Imperial College London
Department of Electrical and Electronic Engineering

Fundamental Limits of Robust Interference Management: From Content-Oblivious to Content-Aware Wireless Networks

Enrico Piovano

2019

Supervised by Dr. Bruno Clerckx

Submitted in part 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

Statement of Originality

I declare that this thesis is the result of my own work. Information and ideas derived from the work of others has been acknowledged in the text and a list of references is given in the bibliography. The material of this thesis has not been and will not be submitted for another degree at any other university or institution.

Copyright Declaration

In loving memory of my dear father Elio Piovano, you will forever remain in my heart.

Abstract

In this thesis we progress towards the understanding of the fundamental limits of wireless networks with partial instantaneous channel state information at the transmitter (CSIT). We first consider classical content-oblivious networks, where the edge-nodes are unaware of the kind of requested content. We study the K-user multiple-input-single-output (MISO) broadcast channel (BC), where a K-antenna transmitter serves K single-antenna users, and we characterize the optimal degrees-of-freedom (DoF) region under arbitrary CSIT levels for the users. We then study the overloaded MISO BC with two groups of CSIT qualities. We propose a transmission scheme where no CSIT codewords are superimposed on top of spatially-multiplexed codewords. We show that the developed strategy outperforms the existing schemes and achieves the entire DoF region.

Next, we move from content-oblivious networks to content-aware networks, where the edge-nodes can predict the most popular content. We first consider the K-user cache-aided MISO BC, where users are equipped with a cache memory. For a symmetric setting, in terms of channel strength levels, partial channel knowledge levels and cache sizes, we characterize the sum-generalized-degrees-of-freedom (sum-GDoF) up to a constant multiplicative factor of 12. We further show that the characterized order-optimal sum-GDoF is also attained in a decentralized setting, where no coordination is required for content placement in the caches. We then study the cache-aided interference channel, where an arbitrary number of cache-equipped transmitters serve an arbitrary number of cache-equipped receivers. Transmitters communicate with receivers over two heterogenous parallel subchannels: one with perfect CSIT, and the other with no CSIT. Under the assumptions of uncoded placement and separable one-shot linear delivery over the two subchannels, we characterize the optimal sum-DoF to within a constant multiplicative factor of 2. We extend the result to decentralized setting, and we characterize the optimal one-shot linear sum-DoF to within a factor of 3.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Dr. Bruno Clerckx, for the constant help, support and guidance during my PhD. Bruno is an excellent supervisor and he always puts the student's interest first and the scientific outcome second. This has been fundamental for me to learn many new areas outside my main PhD track and understand what I really wanted to do in my life. I would also like to express a special thanks to Hamdi, both on a professional side for his impeccable help and guidance, and on a friendship side for supporting me in the most stressful moments. I would like to thank you Morteza, as a colleague in my research group but also, more importantly, as a wonderful friend. I would like to extend my sincere gratitude to all my other friends and colleagues from Bruno's research group.

Outside of my research group, I have to thank you Thanos for our coffee breaks together and infinite discussions. Faheem, no worry, I have not forgotten you. I have to thank you for all your support and friendship you gave me during all this time. I would like to thank you John for the wonderful time we spent together to solve brain teasers and math problems. I would like to extend my sincere gratitude to my friends and colleagues from the office and from the department: Tricia for our philosophical discussions, Maxime for the boundless entropy, Wilhelm, Alex, Antonio, Metin, Roxana, Tiffany, Ludovico, Michael, Charlie, Jun-Jie, Mohammud, Nitish, Peppe, Chris, David, Mudasar, Borzoo, Giulio, Pouria, Harry, Edmund, Steph, Longfei, Li, Jiabo, Mohamed A., Mohamed M., Ming, Edmund and Lina. I am also truly thankful to all the people from the Communications and Signal Processing group and from the Electrical Engineering department: all of you helped me in one way or another during these years. Outside of work I would like to say a warm thank you to my friends Sharon, Valentini, Josu, Ali, Carlos, Jonas, Iraklis, Emilios, Irene, Sophie, John and Peter for supporting me all over these years. Of course, a special dedication to the Italian friends Dorian, Federico, Alberto, Clara, Francesca and Giorgia.

Vorrei ringraziare tutta la mia famiglia per il supporto incondizionato anche a migliaia di kilometri di distanza. Un particolare ringraziamento a mia Madre, a mio fratello Francesco, a mia Nonna, a mio zio Fredino, a mia zia Adriana, a mia zia Emanuela e ovviamente a voi, i miei cuginetti: Angelica (alias Mimma), Francesca, Chiara e Alessandro. A tu Papa, a tu nonno Stefano, a tu nonno Enrico, a tu nonna Angela e a tu zio Stefano che, anche se non fisicamente qua, sono sicuro starete festeggiando insieme questo traguardo. Un grazie enorme a tutti i miei amici di Fossano: Luca, Andrea, Samuele, Mattia, Enrico e ai miei amici di Torino tra i quali ricordo Luca, Umberto, Antonio, Gianluca, Ruggero, Steve, Michele, Federico, Daniele. E infine, un grazie particolare a te Anastasia per tutto il supporto incondizionato in questo lungo periodo. Grazie a tutti!

Enrico Piovano
London, July 2019.

Contents

List of Figures

1. Introduction

Wireless communications have completely revolutionized our lives and our society. Laptops, smartphones and tablets have become an integrated part of our daily routine and no one can imagine a world without them anymore. However, while the world has been already completely penetrated by the advent of communication technologies, the wireless revolution is just at the beginning. In fact, the fifth generation (5G) of mobile communication systems, in addition to supporting the traffic generated by mobile communications, has been envisioned to further enable the connection of billions of devices coming from the most different applications: Internet of things (IoT), sensor networks, smart grids, etc. In particular, a 1000-fold data traffic increase has been predicted by 2020 [1–4]. Nonetheless, supporting all these devices is very challenging as the limited bandwidth resources of the current networks alone cannot accommodate all the generated data traffic. In order to overcome the bandwidth problem, academic and industrial studies have focused their research and development on the design of new technologies in order to accommodate all the data traffic and reduce the interference due to so many devices connected simultaneously [4].

