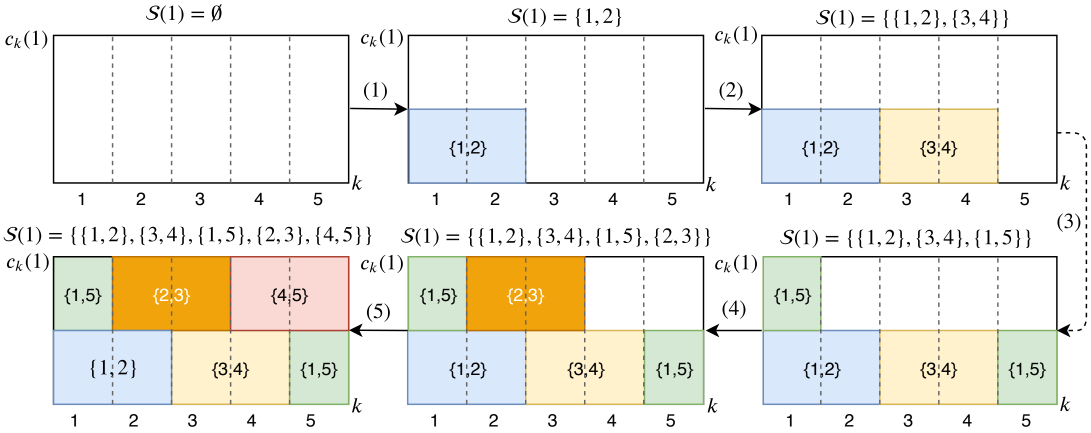
$$
\begin{aligned}
B = \min_{\{\boldsymbol{v}_{\mathcal{T}}\},\boldsymbol{q}} \quad & \mathbf{1}^T \boldsymbol{q} && \text{(4.18a)}\\
\text{s.t.} \quad & \sum_{\mathcal{T} \ni k} v_{\mathcal{T}}(i) \leq s, \forall i \in [L],\ k \in [K], && \text{(4.18b)}\\
& \sum_{i=1}^{L} v_{\mathcal{T}}(i) = 1, \forall \mathcal{T}, && \text{(4.18c)}\\
& \sum_{\mathcal{T}} v_{\mathcal{T}}(i) \leq \binom{K}{t+1} q_i, \forall i \in [L], && \text{(4.18d)}\\
& v_{\mathcal{T}}(i) \in \{0,1\}, \forall \mathcal{T}, i \in [L], && \text{(4.18e)}\\
& \boldsymbol{q} \in \{0,1\}^L, && \text{(4.18f)}
\end{aligned}
$$

where $\mathbf{1}$ denotes a column vector of all ones. Since $\binom{K}{t+1}$ is a bound on $\sum_{\mathcal{T}} v_{\mathcal{T}}(i)$, the optimal q_i is 1 if $\sum_{\mathcal{T}} v_{\mathcal{T}}(i)$ is nonzero, and 0 otherwise, in order to minimize $\sum\limits_{i=1}^{L} q_i$ in the objective. Note that problem (14) can be considered as minimizing the number of time slots employed out of a maximum L available time slots. We can set $L = \binom{K}{t+1}$ which guarantees the existence of a solution; however, choosing a smaller L will reduce the complexity of the problem. The problem in (4.18) is a $0-1$ integer programming problem, which is generally NP-hard [138].

Algorithm 1 Low-complexity greedy delivery scheme

```
Require: $N, K, M, s, R$
Ensure: $B$, $\bigcup_{i=1}^{B}\{S(i)\}$, $\bigcup_{i=1}^{B}\{n_i\}$, $\forall \mathcal{T}$
Set $t = \frac{MK}{N}$, $i = 1$, and $\mathcal{E} = \mathcal{S}$
while $\mathcal{E} \neq \varnothing$ do
    Set $\boldsymbol{c}(i) \triangleq [c_1(i) \cdots c_K(i)] = \mathbf{0}$, $\mathcal{S}(i) = \varnothing$, $\mathcal{C} = \mathcal{E}$
    while $c_k(i) \leq s, \forall k \in [K]$ and $\mathcal{C} \neq \varnothing$ do
        $\mathcal{K} \triangleq \{k \mid \arg\min_{k} \boldsymbol{c}(i)\}$
        Find $\hat{\mathcal{T}} =_{\mathcal{T}\in\mathcal{C}} |\mathcal{K} \bigcap \mathcal{T}|$
        $\mathcal{C} = \mathcal{C} \backslash \{\hat{\mathcal{T}}\}$
        if $c_k(i) + 1 \leq s, \forall k \in \hat{\mathcal{T}}$ then
            $c_k(i) = c_k(i) + 1, \forall k \in \hat{\mathcal{T}}$
            $\mathcal{S}(i) = \mathcal{S}(i) \bigcup \hat{\mathcal{T}}$, $\mathcal{E} = \mathcal{E} \backslash \{\hat{\mathcal{T}}\}$
        else
            break
        end if
    end while
    $i \leftarrow i + 1$
end while
Set $B = i - 1$
for $i = 1 : B$ do
    $n_i = \frac{|\mathcal{S}(i)|}{\binom{K+1}{t}} n$
    $R^{\mathcal{T}}(i) = \begin{cases} \frac{R}{\binom{K}{t}}, \ \forall \mathcal{T} \in \mathcal{S}(i) \\ 0, \ \text{otherwise} \end{cases}$
end for
```

In Algorithm 1, we propose a greedy solution that constructs disjoint $\mathcal{S}(i)$ sets for any s value. Specifically, $\mathcal{S}(i)$'s are generated in a sequential manner: to construct $\mathcal{S}(i)$, we initialize $\boldsymbol{c}(i) \triangleq [c_1(i) \cdots c_K(i)] = \mathbf{0}$, $\mathcal{S}(i) = \varnothing$, and the set $\mathcal{E} = \mathcal{S} \backslash \bigcup_{j=1}^{i-1} \mathcal{S}(j)$ of remaining messages for assignment, we first identify the



$\mathcal{S}(1)=\emptyset$
$\mathcal{S}(1)=\{1,2\}$
$\mathcal{S}(1)=\{\{1,2\},\{3,4\}\}$
$\mathcal{S}(1)=\{\{1,2\},\{3,4\},\{1,5\},\{2,3\},\{4,5\}\}$
$\mathcal{S}(1)=\{\{1,2\},\{3,4\},\{1,5\},\{2,3\}\}$
$\mathcal{S}(1)=\{\{1,2\},\{3,4\},\{1,5\}\}$
$c_k(1)$
k
1 2 3 4 5
(1)
(2)
(3)
(4)
(5)
{1,2}
{3,4}
{1,5}
{2,3}
{4,5}

FIGURE 4.2: Illustration of the proposed low-complexity greedy scheme for the network with $N = K = 5$, and $M = 1$.

user(s) that have decoded the least number of messages so far, i.e., user(s) in set $\mathcal{K} \triangleq \{k|\arg\min_k \boldsymbol{c}(i)\}$, and check whether there exists a message $\hat{\mathcal{T}} \in \mathcal{E}$ such that the condition $c_k(i) + 1 \leq s$ for $\forall k \in \hat{\mathcal{T}}$ holds. If no such $\hat{\mathcal{T}}$ can be found, the process of constructing $\mathcal{S}(i)$ is completed, and we start constructing $\mathcal{S}(i+1)$ in the same manner. The whole procedure is completed when $\mathcal{E} = \varnothing$, i.e., all the messages have been assigned to a subset. Note that our proposed greedy scheme covers the case of $B = 1$, where all the messages are sent simultaneously, if $s \geq \binom{K-1}{t}$.

Next we elaborate the proposed greedy content delivery algorithm in Settings 1 and 2.

Setting 1: $N = 5,\ K = 5,\ M = 1$. $t = \frac{MK}{N} = 1$. Suppose $s = 2$.

As illustrated in Fig. 4.2, the algorithm starts by constructing $\mathcal{S}(1)$, i.e., identifying the coded messages to be delivered in the first time slot, initialized as $\boldsymbol{c}(1) \triangleq [c_1(1) \cdots c_K(1)] = \mathbf{0}$, $\mathcal{S}(1) = \varnothing$. Firstly, it is obvious that $\mathcal{K} = [K]$ since $c_k(1) = 0,\ \forall k$. Hence, one may choose any of the available messages in $\mathcal{E} = \mathcal{S}$. Suppose message $s_{\{1,2\}}$ is chosen. We update $c_1(1) = c_2(1) = 1$, $\mathcal{E} = \mathcal{E} \backslash \{1, 2\}$. The algorithm then identifies the updated $\mathcal{K} = [3, 4, 5]$, according to which one may choose from $s_{\{3,4\}}, s_{\{3,5\}}, s_{\{4,5\}}$ without violating the constraint

$c_k(1)+1 \leq s$, $\forall k$. Suppose $s_{\{3,4\}}$ is chosen, and we have $c_1(1)=c_2(1)=c_3(1)=c_4(1)=1$, $\mathcal{E}=\mathcal{E}\backslash\{3,4\}$, and $\mathcal{K}=[5]$; and accordingly one may choose from $s_{\{1,5\}}, s_{\{2,5\}}, s_{\{3,5\}}, s_{\{4,5\}}$, while still keeping the constraint $c_k(1)+1 \leq s$, $\forall k$. Similarly, messages $s_{\{2,3\}}$ and $s_{\{4,5\}}$ can be chosen, and the algorithm for $\mathcal{S}(1)$ is completed since $c_k(1)=2$, $\forall k$, and adding any of the remaining messages will violate the constraint. The algorithm then turns to construct $\mathcal{S}(2)$ similarly, until all the messages have been chosen, i.e., $\mathcal{E}=\emptyset$.

Remark 1: As it can be seen above, the content delivery scheme obtained via Algorithm 1 for Setting 1 is not unique. As a matter of fact, there may exist more than one content delivery scheme that satisfies the constraints specified by s on the maximum number of messages that can be decoded at each user. For instance, another feasible content delivery scheme with $s=2$ for Setting 1 is:

$$\mathcal{S}(1)=\{\{1,2\},\{3,4\},\{1,5\},\{2,4\},\{4,5\}\}, \text{and}$$
$$\mathcal{S}(2)=\{\{1,3\},\{2,3\},\{3,5\},\{1,4\},\{2,5\}\}.$$

Setting 2: $N=4,\ K=4,\ M=1.\ t=\frac{MN}{K}=1.$

We present the content delivery schemes obtained from Algorithm 1 for different values of s:

Case 1: $s = 1$. Algorithm 1 leads to $B=3$, and

$$\mathcal{S}(1)=\{\{1,2\},\{3,4\}\},$$
$$\mathcal{S}(2)=\{\{1,3\},\{2,4\}\},$$
$$\mathcal{S}(3)=\{\{1,4\},\{2,3\}\},$$

where the transmission takes $B=3$ time slots, each with $\frac{1}{3}n$ channel uses. It is noted that the scheme is the same as the one in [106] obtained for $\alpha=3$ and $\beta=1$, in the sense that the same sets of messages are transmitted over the same number of time slots.

Case 2: $s = 2$. Algorithm 1 leads to $B = 2$, and

$$\begin{aligned}\mathcal{S}(1) &= \{\{1,2\},\{3,4\}\},\{\{1,3\},\{2,4\}\},\\ \mathcal{S}(2) &= \{\{1,4\},\{2,3\}\}.\end{aligned}$$

It is noted that β is not available to induce a scheme in [106] in this scenario, since β can only take the value of 2 to have $s = 2$, making $t + \alpha$ not divisible by $t + \beta$.

Remark 2: The proposed greedy content delivery scheme can be easily extended by limiting the number of active users as in [106]. Specifically, instead of serving as many users as possible, which is up to K, Algorithm 1 can be applied for a user subset of size $t + \alpha$ to obtain a content delivery scheme under the constraints on the number of messages to decode. While it is required $\frac{t+\alpha}{t+\beta} \in \mathbb{N}$ to induce a content delivery scheme in [106], Algorithm 1 always provides a delivery scheme for any s value. Therefore, our proposed greedy scheme can be considered as a generalization of the one in [106].

Remark 3: It is noted that the proposed greedy content delivery scheme in Algorithm 1 may lead to unequal number of messages transmitted in different time slots, which can be highly sub-optimal. An intuitive way to enhance the performance is to allocate more channel uses to the time slot with more messages to deliver. In general, once the non-overlapping partition of $\mathcal{S}$, i.e., $\bigcup_{i=1}^{B}\{\mathcal{S}(i)\}$ is obtained, we can set the total blocklength of the transmission of $\mathcal{S}(i)$ proportionally to the number of messages $|\mathcal{S}(i)|$. For instance, in the case of $s = 2$ in Setting 2, we can allocate $\frac{2n}{3}$ channel uses for $\mathcal{S}(1)$, and $\frac{n}{3}$ channel uses for $\mathcal{S}(2)$.

Numerical results for the minimum required power for the proposed greedy transmission scheme, and the comparison with the one proposed in [106] will be presented in Section 4.6.