Multiantenna Wireless Communication Systems

One of the most developed and powerful technology in order to combat interference is the multiantenna technology, where a transmitter, equipped with multiple antennas, can serve multiple users simultaneously. This technology, which is already standardized and implemented in the current generation of wireless communication systems, has become even further an inevitable necessity to meet the requirements of future wireless networks [5–7]. In fact, multiantenna systems exploit the spatial dimensions of the wireless channel through multiuser-multiantenna techniques in order to increase the capacity of wireless networks, as multiple antennas can help to deliver multiple streams of data simultaneously. Such property is captured by the so-called spatial multiplexing gain, which can be roughly defined as the number of streams which can be simultaneously multiplexed over the channel. The spatial multiplexing gain is re-branded as Degrees of Freedom (DoF) when the signal-to-noise-ratio (SNR) is let go to infinity and the system becomes interference-limited. Among the many references on spatial multiplexing, we recall [5, 7–17] and references therein.

It is well established that achieving such spatial-multiplexing gains is highly dependent on the availability of accurate channel state information at the transmitter (CSIT) [18–21]. Since highly accurate CSIT is not always guaranteed, initial studies and deployments strived to apply multi-antenna schemes that assume perfect CSIT to scenarios with imperfect CSIT [5]. However, recent breakthroughs in the study of the DoF unveiled that such approach is fundamentally flawed as it fails to achieve the information-theoretic limits of the channels [13, 22]. On the other hand, insights drawn from such fundamental works have proved to be very promising for the design of future wireless networks [23]. In fact, the DoF framework has provided important insights towards the characterization of the capacity of wireless networks with different kinds of CSIT deficiencies. In particular, as the DoF analysis is performed in the interference-limited regime, deriving the DoF limits is intimately related to the problem of characterizing the fundamental limits of robust interference management, i.e. designing optimal robust interference management strategies, where robustness is intended with respect to the partial CSIT.

From Content-Oblivious to Content-Aware Networks

While the use of multiple antennas can tremendously increase the performance of wireless communication networks and it has been one of driving factor behind the success of the fourth generation of wireless networks (4G), it is well understood that additional solutions are needed in order to satisfy the massive demand of content expected for the 5G networks. One of the most promising directions to solve this problem is given by *caching*, which is the possibility to store a fraction of the most popular content across the edge-nodes of the network in order to reduce both the backhaul cost and the data traffic over the access channel [24–29]. The applicability of caching is mostly driven by three important factors: 1) the nature of content-oriented traffic, as video-on-demand services, which is largery predictable, and 2) the ubiquity of memories and data storage devices, 3) the temporal variability of the traffic which allows the edge-nodes to store the most popular content during the off-peak times in order to reduce the traffic of the network during the peak times.

While caching has been originally developed in the context of networking systems, a significant effort has been recently made to integrate the caching setup in wireless networks [25, 28, 30–33]. In particular, by assuming that a content library of most popular files could be predicted and that each edge-node (base station or user) could be equipped with a cache memory able to store a fraction of the library, a great deal of research has focused on characterizing the performance limits of these cache-aided networks. As wireless networks are intrinsically affected by CSIT inaccuracies, it is natural that many works have taken CSIT imperfections into account and studied the fundamental limits of robust cache-aided interference management [25, 27, 30–32, 34–39].

In this manuscript we make progress towards the understanding of the fundamental limits of robust interference management for different kinds of wireless networks with imperfect CSIT. In particular, as we will see in the next section, we consider a specific type of imperfect CSIT which is the partial instantaneous channel state information, where the transmitter(s) has corrupted instantaneous estimates of the channels of the users. We will start by considering content-oblivious networks in Chapter 2 and 3, where no popular content can be stored in advance by the edge-nodes of the network (so, no caching is taken into consideration). Then, we will move to content-aware networks in Chapter 4 and Chapter 5, where caching is taken into consideration and the edge-nodes are equipped with memories where they can store a fraction of the content library.

1.1. Robust Interference Management for Content-Oblivious Networks

As aforementioned, in Chapters 2 and 3 we study classical content-oblivious wireless networks with imperfect CSIT, where no caching is taken into consideration. Note that many forms of CSIT inaccuracies have been considered in the literature, such as perfect delayed CSIT starting from the seminal work in [40], partial instantaneous CSIT [14, 15, 17, 22, 41–46] or hybrid settings with both delayed and corrupted instantaneous CSIT [13, 47, 48]. In this work we focus on the case where the transmitter has a partial knowledge of the instantaneous channel of the users.

In the current generation of wireless networks, such as Long Term Evolution (LTE), there are two different ways the transmitter can acquire the channel state information (CSI) of the users. In the the first and most used way, called Frequency Division Duplexing (FDD), the users estimate their CSI using pilot symbols, and the estimated CSI are then quantized and reported to the transmitter over a standardized number of bits. In the second way, denoted as Time Division Duplexing (TDD), CSI is measured in the uplink by the transmitter and used in downlink by assuming uplink-downlink reciprocity. It is clear that in both cases, and in particular in the mostly used FDD case, the transmitter can only get an approximate knowledge of the CSI. Note that the difference between the estimate and the real value of the channel is denoted as channel estimation error.

The partial instantaneous CSIT has been widely studied in the information-theoretic literature [15, 17, 22, 41–46]. While studying the capacity limits of wireless network with partial instantaneous CSIT is often intractable with the known information-theoretic techniques, many works have made significant contribution towards the characterization of the fundamental limits of these networks in interference-limited frameworks such as the DoF or the Generalized Degrees of Freedom framework (GDoF) frameworks. We focus here on the multiple-input-single-output (MISO) broadcast channel (BC), where a multiantenna transmitter communicates with multiple single-antenna users (or receivers). Note that the transmit antennas in the considered setup are not necessarily physically co-located, and may generally represents radio heads (or remote antennas) connected through a strong fronthaul. This setting is also denoted as full transmitter cooperation.