4.5 Joint Optimization of Beamforming and Coded Content Delivery

In this section, we formulate a sparsity constrained power minimization problem to jointly optimize the beamformers and the content delivery scheme. The sparsity induced problem directly limits the number of messages to be decoded by each user at any time slot, and the indicator function $v_{\mathcal{T}}(i)$ is identified by setting $v_{\mathcal{T}}(i) = |R^{\mathcal{T}}(i)|_0$, for $\forall i, \mathcal{T}$, where $|\cdot|_0$ denotes the ℓ_0-norm and is equal to the number of non-zero elements of a vector. Therefore, we impose an ℓ_0-norm constraint on the rates of messages at any time slot i as follows:

$$\sum_{\mathcal{T}\in\mathcal{S}_k} |R^{\mathcal{T}}(i)|_0 \leq s, \ \forall k, i. \tag{4.20}$$

In this section, we assume equal blocklength allocation over all the B time slots for simplicity. Then, the minimum required power problem with the constrains on the number of messages to be decoded by any user at any time slot can be formulated as follows:

$$\min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\},\{R^{\mathcal{T}}(i)\}} \ \frac{1}{B}\sum_{i=1}^{B}\sum_{\mathcal{T}\in\mathcal{S}} \|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.21a}$$

$$\text{s.t.} \ \sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} R^{\mathcal{T}}(i) \leq \frac{1}{B}\log_2\Bigg(1+\sum_{\mathcal{T}\in\pi^j_{\mathcal{S}_k}} \gamma_k^{\mathcal{T}}(i)\Bigg), \ \forall \pi^j_{\mathcal{S}_k} \in \Pi_{\mathcal{S}_k}, \tag{4.21b}$$

$$\textstyle\sum_{i=1}^{B} R^{\mathcal{T}}(i) \geq \frac{R}{\binom{K}{t}}, \ \forall \mathcal{T}, \tag{4.21c}$$

$$\sum_{\mathcal{T}\in\mathcal{S}_k} |R^{\mathcal{T}}(i)|_0 \leq s, \ \forall k, i, \tag{4.21d}$$

where $\gamma_k^{\mathcal{T}}(i)$ is defined in (4.9). In this problem, the objective is to minimize the average transmission power over all the time slots; constraints (4.21b)-(4.21c) guarantee successful decoding of all the required messages at each user in each time slot; and constraint (4.21d) limits the number of messages decoded by each user in any time slot. Since (4.21d) limits only the number of messages decoded by each user in any time slot, without assuming any specific content delivery scheme, the problem in (4.21) includes the content delivery schemes in [106] as a special case. This formulation also generalizes the one presented in Section 4.3 when the time slots are of equal duration. However, note that the number of time slots B is a free variable for the greedy scheme, while it is assumed to be given in (4.21).

To deal with the discontinuous ℓ_0-norm constraint in (4.21d), we approximate it with a differentiable continuous function [139]

$$f(R^{\mathcal{T}}(i), t) \triangleq \frac{2}{\pi} \text{arctan} \frac{R^{\mathcal{T}}(i)}{\xi}, \tag{4.22}$$

where $\xi > 0$ is a prescribed constant that determines the approximation accuracy. The function in (4.22) is concave with respect to $R^{\mathcal{T}}(i)$, therefore the approximate constraint for (4.21d)

$$\sum_{\mathcal{T} \in \mathcal{S}_k} f(R^{\mathcal{T}}(i), t) \leq s, \ \forall k, i, \tag{4.23}$$

is concave and can be treated as a difference of convex function. Overall, the original problem in (4.21) can be approximated by the following problem:

$$\min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\}, \{R^{\mathcal{T}}(i)\}, \{\eta_{\pi^j_{\mathcal{S}_k}}(i)\}} \sum_{i=1}^{B} \sum_{\mathcal{T} \in \mathcal{S}} \frac{1}{B} \|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.24a}$$

$$\text{s.t.} (4.21\text{b}), (4.21\text{c}) \text{ and } (4.23).$$

Similarly to (4.12), the problem in (4.21) can be solved via the SCA method. Specifically, the convex subproblem to be solved in the $(\nu + 1)$-th iteration is

$$\min_{\{\boldsymbol{w}_{\mathcal{T}}(i)\}, \{R^{\mathcal{T}}(i), \{\eta_{\pi^j_{\mathcal{S}_k}}(i)\}} \frac{1}{B} \sum_{i=1}^{B} \sum_{\mathcal{T} \in \mathcal{S}} \|\boldsymbol{w}_{\mathcal{T}}(i)\|^2 \tag{4.25a}$$

$$\text{s.t.} \quad \sum_{\mathcal{T}\in\mathcal{S}_k} \arctan\frac{R^{\mathcal{T}^{\nu}}(i)}{t} + \frac{t}{\left(t^2 + R^{\mathcal{T}^{\nu 2}}(i)\right)}\left(R^{\mathcal{T}}(i) - R^{\mathcal{T}^{\nu}}(i)\right) \leq \frac{\pi s}{2},\ \forall k, i, \tag{4.25b}$$

$$(4.14\text{b}), (4.14\text{c}), \text{and}\, (4.21\text{c}). \tag{4.25c}$$

The initialization of the SCA algorithm for problem (4.25) for a given value of s requires a content delivery scheme with less or equal complexity, which can be obtained via Algorithm 1 in Section 4.4. The associated beamforming design can be readily obtained similarly to (4.14). From problem (4.21), we can also conclude that the minimum required power of a content delivery scheme is a non-decreasing function of s, as the problem becomes more relaxed as s increases.

4.6 Simulation Results

We consider a single-cell with radius 500m, and users uniformly randomly distributed in the cell. Channel vectors $\boldsymbol{h}_k$ are written as $\boldsymbol{h}_k = (10^{-\text{PL}/10})^{1/2}\tilde{\boldsymbol{h}}_k$, $\forall k$, where $\tilde{\boldsymbol{h}}_k$ denotes an i.i.d. vector accounting for Rayleigh fading of unit power, and the path loss exponent is modeled as $\text{PL} = 148.1 + 37.6\log_{10}(v_k)$, with v_k denoting the distance between the BS and the user (in kilometers). The noise variance is set to $\sigma_k^2 = \sigma^2 = -134$ dBW for all the users. All simulation results are averaged over 300 independent trials computed with CVX [140].

The scheme with $B = 1$ time slot will be referred to as the full superposition (FS) scheme. FS has the best performance in terms of transmit power given enough spatial DoF, and serves as a baseline, but it also has the highest complexity. To compare our results with those in [106], same number of coded messages are transmitted to each user in each time slot for both schemes. We note here that with the use of β parameter, the scheme in [106] can be improved by serving disjoint subsets of users simultaneously without increasing the complexity, but the improvement is only applicable when the size of user subset can be partitioned equally and exactly. Therefore, the scheme in [106] cannot handle certain settings such as the case of $s = 2$ in Setting 1.

We first present the average transmit power as a function of the target rate R in Fig. 4.3 for Setting 1, assuming that the BS is equipped with $N_T = 6$ antennas. The scheme in [106] that satisfies $s = 2$ is adopted for fair comparison, where $t + \alpha = 3$ users are served in each time slot. We observe that the proposed greedy scheme provides significant savings in the transmit power compared to the one in [106] at all rates. The power savings increase with rate R as a result of the increased superposition coding gain. Furthermore, the gap between the proposed greedy scheme and FS is quite small, and remains almost constant with rate. At $R = 8$ bps/Hz, the power loss of the scheme in [106] and ours compared to FS are about 8.5 dB and 0.5 dB, respectively. Hence, we can conclude that the proposed greedy scheme provides significant reduction in the computational complexity without sacrificing the performance much.