It is well known that zero-forcing can achieve full DoF for the perfect CSIT case. While considering zero-forcing precoding scheme and partial instantaneous CSIT, important results were found in [19]. It was shown in [19] that full multiplexing gain can still be maintained if the variance of the channel estimation error at the transmitter scales as $O(\text{SNR}^{-1})$ as the SNR grows infinitely large [19, 21]. On the other hand, it was shown in [19] that if the number of feedback bits scales less than the logarithm of the SNR, which corresponds to the case where the variance of the channel estimation error scales as $O(1)$ with the SNR, for instance in case of a constant number of feedback bits, the achievable rate saturates at high SNR and this corresponds to zero DoF.

By considering more general schemes than zero-forcing, a great deal of research has made effort towards characterizing robust information-theoretic sum-DoF upper-bounds, while assuming the variance of the channel estimation error of the users to scale in general as $O(\text{SNR}^{-\beta})$ for some $\beta \in [0, 1]$. In particular a problem proposed in [49], which remained open for nearly one decade, conjectured the total collapse of the sum-DoF to 1 under finite precision CSIT for all users, i.e. $\beta = 0$ for all users. This conjecture was only recently proved correct by the seminal work of Davoodi and Jafar in [22]. This in turn implies that final precision CSIT is as (un)useful as no CSIT from a DoF perspective. On the other hand, it was shown that a CSIT quality $\beta \in (0, 1)$ helps to save some of the spatial multiplexing gains and to achieve a sum-DoF greater than 1.

While zero-forcing is in fact sub-optimal for $\beta < 1$, a scheme which has been proved to better tackle the interference originated by partial instantaneous CSIT is rate-splitting, where a common codeword decoded by all the users is delivered on top of private codewords intended for the specific users only [13, 15, 44–46]. Rate-splitting finds its roots in a fundamental technique in the information theoretic literature called superposition coding [9]. In the next paragraph we will first start by describing the historical context behind superposition coding. On the basis of this explaination, we will then introduce rate-splitting and the main intuition behind its application to the MISO BC with partial CSIT.

1.1.1. Superposition Coding and Rate-Splitting

As known, the broadcast channel refers to the setup where a single transmitter sends independent information to uncoordinated receivers through a shared medium. From an information-theoretic perspective, the first definition of broadcast channel was given in the seminal work by Thomas Cover in [50]. This opened the door of one of the most important area of research in Information Theory which was then followed-up by many researchers. Among the main works, we recall [9, 10, 12, 50–57]. One of the main problems related to the broadcast channel is the characterization of the (information-theoretic optimal) sum-rate or, even more importantly, the capacity region. We recall that the capacity region is the set of all simultaneously achievable user rates. In the context of a single-antenna transmitter, i.e. the single-input single-output broadcast channel (SISO BC), the sum-rate is achieved by serving the strongest user only. However, while considering the capacity region, techniques which allow to serve all users have to be taken into consideration. Among them there is time-division multiplexing, also called as TDMA, which is a technique where different users are served in different slots, and in each slot a single user only is served. Another important technique was formally introduced by Thomas Cover in [50] and referred as superposition coding [58]. In superposition coding, differently from time-division multiplexing, users are simultaneously served by superimposing their codewords in the power domain. Each user decodes all the codewords from the one transmitted with the highest power to its own, where the decoding is performed at decreasing order of power level.

One the main results in [50] was the proof that superposition coding outperforms time-division multiplexing. Moreover, by considering a degraded channel, where degraded means that the users can be ordered from the strongest to the weakest, as the SISO BC where the users can be ordered from the strongest to the weakest on the basis of their channel strengths, superposition coding achieves the entire capacity region [58]. This is obtained by superimposing the codewords for different users from the strongest to the weakest with increasing power levels, i.e. the codeword of a weaker user is on top of the codeword of a stronger user. Each user decodes then all the codewords, starting from the codeword intended for the weakest user until its own codeword. Hence, the weakest user only decodes its own codeword while the strongest user decodes all the codewords. With this technique is in fact possible to achieve the entire capacity region of the SISO BC, while time-division multiplexing can in fact only achieve a fraction of it. Moving from the SISO BC to the MISO BC with perfect CSIT adds extra challenges to the problem as the latter is a non-degraded broadcast channel. For the MISO BC with perfect CSIT the sum-DoF and the capacity region have only been recently characterized [10,51–53] by utilizing a technique called dirty paper coding [59], a different scheme from superposition coding where the interference from other users are pre-cancelled at the transmitter. Afterward, dirty paper coding has also been shown to achieve the capacity region of the multiple-input multiple-output broadcast channel (MIMO BC) with perfect CSIT [12].

From Superposition Coding to Rate-Splitting

Moving from the broadcast channel, another very important setup studied in the information-theoretic literature is the interference channel (IC). In the interference channel, multiple transmitters aim to deliver their separate messages to multiple of receivers through a common channel, and each transmitters aims to send its message to a specific receiver. As in the interference channel there is no cooperation between transmitters and receivers, the codeword sent by each transmitter to its corresponding receiver generates interference to the other receivers (see, for instance, [60–65] and references therein).

The idea of rate-splitting dates back the studies of the interference channel by Carleial in [60] and by Han and Kobayashi in [61]. Let us consider the work in [61]. In their paper, Han and Kobayashi considered the two-user single-input single-output interference channel (SISO IC). Each transmitter splits its message into a common and private part. Each receiver jointly decodes the two common messages and its intended private message. The main idea behind this scheme is the fact that decoding part of the interference (in the form of common message) can enhance the performance. We can now see how this scheme finds its roots in superposition coding. With a proper power allocation for the private and common codewords, the Han and Kobayashi scheme leads to an achievable region which is to within 1 bit to the capacity region, and this is the best achievable rate region to date [66].

Now, let us consider the MISO BC with partial instantaneous CSIT. For simplicity, we consider the setup of the two-user case and zero-forcing transmission strategy. In case of perfect CSIT, the transmitter can deliver the codeword to the intended receiver by removing interference to the unintended receiver by simply choosing a precoding vector for the codeword which is orthogonal to the channel vector of the unintended receiver. However, in case of partial instantaneous CSIT, the transmitter only knows an estimate of the channel of the users. Hence, it is not provided with enough information to design a precoding vector which is perfectly orthogonal to the channel of the unintended receiver. It follows that the codeword intended for a specific receiver generates interference at the unintended receiver, as in the IC. We can then see the similarity between the two-user MISO BC with partial CSIT and the two-user IC. This motivates the application of rate-splitting, i.e. the idea of decoding part of the interference in the form of common message, to the MISO BC with partial CSIT.