The average transmit power as a function of file rate R is further investigated for the setting with $N = 6$ files, $K = 6$ users, $M = 1$, and $N_T = 6$ antennas. Similarly to Fig. 4.3, it is observed in Fig. 4.4 that our proposed low-complexity greedy scheme substantially outperforms the scheme in [106] with the same value of s in the high SNR regime. For example, the power savings of the greedy delivery scheme compared to [106] are 8dB and 2dB, for $s = 3$ and $s = 4$, respectively, at $R = 10$ bps/Hz. The power gain is again observed to be larger as the rate increases, while in the low SNR/rate regime, all the schemes achieve comparable performance regardless of s. Also, for the proposed greedy scheme, a larger s allows achieving the same rate with lower transmit power in the high SNR regime, at the expense of increased complexity at the receivers. It is noted that the proposed greedy scheme yields the same content delivery scheme in terms of the transmitted coded messages in each time slot as the one in [106] when $s = 1$ and $s = 2$, which correspond to $\alpha = 5, \beta = 1$, and $\alpha = 5, \beta = 2$

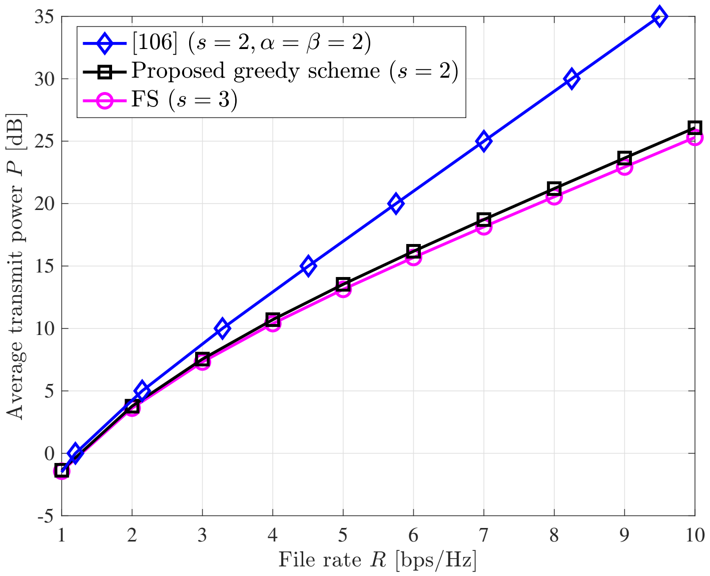


FIGURE 4.3: Average transmit power P as a function of rate R for the network for $N = K = 5$, $M = 1$, and $N_T = 6$.

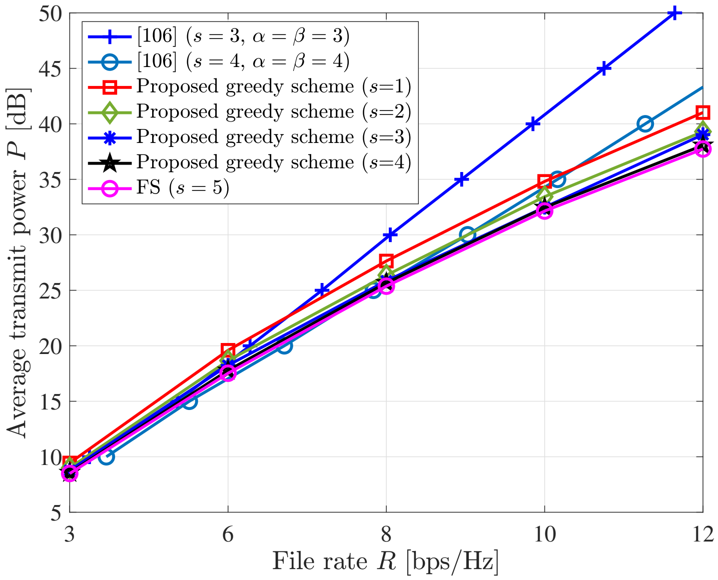


FIGURE 4.4: Average transmit power P as a function of rate R for $N = K = 6$, $M = 1$, and $N_T = 6$.

in [106], respectively.

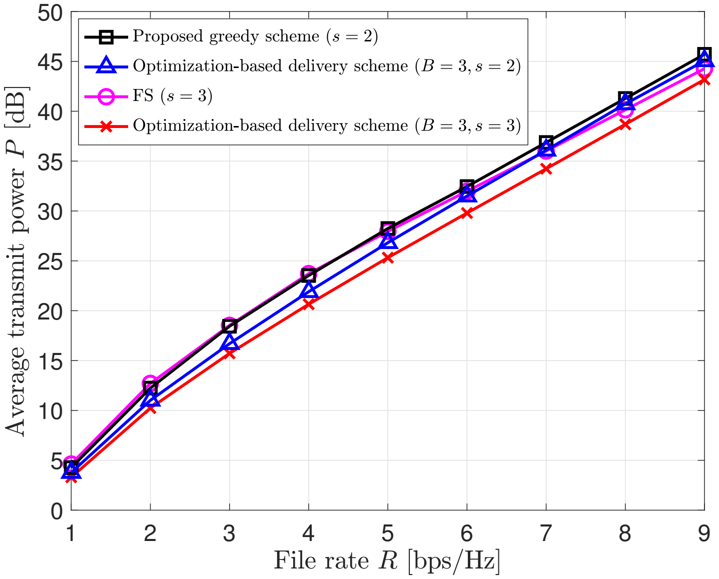


FIGURE 4.5: Average transmit power P as a function of rate R for $N = K = 4$, $M = 1$, and $N_T = 3$.

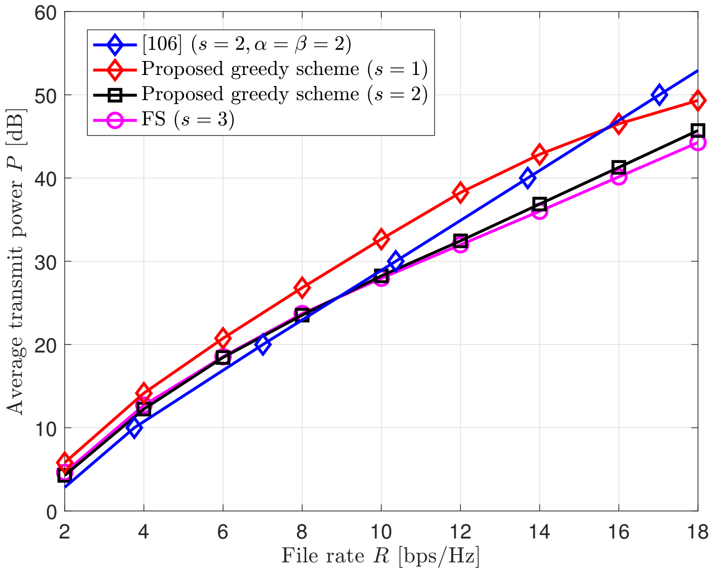


FIGURE 4.6: Average transmit power P as a function of rate R for $N = K = 4$, $M = 1$, and $N_T = 3$.

Fig. 4.5 and Fig. 4.6 show the average transmit power versus rate R for Setting 2, with $N = 4$ files, $K = 4$ users, $M = 1$, and $N_T = 3$ antennas. In Fig. 4.5, we compare our greedy content delivery scheme with the one obtained by solving the problem in (4.21) for $s = 2$ and $s = 3$, by setting $B = 3$. It is seen that the greedy scheme can achieve comparable performance, and the performance gap is small especially for high rates. The optimization-based content delivery scheme with $s = 3$ is found to outperform the one with $s = 2$ as expected, and the improvement is larger as the rate increases. In Fig. 4.6, it is interesting to see that when the rate is low, the scheme in [106] slightly outperforms both the FS and the proposed schemes. Due to insufficient spatial degrees of freedom, both the FS and the proposed schemes fail to manage the interference between data streams. We conclude that this effect occurs only for low rates, as the benefit of superposition coding becomes more dominant at higher rates. We note that the greedy scheme coincides with the one in [106] for $s = 1$ in terms of the transmitted coded messages in each time slot, but this does not always happen. For instance, when $s = 2$, the only option in [106] to keep the same level of complexity is to serve 3 users in each time slot.

We plot in Fig. 4.7 the power loss of the proposed scheme in Algorithm 1 compared to FS as a function of s. Assuming $N = 6$, $K = 6$, $M = 1$, we let s take values from $\{1, 2, 3, 4, 5\}$, where $s = 5$ corresponds to the FS scheme in which all the $\binom{K}{t+1} = 15$ coded messages are transmitted simultaneously. At the other extreme, when $s = 1$, the model boils down to the single-cell multigroup multicasting problem, which has the lowest computation and implementation complexity. In general, Fig. 4.7 can be considered as the trade-off curve between the performance and complexity for each rate value, both of them increasing with s.

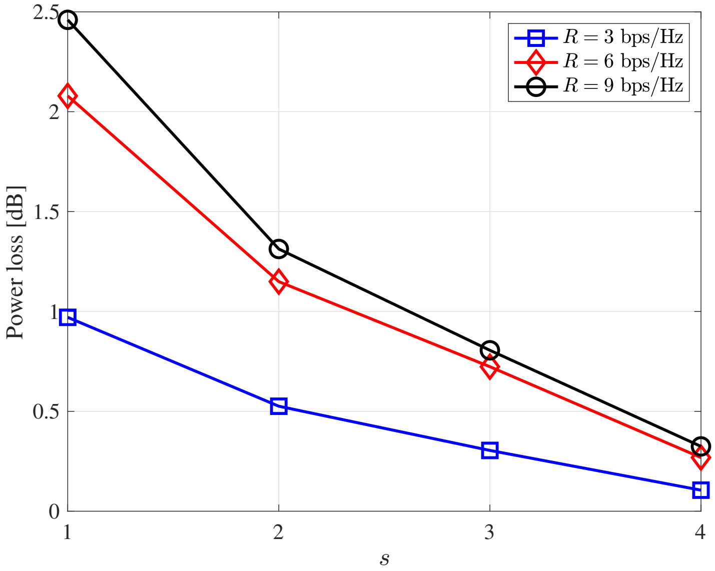


FIGURE 4.7: Power loss with respect to FS as a function of s for $N = K = 6$, $M = 1$, and $N_T = 6$.

4.7 Conclusions

In this chapter, we have studied cache-aided content delivery from a multi-antenna BS in the finite SNR regime. We have formulated a general beamforming scheme that multicasts coded files over multiple orthogonal time slots. We have then specialized this general formulation to a low-complexity greedy scheme by limiting the number of coded messages targeted at each user at each time slot. This scheme provides the flexibility to adjust the computational complexity of the optimization problem and the receiver complexity. We have then formulated the constraint on the number of coded messages targeted at each user at each time slot as a sparsity constraint, and solved the resulting mixed-integer non-convex optimization problem using the SCA method. Compared with FS, where all the coded messages are transmitted simultaneously, and the scheme obtained via the sparsity-constrained optimization framework, the greedy scheme achieves comparable performance, and outperforms the one proposed in [106] for all values of SNR and rate with sufficient spatial degrees of freedom, while the improvement is limited to high data rate values when the BS does not have sufficiently many transmit antennas. Furthermore, the gap between the greedy delivery scheme and the optimization-based delivery scheme decreases as the SNR/power increases. When considering practical implementations, one must choose a suitable value of s that yields an acceptable performance while keeping the complexity feasible.

Chapter 5

Conclusions and Future Work

MIMO has been very successful in providing highly efficient and reliable content delivery in mobile networks. To face the forthcoming challenges brought up by 5G, MIMO techniques can be investigated together with non-orthogonal multiple access approaches, and coded content delivery schemes for further performance enhancement. In this dissertation, we have studied non-orthogonal content delivery in MISO broadcast channels with transmit beamforming, successive decoding, and coded content delivery in the presence of cache memories at the receivers. In the following we conclude the results presented in each technical chapter and discuss possible directions for future research.

In Chapter 2, we have studied joint broadcast and unicast transmission via beamforming and LDM. As a NOMA approach for simultaneous transmission of broadcast and unicast messages over cellular networks, LDM has a significant performance gain over orthogonal approaches such as TDM/FDM. We have formulated a joint beamforming design and power allocation problem with the objective to minimize the total power consumption, and with distinct QoS constraints for the individual unicast messages and the common broadcast message. The formulated problem is non-convex and NP-hard, therefore we have provided upper and lower bounds on the optimal solution via SCA and S-procedure, respectively. Numerical results have demonstrated that the upper and lower bounds are tight, which indicates the near-optimality of the proposed solutions via SCA. We have observed that the benefit of the increased bandwidth available for the broadcast layer outweighs the interference caused by unicast transmissions, indicating that LDM has better interference management capability than

TDM/FDM. We have also observed that the CSI errors have less impact on the power consumption of LDM than that of TDM, indicating that LDM also provides better robustness against CSI uncertainties commonly experienced in real systems. Motivated by the advantageous of LDM over TDM/FDM, we have also derived a dual decomposition-based distributed algorithm for LDM, which facilitates efficient distributed implementation for this promising non-orthogonal technique.

In Chapter 4, we have studied the multicast beamforming design for coded content delivery. The coded content placement and delivery scheme in [21] has attracted a great deal of attention, thanks to the global caching gain provided by creating and exploiting multicasting opportunities. Having observed the exponentially increasing complexity of the coded delivery scheme as the number of subfiles decoded by users increases, we have first proposed a low-complexity coded content delivery scheme in the downlink of MISO networks, which limits the number of coded messages targeted at each user in each time slot. The proposed greedy coded content delivery scheme strikes a balance between the system performance and the receiver complexity, via the flexibility to adjust the computational complexity of the optimization problem. Moreover, the greedy scheme has been shown to significantly outperform the state-of-the-art proposed in [106] for all data rates when there is sufficient spatial degrees of freedom, while the improvement is limited to high data rate values when the BS does not have sufficiently many transmit antennas. Finally, we have considered the joint design of content delivery scheme and multicast beamforming, by tackling the power minimization problem under QoS constraints and additional sparsity constraints, which specify the maximum number of messages for each user to decode in each time slot. Simulation results have shown that our proposed greedy scheme has minimal gap to the optimization-based content delivery design, and the gap decreases as the target rate increases.