As already mentioned earlier, the characterization of the capacity limits of wireless networks under partial CSIT seems beyond the capabilities of the known information-theoretic techniques. This motivates the use of capacity approximations such as the DoF. Rate-splitting was first introduced in the DoF framework in the context of the MISO BC with imperfect CSIT in the work in [13]. Note that, before [13], the idea of a layered transmission for the broadcast channel to deal with imperfect CSIT was described in [67–70], where the transmitter performs multi-layered coding, which is the essence of the broadcast approach [67]. Coming back to the work in [13], the achievable sum-DoF by rate-splitting for the two-user case was characterized. In [45, 46] the result was generalized for the K-user case. These achievable sum-DoF were then proved to be the optimal ones thanks to the sum-DoF upper-bound derived in [22]. A main result was found in [16], where it was shown that rate-splitting achieves the optimal DoF region of the K-user MISO BC with partial CSIT. This will be the topic of Chapter 2. Note that many other works have considered rate-splitting for the MISO BC with partial CSIT, for instance we recall [14, 15, 17, 23, 44–46, 71] and references therein.

In the next section, we introduce the system model and the DoF definition for the MISO BC with partial instaneous CSIT which will be utilized in Chapters 2 and 3. The system model will be then specialized for the specific settings considered in each chapter.

1.1.2. MISO BC with Partial CSIT

In the general MISO BC a transmitter equipped with K_{T} antennas communicates with K_{R} single-antenna users (or receivers). The users are indexed by the set $\mathcal{K}_{\mathrm{R}} = \{1, \dots, K_{\mathrm{R}}\}$. Often K_{T} is assumed equal to K_{R} and in this case the setup is simply referred as the K-users MISO BC, where a K-antenna transmitter serves K single-antenna users (the subscripts referring to the transmitter or receivers are omitted). The communication between the transmitter and the receivers lasts for T channel uses, where T can grow infinitely large. The input-output relationship at the t-th use of the physical channel, also denoted as t-th channel-use, $t \in [T]$, is modeled by

$$Y_i(t) = \sum_{j=1}^{K_{\mathrm{T}}} G_{ij}(t) X_j(t) + Z_i(t) \tag{1.1}$$

where $Y_i(t) \in \mathbb{C}$ is the signal received by the i-th receiver, $X_j(t) \in \mathbb{C}$ is the signal transmitted from antenna j, $G_{ij}(t)$ is the time-varying fading channel coefficient between transmit antenna j and receiver i and $Z_i(t) \sim \mathcal{N}_{\mathbb{C}}(0, 1)$ is the normalized additive white Gaussian noise (AWGN), which is i.i.d. across all dimensions. All the signals and channel coefficients are complex. We denote the transmitted signal across all transmitting antennas at the t-th channel use as $\mathbf{X}(t) \triangleq [X_1(t) \cdots X_K(t)]^{\mathrm{T}}$. The transmitter is then subject to the power constraint $\frac{1}{T}\sum_{t=1}^{T} |\mathbf{X}(t)|^2 \leq P$. To avoid degenerate situations, we assume that the instantaneous value $|G_{ij}(t)|$ is bounded away from zero and infinity for all $i, j \in [K_{\mathrm{R}}] \times [K_{\mathrm{T}}]$, and $t \in [T]$.

As we will see later, the DoF metric is defined by considering the ergodic-rate. Hence, as we consider the ergodic-rate, we assume that the channel varies over time. In particular, the channel can vary over each channel use or in blocks, where each block spans over a finite number of channel uses where the channel is constant [45,46].

Partial Instantaneous Channel Knowledge

As already mentioned earlier, acquiring an accurate partial instantaneous CSIT is very challenging in practical systems. For instance, in the current wireless networks, the users estimate their CSI using pilot symbols, and the estimated CSI are then quantized and reported to the transmitter over a standardized number of bits. It is clear that this leads to a knowledge of the channel by the transmitter which is not perfect. We model the partial instantaneous CSIT by following the classical assumption made in [13–16,22,46,72,73]. In particular, by denoting as $G_{ij}(t)$ the channel between the j-th transmitting antenna and the i-th receiver, we assume that the variance of the channel estimation error scales as $O(\mathrm{SNR}^{-\beta_i})$, where β_i is the CSIT quality level of user i, as already explained earlier. Hence, we can write $G_{ij}(t) = \hat{G}_{ij}(t) + \sqrt{\mathrm{SNR}^{-\beta_i}}\tilde{G}_{ij}(t)$, where $\tilde{G}_{ij}(t)$ is the estimation error term.

Note that this definition can be linked to the number of feedback bits as described in [19], where it was assumed that each user performs vector quantization on its channel realization using random quantization codebooks, also called as random vector quantization [74,75]. In particular, following [19], a variance of the channel estimation error which scales as $O(\mathrm{SNR}^{-\beta_i})$ corresponds to the case where the number of feedback bits for user i scales linearly with the logarithm of 2 of the SNR at rate $\beta_i(K_\mathrm{T} - 1)\log_2 \mathrm{SNR}$, where K_T is the number of transmitting antennas. Hence, by considering the case $\beta_i = 0$, it is clear that if the number of feedback bits is kept constant, the variance of the channel estimation error scales with the SNR as $O(1)$. As shown in [19] for the zero-forcing case and generalized later in [22] for any transmission scheme, $\beta_i = 0$ corresponds to no CSIT from a DoF sense. Hence, as shown in [19], if $\beta_i = 0$ for all users, considering a zero-forcing transmission strategy the sum-rate is bounded as the SNR goes to infinity. On the other hand, the case $\beta_i = 1$ corresponds to the case where the number of feedback bits scales with the logarithm 2 of the SNR at rate $(K_\mathrm{T} - 1)\log_2 \mathrm{SNR}$. As already mentioned aboved and shown in [19] by considering zero-forcing, this corresponds to perfect CSIT in a DoF sense.

We can mathematically formalize what just explained in the following. Let $\mathcal{G} \triangleq \big\{G_{ij}(t) : i,j \in [K_\mathrm{R}] \times [K_\mathrm{T}],\ t \in [T]\big\}$ be the set of all channel coefficient variables. Under partial CSIT, and considering that the SNR is equal to P given the assumption on the unitary variance of the noise, the channel coefficients are modeled as

$$G_{ij}(t) = \hat{G}_{ij}(t) + \sqrt{P^{-\beta_i}}\tilde{G}_{ij}(t) \tag{1.2}$$

where $\hat{\mathcal{G}} \triangleq \big\{\hat{G}_{ij}(t) : i,j \in [K_\mathrm{R}] \times [K_\mathrm{T}],\ t \in [T]\big\}$ are channel estimates, $\tilde{\mathcal{G}} \triangleq \big\{\tilde{G}_{ij}(t) : i,j \in [K_\mathrm{R}] \times [K_\mathrm{T}],\ t \in [T]\big\}$ are estimation error terms and $\beta_i \in \mathbb{R}$ is a parameter capturing the CSIT quality level for receiver i. The channel knowledge available to the transmitters includes the CSIT quality levels β_i and the estimates in $\hat{\mathcal{G}}$, but does not include the error terms in $\tilde{\mathcal{G}}$.

All variables in $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$ are subject to the bounded density assumption as explained in [22, 41]. The difference between $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$, as pointed out earlier, is that the former is revealed to the transmitters while the latter is not. Hence, given the estimates $\hat{\mathcal{G}}$, the variance of each channel coefficient of receiver i in $\mathcal{G}$ behaves as $\sim P^{-\beta_i}$ and the peak of the probability density function behaves as $\sim \sqrt{P^{\beta_i}}$. As already mentioned, we assume that $\beta_i \in [0, 1]$. In particular, $\beta_i = 0$ and $\beta_i = 1$ capture the two extremes where channel knowledge at the transmitters is absent and perfectly available, respectively, in a DoF sense. Please keep in mind that $\beta_i = 0$ is also denoted as finite precision CSIT. On the other hand, note that $\beta_i > 1$ would not lead any benefit compared to $\beta_i = 1$. Before we proceed, it is worth highlighting that channel state information at the receivers (CSIR) is always assumed here to be perfect.

Degrees of Freedom (DoF) Metric

The transmitter aims to send the independent messages $W_1, W_2, ..., W_{K_\mathrm{R}}$ to the corresponding users. The messages $\{W_i\}_{i=1}^{K_\mathrm{R}}$ are jointly encoded over n channel uses, at the respective rates of $\{R_i\}_{i=1}^{K_\mathrm{R}}$, into a codebook matrix over the input alphabet, where the codebook matrix has size $2^{nR_1+\cdots+nR_{K_\mathrm{R}}} \times n$. We denote the codebook matrix as $Co(P, \{R_i\}_{i=1}^{K_\mathrm{R}}, n)$, where P indicates the power constraint at the transmitter. For a given power constraint P, the rate vector $(R_1, R_2, \ldots, R_{K_\mathrm{R}})$ is denoted as achievable if there exists a sequence of codebooks $Co(P, \{R_i\}_{i=1}^{K_\mathrm{R}}, n)$, indexed by n, such that the probability of all the messages being successfully decoded by the respective receivers goes to 1 as n approaches infinity. Note that, as the transmitter only has a partial knowledge of the CSI, we consider here the *ergodic-rate* as we mentioned earlier. Hence, as explained earlier, we assume the channel to vary in each channel use or in blocks, where each block spans over a finite number of channel uses and in each block the channel is constant. As n approaches infinity, coding is performed across different blocks, where the number of blocks goes to infinity with n. The closure of all the achievable rate vectors $(R_1, R_2, \ldots, R_{K_\mathrm{R}})$ is called capacity region $\mathcal{C}$.

The per-user DoF of user i is defined as the following asymptotic ratio with respect the SNR P, where the notation $R_i(P)$ is used to indicate the dependency of the rate of user i on the SNR P

$$d_i \triangleq \lim_{P\to\infty} \frac{R_i(P)}{\log(P)}. \tag{1.3}$$

The DoF tuple $(d_1, d_2, \ldots, d_{K_\mathrm{R}})$ is said to be achievable if there exists a rate tuple which is given by $(R_1(P), R_2(P), \cdots, R_{K_\mathrm{R}}(P)) \in \mathcal{C}(P)$ such that $d_i \triangleq \lim_{P\to\infty} \frac{R_i(P)}{\log(P)}$ for $i = 1, 2, \ldots, K_\mathrm{R}$, where P is used to highlight the dependency on the SNR P. The closure of all achievable DoF tuples $(d_1, d_2, \ldots, d_{K_\mathrm{R}})$ is called the DoF region, which we indicate as $\mathcal{D}^*$ and we define as follows.

$$\mathcal{D}^* = \left\{ (d_1, d_2, \ldots, d_{K_\mathrm{R}}) | \exists\, (R_1(P), R_2(P), \ldots, R_{K_\mathrm{R}}(P)) \in \mathcal{C}(P), \text{s.t. } \forall i = 1, \ldots, K_\mathrm{R}, d_i = \lim_{P\to\infty} \frac{R_i(P)}{\log(P)} \right\}.$$

The sum-DoF d_Σ is defined as

$$d_\Sigma = \max_{(d_1, d_2, \ldots, d_{K_\mathrm{R}}) \in \mathcal{D}^*} d_1 + \cdots + d_{K_\mathrm{R}}. \tag{1.4}$$

Practical Implications of the DoF Studies

The study of the DoF allows to characterize the optimal number of signalling dimensions, or spatial multiplexing, of wireless networks in the asymptotically high SNR regime. This in turn has allowed to make progress in the characterization of the fundamental limits of wireless networks for which capacity studies seem unfeasible with the known information-theoretic techniques. From a more practical perspective, while the DoF studies can allow to roughly approximate the behaviour of wireless networks in the high SNR regime, caution has to be taken while directly translating DoF results into practical insights. For instance, in order to achieve many DoF results, trivial choices of the precoders such as zero-forcing are indeeded sufficient. Unfortunately, this does not hold true when translating the results to a finite SNR regime scenario, where for instance the kind of precoder can significantly influence the performance of the system.