Improving the spectral efficiency in wireless networks is expected to remain a key research issue in communications in the foreseeable future, driven by the explosively increasing demand on high quality content, and the ongoing trend of Internet of Things (IoT) that expects 29 billion connected devices by 2022. This thesis aims to investigate physical layer techniques to increase the network throughput, for the purpose of efficient content delivery. While our research is carried out for specific yet general application scenarios, such as delivery of TV services over wireless networks and synchronized cache-aided multicasting, the superiority of multicast beamforming, NOMA, and coded caching presented for our considered scenarios exhibits potentials of these techniques for other important settings, and motivates further explorations on network performance enhancement with these techniques.

Potential Extensions

In our work on joint beamforming for broadcast and unicast transmission, we have considered the setting that a single broadcasting stream is requested by all the users. A more general setting is that multiple broadcasting services, e.g., multiple TV programs, are available in the network, and users randomly request access to one of the broadcasting streams, in addition to the unicast services. Different levels of cooperation between BSs can be performed for the purpose of joint broadcast and unicast transmission, which have not been studied in the literature.

In this thesis, we have considered multicast beamforming with the centralized coded caching scheme in [21], where the placement phase is designed with the knowledge of the number and identity of users at the server. However, practical scenarios are highly dynamic as the users can join the network and request files at any time, which may not be known in advance by the server. To tackle this issue, decentralized caching has been studied in various settings, and efficient multicast beamforming designs with decentralized caching can be investigated.

In addition to the immediate extensions mentioned above, we briefly discuss a few possible extensions that can be considered in the future. In this thesis, we have primarily focused on conventional MIMO techniques with NOMA and coded content delivery. One major direction for future work can be exploiting hybrid precoding in millimeter waves for joint non-orthogonal broadcast and unicast transmission, which is a timely extension with the recent efforts to develop 5G in millimeter wave bands.

In the work of multicast beamforming for coded content delivery, we have adopted the content placement and delivery design proposed by [21], which has not taken the impact of content popularity into consideration. Another extension for this work can be exploiting popularity-aware caching with multicast beamforming, with the goal to maximize the multicasting opportunities for efficient transmission.

Bibliography

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] Qualcomm, “5G vision for the next generation of connectivity,” https://www.qualcomm.com/media/documents/files/whitepaper-5g-vision-for-the-next-generation-of-connectivity.pdf, 2015.

[3] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022 white paper,” https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html, Feb. 2019.

[4] 3GPP, TS 25.308, “High speed downlink packet access (HSDPA),” https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1171, 2004.

[5] 3GPP, TS 36.212, “Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding (Release 8),” https://portal.3gpp.org/ngppapp/CreateTDoc.aspx?mode=view&contributionUid=SP-190253, Mar. 2010.

[6] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[7] E. Bjornson, L. Van der Perre, S. Buzzi, and E. G. Larsson, “Massive MIMO in sub-6 GHz and mmWave: Physical, practical, and use-case differences,” *IEEE Wirel. Commun.*, vol. 26, no. 2, pp. 100–108, Apr. 2019.

[8] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, “Millimeter-wave massive MIMO communication for future wireless systems: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, Second quarter 2018.

[9] D. Kim, F. Khan, C. V. Rensburg, Z. Pi, and S. Yoon, “Superposition of broadcast and unicast in wireless cellular systems,” *IEEE Commun. Mag.*, vol. 46, no. 7, pp. 110–117, Jul. 2008.

[10] P. Bergmans, “Random coding theorem for broadcast channels with degraded components,” *IEEE Trans. Inf. Theory*, vol. 19, no. 2, pp. 197–207, Mar. 1973.

[11] J. Kim and I. Lee, “Capacity analysis of cooperative relaying systems using non-orthogonal multiple access,” *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1949–1952, Nov. 2015.

[12] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, “Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[13] L. Yin, W. O. Popoola, X. Wu, and H. Haas, “Performance evaluation of non-orthogonal multiple access in visible light communication,” *IEEE Trans. Wireless Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.

[14] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, “Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[15] K. C. Almeroth and M. H. Ammar, “The use of multicast delivery to provide a scalable and interactive video-on-demand service,” *IEEE J. Sel. Areas Commun.*, vol. 14, no. 6, pp. 1110–1122, Aug. 1996.

[16] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

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

[18] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[19] A. Liu and V. K. N. Lau, "Exploiting base station caching in mimo cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.

[20] J. Qiao, Y. He, and X. S. Shen, "Proactive caching for mobile video streaming in millimeter wave 5G networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[21] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May. 2014.

[22] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

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

[24] Wei Yu and J. M. Cioffi, "Sum capacity of gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.

[25] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, achievable rates, and sum-rate capacity of gaussian MIMO broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[26] P. Viswanath and D. N. C. Tse, “Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality,” *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[27] H. Weingarten, Y. Steinberg, and S. S. Shamai, “The capacity region of the gaussian multiple-input multiple-output broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[28] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[29] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Salzer, “Multiuser diversity - multiplexing tradeoff in MIMO broadcast channels with limited feedback,” in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, Oct. 2006, pp. 364–368.

[30] T. Yoo and A. Goldsmith, “Optimality of zero-forcing beamforming with multiuser diversity,” in *IEEE International Conference on Communications, 2005. ICC 2005. 2005*, vol. 1, May. 2005, pp. 542–546 Vol. 1.

[31] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, “Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[32] M. Schubert and H. Boche, “Solution of the multiuser downlink beamforming problem with individual SINR constraints,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[33] C. Geng, N. Naderializadeh, A. S. Avestimehr, and S. A. Jafar, "On the optimality of treating interference as noise," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1753–1767, Apr. 2015.

[34] S. Verdu, *Multiuser Detection*, 1st ed. USA: Cambridge University Press, 1998.

[35] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* USA: Wiley, 2006.

[37] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," vol. 42, no. 2, pp. 364–375, Mar. 1996.

[38] A. Carleial, "A case where interference does not reduce capacity (corresp.)," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 569–570, Sep. 1975.

[39] H. Sato, "The capacity of the gaussian interference channel under strong interference (corresp.)," *IEEE Trans. Inf. Theory*, vol. 27, no. 6, pp. 786–788, Nov. 1981.

[40] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49–60, Jan. 1981.

[41] S. A. Jafar, G. J. Foschini, and A. J. Goldsmith, "Phantomnet: Exploring optimal multicellular multiple antenna systems," *EURASIP J. Adv. Signal Process*, vol. 2004, no. 5, pp. 591–604, May. 2004.

[42] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, 2005.

[43] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser

communication-part i: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[44] F. Rashid-Farrokhi, K. J. R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1437–1450, Oct. 1998.

[45] E. Visotsky and U. Madhow, "Optimum beamforming using transmit antenna arrays," in *1999 IEEE 49th Vehicular Technology Conference (Cat. No.99CH36363)*, vol. 1, May 1999, pp. 851–856 vol.1.

[46] M. Bengtsson and B. Ottersten, "Optimal downlink beamforming using semidefinite optimization," in *37th Annual Allerton Conference on Communication, Control, and Computing*, 1999, pp. 987–996.

[47] ——, "Optimal and suboptimal transmit beamforming," 2001.