However, the DoF metric can be very useful to inspire new results, by showing that such results hold in the asymptotically large high SNR regime. New techniques can be then developed to see how these results translate in the finite SNR regime. An example of this is the following. It was shown that in the DoF regime a rate-splitting scheme with private and common messages can strictly outperform in many cases zero-forcing, where only private messages are considered. However, the question was whether this gain translates in something meaningful in the finite SNR regime. The works in [45,46] have shown that, by considering precoder optimization, the DoF gain translates in a significant sum-rate gain at finite SNR regime. This is a DoF inspired result, as the result was first noticed in the DoF regime and then translated in the finite SNR regime.

1.1.3. Main Results of Our Work for Content-Oblivious Networks

DoF Region of the K-user MISO BC with Partial CSIT

In Section 2 we address the problem of characterizing the optimal DoF region of the K-user MISO BC with arbitrary CSIT levels of the users, i.e. each user i has a CSIT quality $\beta_i \in [0, 1]$. While the optimal sum-DoF was established based on the seminal work by Davoodi and Jafar in [22], no attempt has been made to characterize the entire DoF region. Indeed the main result of Chapter 2 is the characterization of the optimal DoF region. To derive this result, we employ a two-steps approach: we first derive a polyhedral outer-bound of the optimal DoF region and we then prove the achievability of this outer-bound. The outer-bound is obtained by applying the sum-DoF upper-bound in [22] to each subset of users. The achievability of the outer-bound turns out to be more challenging and it can be obtained by employing a rate-splitting strategy [13–15,44–46]. Note that rate-splitting was already shown to be able to achieve the optimal sum-DoF.

Conventional methods to show the achievability of a DoF region rely on characterizing and showing the achievability of the corner points, as the achievability of any other point is then obtained by time-sharing over the corner points [76]. However, this method fails to succeed in the considered setup as the number of corner points scales exponentially with the number of users K. In order to overcome this problem, we introduce a novel approach. Instead of characterizing the corner points, we characterize the facets of the polyhedral region. Surprisingly, utilizing mathematical tricks, it is possible to rewrite each facet of the polyhedral region as a set of inequalities which bound the per-user DoF of each individual user. Such characterization is then suitable to show the achievability by rate-splitting [16]. The result in this chapter has been published in:

- E.Piovano, B. Clerckx "Optimal DoF region of the K-User MISO BC with Partial CSIT", *IEEE Communication Letters*, 2017.

DoF Behavior of the Overloaded MISO BC

In Chapter 3 we extend the DoF analysis of Chapter 2 and we make progress towards the understanding of the fundamental limits of the overloaded MISO BC, where the number of users K_{R} is larger than the number of antennas at the transmitter K_{T}. We consider a setup where a group of users with size K_{T} has partial CSIT and the remaining users have no CSIT. The most natural way to serve these users is to utilize an orthogonal time partitioning approach, where the two groups of users are served in two different phases. The group of users with partial CSIT is served utilizing a rate-splitting transmission strategy, while the group of users with no CSIT is served utilizing a no CSIT transmission layer due to the collapse to 1 of the DoF [22]. We propose a non-orthogonal transmission scheme based on power partitioning where the signals carrying the messages for the users with and without CSIT are superimposed and separated in the power domain [17]. Users with no CSIT decode their own codeword first by treating the codewords of the users with partial CSIT as noise. Users with partial CSIT decode the codewords intended for the users with no CSIT first, they remove them and then they finally decode their own codewords. First, we show that such strategy achieves a strict DoF gain over time partitioning. Second, we show that a generalized version of this power partitioning strategy achieves in fact the optimal DoF region for the considered overloaded MISO BC. The result in this chapter has been published in:

- E.Piovano, H. Joudeh, B. Clerckx, "Overloaded MU-MISO transmission with imperfect CSIT", *Asilomar Conference on Signals, Systems, and Computers*, Asilomar, 2016.

1.2. Robust Cache-Aided Interference Management for Content-Aware Networks

While in Chapter 2 and Chapter 3 we deal with content-oblivious networks, the focus of the chapters 4 and 5 is on content-aware networks, where caching is taken into the picture. As the study of content-aware networks, in particular of caching, has been one of the main topics in the information-theoretic literature over the last few years, we start by revisiting the main works and the main consequences. The idea of caching the popular content has been mostly driven by the massive increase in the expected content traffic in the next generation of wireless networks. In particular, thanks to the predictable nature of content-oriented traffic, alongside the temporal variability, nodes across the network can *cache* popular content in their cache memories during off-peak times, in which network resources are under-utilized, and then use this cached content (sometimes in surprisingly novel ways) to alleviate the traffic load of the network during congested peak times, when users are actively requesting content and competing for wireless spectrum [77]. This scheme can lead to a dramatic alleviation in the network load both in the access channel (wireless) and in the backhaul network (wired) [78].

Caching has been extensively deployed in wired networks since late 90s, via the so-called Content Distribution Networks (CDNs) and the utilization of web-caching. These replicate content across different locations of the network, which in turn allows to place popular content close to the users and avoid multiple request of the same data at the content distribution server. Moreover, it also reduces the distance between the users and the location of the content which in turn reduces the latency of the communication. CDNs are, in general, efficient solutions when the local communication link is not the bottleneck of the performance, as for the case of wired networks. However, this limits the direct applicability of CDNs in wireless networks, where the bottleneck is given by the wireless access channel between the access nodes of the communication network (base station, femtocell, WiFi access point, etc.) and the users.

In order to also utilize caching at the edge-nodes of a wireless network, the properties of the wireless access channel have to be taken into consideration to efficiently design and combine storage and transmission strategies. In particular, the broadcast nature of the wireless channel allows to simultaneously deliver useful information for multiple users with a single transmission. On the other hand, access nodes can be equipped with multiple antennas, which in turn allow to leverage the fading nature of the wireles channel to create spatial multiplexing opportunities. Wireless networks which integrate caching are often denoted as *cache-aided wireless networks*.