[48] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

[49] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Commun.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.

[50] M. Stojnic, H. Vikalo, and B. Hassibi, "Rate maximization in multiantenna broadcast channels with linear preprocessing," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2338–2342, Sep. 2006.

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

[52] R. Zhang, Y. Liang, and S. Cui, "Dynamic resource allocation in cognitive radio networks," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 102–114, May. 2010.

[53] W. Liao, T. Chang, W. Ma, and C. Chi, "QoS-based transmit beamforming in the presence of eavesdroppers: An optimized artificial-noise-aided approach," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1202–1216, Mar. 2011.

[54] J. Xu, L. Liu, and R. Zhang, "Multiuser MIMO beamforming for simultaneous wireless information and power transfer," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4798–4810, Sep. 2014.

[55] S. Shamai and B. M. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," in *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings*, vol. 3, May. 2001, pp. 1745–1749 vol.3.

[56] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May. 2010.

[57] R. Zhang and S. Cui, "Cooperative interference management with MIMO beamforming," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5450–5458, Oct. 2010.

[58] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

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

[60] Y. Liu, Y. Dai, and Z. Luo, "Coordinated beamforming for MIMO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, Mar. 2011.

[61] M. Hong, R. Sun, H. Baligh, and Z. Luo, “Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.

[62] Y. Cheng, M. Pesavento, and A. Philipp, “Joint network optimization and downlink beamforming for comp transmissions using mixed integer conic programming,” *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, Aug. 2013.

[63] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, “Transmit beamforming for physical-layer multicasting,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[64] K. T. Phan, S. A. Vorobyov, N. D. Sidiropoulos, and C. Tellambura, “Spectrum sharing in wireless networks via qos-aware secondary multicast beamforming,” *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2323–2335, Jun. 2009.

[65] E. Karipidis, N. D. Sidiropoulos, and Z. Luo, “Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups,” *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[66] D. Christopoulos, S. Chatzinotas, and B. Ottersten, “Weighted fair multicast multigroup beamforming under per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5132–5142, Oct. 2014.

[67] ——, “Multicast multigroup beamforming for per-antenna power constrained large-scale arrays,” in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2015, pp. 271–275.

[68] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen, and T. L. Marzetta, “Max–min fair transmit precoding for multi-group multicasting in massive

MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, Feb. 2018.

[69] H. Joudeh and B. Clerckx, "Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7276–7289, Nov. 2017.

[70] M. Alodeh, D. Spano, A. Kalantari, C. G. Tsinos, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1733–1757, thirdquarter 2018.

[71] M. Jordan, X. Gong, and G. Ascheid, "Multicell multicast beamforming with delayed SNR feedback," in *GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference*, Nov. 2009, pp. 1–6.

[72] G. Dartmann, X. Gong, and G. Ascheid, "Low complexity cooperative multicast beamforming in multiuser multicell downlink networks," in *2011 6th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, Jun. 2011, pp. 370–374.

[73] ——, "Cooperative beamforming with multiple base station assignment based on correlation knowledge," in *2010 IEEE 72nd Vehicular Technology Conference - Fall*, Sep. 2010, pp. 1–5.

[74] Z. Xiang, M. Tao, and X. Wang, "Coordinated multicast beamforming in multicell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 12–21, Jan. 2013.

[75] Z. Xiang, M. Tao, and X. Wang, "Massive MIMO multicasting in noncooperative cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1180–1193, Jun. 2014.

[76] O. Tervo, H. Pennanen, D. Christopoulos, S. Chatzinotas, and B. Ottersten, “Distributed optimization for coordinated beamforming in multicell multigroup multicast systems: Power minimization and sinr balancing,” *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 171–185, Jan. 2018.

[77] Y. Shi, J. Zhang, and K. B. Letaief, “Robust group sparse beamforming for multicast green cloud-ran with imperfect CSI,” *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4647–4659, Sep. 2015.

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

[79] T. Weber, A. Sklavos, and M. Meurer, “Imperfect channel-state information in MIMO transmission,” *IEEE Trans. Commun.*, vol. 54, no. 3, pp. 543–552, Mar. 2006.

[80] Yue Rong, S. A. Vorobyov, and A. B. Gershman, “Robust linear receivers for multiaccess space-time block-coded MIMO systems: A probabilistically constrained approach,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1560–1570, Aug. 2006.

[81] X. Zhang, D. P. Palomar, and B. Ottersten, “Statistically robust design of linear MIMO transceivers,” *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3678–3689, Aug. 2008.

[82] M. B. Shenouda and T. N. Davidson, “On the design of linear transceivers for multiuser systems with channel uncertainty,” *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 1015–1024, Aug. 2008.

[83] ——, “Convex conic formulations of robust downlink precoder designs with quality of service constraints,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 714–724, Dec. 2007.

[84] A. Mutapcic, S. . Kim, and S. Boyd, “A tractable method for robust downlink beamforming in wireless communications,” in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, Nov. 2007, pp. 1224–1228.

[85] N. Vucic and H. Boche, “Robust QoS-constrained optimization of downlink multiuser MIMO systems,” *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 714–725, Feb. 2009.

[86] Y. C. Eldar and N. Merhav, “A competitive minimax approach to robust estimation of random parameters,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1931–1946, Jul. 2004.

[87] Yongfang Guo and B. C. Levy, “Robust MSE equalizer design for MIMO communication systems in the presence of model uncertainties,” *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1840–1852, May. 2006.

[88] M. B. Shenouda and T. N. Davidson, “Robust linear precoding for uncertain MIMO broadcast channels,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 4, May. 2006, pp. IV–IV.

[89] G. Zheng, K. Wong, and B. Ottersten, “Robust cognitive beamforming with bounded channel uncertainties,” *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4871–4881, Dec. 2009.

[90] E. A. Gharavol, Y. Liang, and K. Mouthaan, “Robust downlink beamforming in multiuser MIMO cognitive radio networks with imperfect channel-state information,” *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 2852–2860, Jul. 2010.

[91] N. Vucic, H. Boche, and S. Shi, “Robust transceiver optimization in downlink multiuser MIMO systems,” *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3576–3587, Sep. 2009.

[92] A. Tajer, N. Prasad, and X. Wang, “Robust linear precoder design for multi-cell downlink transmission,” *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 235–251, Jan. 2011.

[93] C. Shen, T. Chang, K. Wang, Z. Qiu, and C. Chi, “Distributed robust multicell coordinated beamforming with imperfect CSI: An ADMM approach,” *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2988–3003, Jun. 2012.

[94] E. Björnson, G. Zheng, M. Bengtsson, and B. Ottersten, “Robust monotonic optimization framework for multicell MISO systems,” *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2508–2523, May. 2012.

[95] H. Joudeh and B. Clerckx, “Sum-rate maximization for linearly precoded downlink multiuser MIMO systems with partial CSIT: A rate-splitting approach,” *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.

[96] ——, “Robust transmission in downlink multiuser MIMO systems: A rate-splitting approach,” *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6227–6242, Dec. 2016.