1.2.1. Information Theoretic Limits of Caching

While this manuscript is mostly related to the application of caching in wireless communication networks, the information-theoretic limits of caching were first established, by the seminal work in [24], in the context of a broadcast network in which one transmitter (server) communicates with multiple users, each of these users equipped with cache memories, over a shared noiseless link. In this section we will revisit the main results and insights from the work in [24].

Placement and Delivery Phases

Traffic over communication networks is highly variable and significantly fluctuates over different times of the day. This leads to an asymmetric utilization of the network resources and, in particular, resources are often underutilized during off-peak times, for instance over night hours, and overloaded during the the peak times, for instance over day hours. This disparity can be exploited by storing the popular content in the memories distributed across the networks (users, base stations, servers, routers, etc.) during the off-peak hours, when the resurces of the network are abundant, in order to reduce the load of the network during the peak hours, when users are actively requesting for content. Caching shifts then a portion of the traffic from the peak hours to the off-peak hours, which in turn allows to reduce the traffic variability over time.

Starting from the above discussion, the work in [24] formulates the caching problem as the succession of two distinct phases: the *placement phase* and the *delivery phase*. The *placement phase* takes place during the off-peak times, when the memories across the networks are filled with the predicted most popular content. The *delivery phase* takes place during the peak hours, when users reveal their demand and actively request for content. The main objective of caching, as presented in [24], is to jointly optimize the placement and delivery phase in order to maximize the performance of the network. Different metric of performance can be used. In the original papers in [24], the utilized metric of performance was the overall amount of information needed to be delivered to the users during the delivery phase. While this metric is common for wired networks, in wireless settings it is more customary to use the delivery time or the DoF, as we will see in the next sections.

Coded-Caching

In single-user systems, the caching gain comes from making part of the content locally available to the user. Such *local caching gain* scales with the cache memory size, and extends to networked systems with no interference, i.e. where each user enjoys a dedicated and isolated communication link. The picture, however, is very different when users share communication links. As aforementioned, this was taken up by Maddah-Ali and Niesen in [24], in the context of a broadcast network setting where a transmitter communicates with multiple users through a shared noiseless link. In addition to the oblivious local caching gains, Maddah-Ali and Niesen revealed a (hidden) *global caching gain* which scales with the aggregate size of user cache memories, despite the lack of cooperation amongst users during transmissions. Such global caching gain is exploited through careful placement of content during the placement phase, creating (coded) multicasting opportunities during the delivery phase, that would not naturally occur otherwise. This in turn allows serving multiple distinct user demands using fewer transmissions.

To formally explain the local and global caching gains and the result in [24], let us consider the mentioned above broadcast network where a transmitter (server) has access to a library of N files, each of size F bits, and communicates with K cache-equipped receivers through a shared, noiseless link. Let us assume that $N \geq K$, i.e. each user can choose a different file. Note that, in [24], also the case $N < K$ was discussed. However, we will not explain this latter case here as, throughput our work, we will always assume $N \geq K$. The cache of each user can store up to M files. The fraction $\mu = \frac{M}{N}$ denotes the normalized cache size, which represents the fraction of the library which each user can store in its cache memory. During the placement phase, the caches of the users are filled as a function of the library. During the delivery phase, each user requests a single file from the library. It is assumed that the requests are independent across users and that each user selects one of the files in a uniform manner. It is readily seen that, as each user can store a fraction μ of each file and by assuming the worst-case scenario where each user chooses a different file, a conventional uncoded scheme leads to a normalized load of the shared link given by

$$K \cdot (1 - \mu) \tag{1.5}$$

where the normalization is with respect the file size F. The factor $1 - \mu$ is the *local caching gain*[1] explained earlier, where the name comes from the fact that it only scales with the size of the local cache of each users.

Surprisingly, it was shown in [24] that there exists another (hidden) gain which scales with the aggregate cache memory of all the users. Such gain is obtained by carefully designing the placement of the content at the caches of the users, in order to create coded multicasting opportunities where multiple demands can be simultaneously satisfied with a single transmission. Such framework is often indicated as *coded-caching* and the derived gain as *coded-caching gain* or *global caching gain*. The application of coded-caching leads to normalized load of the shared link given by

$$K \cdot (1 - \mu) \cdot \frac{1}{1 + K\mu}. \tag{1.6}$$

From Eq. (1.6) it is readily seen that, in addition to the local caching gain already described in Eq. (1.5), coded-caching achieves a surprising further reduction of the load by a factor $1 + K\mu$. The term *global caching gain* comes from the fact that this gain, as already anticipated earlier, scales with overall cache memory of all the users, despite the lack of cooperation amongst users. Note that the idea of coded-caching has been inspired by the fundamentals of network coding, specifically index coding with side information [24].

The placement and delivery strategies designed to obtain the result in (1.6) will be revisited in Section 4.6 in Chapter 4 and Section 5.6 in Chapter 5, as they are essential building blocks of our work.

Information-Theoretic Outer-Bound

While the result in Eq. (1.6) provides an upper-bound of the information-theoretic optimal load of the shared link, an outer-bound has to be established to evaluate how far this upper-bound is from the optimal performance. A lower-bound on the information-theoretic normalized optimal load of the shared link was then also established in [24] and it is given by

$$\max_{s \in \{1,2,\ldots,K\}} \left(s - \frac{sM}{\left\lfloor \frac{N}{s} \right\rfloor} \right)^{+}. \tag{1.7}$$

Moreover, in [24] it was shown that the ratio between the upper-bound in (1.6) and the lower-bound in (1.7) cannot exceed a constant factor for all system parameters, in this case a factor of 12, which in turn guarantees that the result in (1.6) is information-theoretic order-optimal. As a consequence, coded-caching not only dramatically improves the performance compared to an uncoded scheme, but it also attains order-optimal performance. Moreover, it also follows that *local caching gain* and *global caching gain* are the two fundamental gains for the considered setting in [24], i.e. there are no other gains which scale with the system parameters.

[1] Note that the term *gain* here indicates the reduction in the load of the shared link.