[97] S. Saeedi Bidokhti, M. Wigger, and R. Timo, “Noisy broadcast networks with receiver caching,” *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6996–7016, Nov. 2018.

[98] M. Mohammadi Amiri and D. Gündüz, “Cache-aided content delivery over erasure broadcast channels,” *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 370–381, Jan. 2018.

[99] J. Zhang and P. Elia, “Wireless coded caching: A topological perspective,” in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 401–405.

[100] M. Mohammadi Amiri and D. Gündüz, “Caching and coded delivery over Gaussian broadcast channels for energy efficiency,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1706–1720, Aug. 2018.

[101] A. Ghorbel, K. Ngo, R. Combes, M. Kobayashi, and S. Yang, "Opportunistic content delivery in fading broadcast channels," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec. 2017, pp. 1–6.

[102] K. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.

[103] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May. 2019.

[104] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[105] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[106] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *CoRR*, vol. abs/1711.03364v3, 2018. [Online]. Available: http://arxiv.org/abs/1711.03364v3

[107] J. Zhao, M. M. Amiri, and D. Gündüz, "Multi-Antenna Coded Content Delivery with Caching: A Low-Complexity Solution," *arXiv e-prints*, p. arXiv:2001.01255, Jan. 2020.

[108] E. Piovano, H. Joudeh, and B. Clerckx, "On coded caching in the overloaded MISO broadcast channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 2795–2799.

[109] J. Zhang and P. Elia, “Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback,” *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May. 2017.

[110] S. Zhong and X. Wang, “Joint multicast and unicast beamforming for coded caching,” *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3354–3367, Aug 2018.

[111] F. Hartung, U. Horn, J. Huschke, M. Kampmann, T. Lohmar, and M. Lundevall, “Delivery of broadcast services in 3G networks,” *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 188–199, Mar. 2007.

[112] Qualcomm, “LTE broadcast,” https://www.qualcomm.com/documents/lte-broadcast-white-paper-idc, Sep. 2014.

[113] J. F. Monserrat, J. Calabuig, A. Fernandez-Aguilella, and D. Gómez-Barquero, “Joint delivery of unicast and E-MBMS services in LTE networks,” *IEEE Trans. Broadcast.*, vol. 58, no. 2, pp. 157–167, Jun. 2012.

[114] G. K. Walker, J. Wang, C. Lo, X. Zhang, and G. Bao, “Relationship between LTE broadcast/eMBMS and next generation broadcast television,” *IEEE Trans. Broadcast.*, vol. 60, no. 2, pp. 185–192, Jun. 2014.

[115] L. Shi, E. Obregon, K. W. Sung, J. Zander, and J. Bostrom, “CellTV - on the benefit of TV distribution over cellular networks: A case study,” *IEEE Trans. Broadcast.*, vol. 60, no. 1, pp. 73–84, Mar. 2014.

[116] L. Shi, K. W. Sung, and J. Zander, “Future TV content delivery over cellular networks from urban to rural environments,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6177–6187, Nov. 2015.

[117] 3GPP TR 22.816 V2.0.0, “3GPP enhancement for TV service,” http://www.3gpp.org/DynaReport/22816.htm, Dec. 2015.

[118] 3GPP, “Enhanced television services over 3GPP eMBMS,” http://www.3gpp.org/news-events/3gpp-news/1905-embms_r14, Oct. 2017.

[119] D. Gomez-Barquero, D. Navratil, S. Appleby, and M. Stagg, "Point-to-multipoint communication enablers for the fifth generation of wireless systems," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 53–59, Mar. 2018.

[120] 3GPP SP-190253, "New SID: Architectural enhancements for 5G multicast-broadcast services," https://portal.3gpp.org/ngppapp/CreateTDoc.aspx?mode=view&contributionUid=SP-190253, Mar. 2019.

[121] L. Fay, L. Michael, D. Gómez-Barquero, N. Ammar, and M. W. Caldwell, "An overview of the ATSC 3.0 physical layer specification," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 159–171, Mar. 2016.

[122] S. I. Park *et al.*, "Low complexity layered division multiplexing for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 233–243, Mar. 2016.

[123] D. Gómez-Barquero and O. Simeone, "LDM versus FDM/TDM for unequal error protection in terrestrial broadcasting systems: An information-theoretic view," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 571–579, Dec. 2015.

[124] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, Apr. 2017.

[125] S. Boyd and L. Vandenberghe, *Convex Optimization.* New York, NY, USA: Cambridge University Press, 2004.

[126] Y. C. B. Silva and A. Klein, "Adaptive beamforming and spatial multiplexing of unicast and multicast services," in *2006 IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep. 2006, pp. 1–5.

[127] H. Weingarten, Y. Steinberg, and S. Shamai, "On the capacity region of the multi-antenna broadcast channel with common messages," in *2006 IEEE International Symposium on Information Theory*, Jul. 2006, pp. 2195–2199.

[128] E. Ekrem and S. Ulukus, "On Gaussian MIMO broadcast channels with common and private messages," in *2010 IEEE International Symposium on Information Theory*, Jun. 2010, pp. 565–569.

[129] ——, "An outer bound for the gaussian mimo broadcast channel with common and private messages," *IEEE Transactions on Information Theory*, vol. 58, no. 11, pp. 6766–6772, Nov. 2012.

[130] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[131] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends® in Communications and Information Theory*, vol. 9, no. 2–3, pp. 113–381, 2013.

[132] J. Zhao, D. Gündüz, O. Simeone, and D. Gómez-Barquero, "Non-orthogonal unicast and broadcast transmission via joint beamforming and ldm in cellular networks," *to appear in IEEE Trans. Broadcast.*, 2019.

[133] J. Zhao, O. Simeone, D. Gunduz, and D. Gómez-Barquero, "Non-orthogonal unicast and broadcast transmission via joint beamforming and ldm in cellular networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–6.

[134] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May. 2010.

[135] A. L. Yuille and A. Rangarajan, “The concave-convex procedure,” *Neural Computation*, vol. 15, no. 4, pp. 915–936, Apr. 2003.

[136] T. X. Vu, S. Chatzinotas, and B. Ottersten, “Edge-caching wireless networks: Performance analysis and optimization,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827–2839, Apr. 2018.

[137] T. X. Vu, L. Lei, S. Chatzinotas, B. Ottersten, and T. A. Vu, “Energy efficient design for coded caching delivery phase,” in *2019 3rd International Conference on Recent Advances in Signal Processing, Telecommunications Computing (SigTelCom)*, Mar. 2019, pp. 165–169.

[138] R. Kannan and C. L. Monma, “On the computational complexity of integer programming problems,” in *Optimization and Operations Research*, R. Henn, B. Korte, and W. Oettli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1978, pp. 161–172.

[139] F. Rinaldi, F. Schoen, and M. Sciandrone, “Concave programming for minimizing the zero-norm over polyhedral sets,” *Computational Optimization and Applications*, vol. 46, no. 3, pp. 467–486, Jul. 2010. [Online]. Available: https://doi.org/10.1007/s10589-008-9202-9

[140] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” http://cvxr.com/cvx, Mar. 2014.