Centralized vs Decentralized Placement

The coded-caching framework in [24] was introduced by assuming that the transmitter could design the placement of content in the user caches to create coded multicasting opportunities during the delivery phase. However, this requires the transmitter to know the number and the identity of the users already during the placement phase, setting which is commonly named as *centralized placement*. This setting, while helpful to establish the coded-caching technique, has very limited applications. In fact, in practical networks and more specifically in wireless networks, users enjoy an high-degree of mobility which makes impossible to know the identity of the users already during the off-peak times.

To overcome this problem, a decentralized version of coded-caching was developed in [79], where placement is randomized and hence independent of the identity and number of active users during the delivery phase. This is commonly named as *decentralized placement*. In particular, it was shown that, by letting the users cache a fraction μ of the bits of each file during the placement phase, where such bits have to be chosen uniformly at random, still creates coded multicasting opportunities during the delivery phase. This in turn allows to achieve very close performance to the centralized setting, which limits the loss due to decentralization and makes coded-caching more practical. Moreover, even more surprisingly, in [79] it was shown that decentralized placement still achieves order-optimal performance, hence to within a constant multiplicative gap from the upper-bound in (1.7). To summarize, not only coded-caching applied to decentralized placement comes at low price, but it also still attains order-optimal performance.

In our work, as we are considering wireless networks where an high-degree of mobility of the users has to be taken into account, we will establish the results for centralized placement first and then we will extend to decentralized placement. Hence, more details regarding the decentralized settings will be given in Section 4.7 in Chapter 4 and in Section 5.7 in Chapter 5.

Extensions

The results in [24] have been improved and extended in a number of directions. From a more fundamental perspective, effort has been made in order to design better achievable schemes [80–84] and/or characterize tighter lower-bounds [85–89]. For instance, among the many works which have improved the existing lower-bounds, we recall the one in [89], which has tightened the order-optimal result in [24] from a factor 12 to a factor 2 for all system parameters. On the other hand, the work in [88] has shown that the achievable scheme in [79] is indeed optimal for decentralized and uncoded placement.

Another line of research has instead focused on extending the work in [24] to more general settings and scenarios, while still maintaining the setup of a broadcast network with error-free links. For instance, the works in [90–93] have considered non-uniform popularity of the files, the work in [94] has considered caches of the users with different size and the works in [95, 96] has considered the case where each user requests for multiple files. Other extensions of coded-caching in more general settings include hierarchical coded-caching [97–99], multi-library coded caching [82] and multi-server wired (noiseless) networks [100].

Moving beyond broadcast networks with error-free links, a great deal of research has also aimed to extend the caching framework by taking into consideration the impairments of the channel and the noise of the links. In particular, we recall the extension of caching in the context of noisy broadcast networks [101–105] and wireless networks. Regarding wireless networks, this fundamental approach to caching was extended in a number of scenarios: wireless device-to-device (D2D) networks [27], wireless interference networks with caches at the transmitters only or at both ends [25,30,31,34,106–109], multi-antenna wireless networks under a variety of assumptions regarding the availability of transmitter channel state information (CSIT) [32,36–39,72,110,111,111–114], and fog radio access networks (F-RANs), in which a cloud processor connects to edge nodes through front-haul links, under different assumptions and settings [109, 115–120]. Some of the progresses in the coded-caching framework above have been surveyed in [26], in which challenges and open problems are also discussed.

1.2.2. Cache-Aided Interference Management

The capacity of wireless networks is one of the longest standing open problems in network information theory. The intractability of the problem, in its generality, motivated the use of capacity approximations, e.g. the DoF metric as already explained in Section 1.1.2. A more general metric also utilized for capacity approximation is given by the Generalized Degrees of Freedom (GDoF), which extends the DoF by taking into consideration the difference in the link strengths. However, most of the results about the GDoF are known for symmetric settings for the link strengths to avoid an explosion in the system parameters, which is the main reason why the GDoF is not utilized in Chapters 2 and 3. The introduction of such metrics allowed significant progress in capacity studies. Since incorporating caches adds an extra layer of complexity to the network, it is not surprising to see that the utilization of the above approximations is inherited by works studying cache-aided wireless networks. Examples of such studies in different scenarios are given in [25,27,30–32,34–39].

Another metric commonly used in cache-aided wireless networks is the so-called *normalized delivery time* (NDT), which is defined as the amount of time needed to conclude the delivery phase for any user demand [106, 110, 115, 116]. Note that the NDT and the DoF metrics are intimately related and the DoF is usually defined as the reciprocal of the NDT (based on the definition, the local caching may need to be included in the expression).

Amongst the main insights derived from the above studies of cache-aided networks in the DoF framework is that caching at the transmitters creates interference alignment and zero-forcing opportunities, enabled through partial and full transmitter cooperation. For example, interference channels start resembling X channels and eventually turn into multi-antenna broadcast channels [25,30, 31, 34, 106–109]. On the other hand, caching at receivers creates coded-multicasting opportunities, which are particularly useful in scenarios where spatial degrees of freedom cannot sufficiently create parallel interference free links. For example, coded-multicasting gains are pronounced in multi-antenna broadcast channels with more receivers than transmitting antennas [31, 113] and/or where channel state information at the transmitter (CSIT) is imperfect [32,36,37,110].

From Cache-Aided to Robust Cache-Aided Interference Management

As seen in the previous sections, the study of the fundamental limits of interference management for classical networks has been pushed as far as characterizing entire DoF or GDoF regions, or even capacity regions for specific cases. However, given the extra challenges arising by further introducing caching into the picture, the studies of the fundamental limits of interference management for cache-aided wireless networks have mostly focused on characterizing (order) optimal sum-DoF or sum-GDoF.

Over the recent years, a significant progress has been made in the study of cache-aided interference management strategies for different kind of networks and setups. By considering perfect CSIT at the transmitters, order-optimal results have been derived for a plethora of network settings. For the K-user cache-aided MISO, the optimal NDT was derived in [110]. For cache-aided wireless interference networks, where multiple transmitters serve multiple users and each transmitter can store a fraction of the library (while the receivers have no caches), order-optimal NDT for different network configurations were derived in [107,115,121]. For cache-aided wireless interference networks with both caches at the transmitters and receivers, an order-optimal DoF was first established by restricting to one-shot linear delivery schemes with uncoded placement in [31]. The order-optimal information-theoretic DoF was instead then characterized by the works in [34, 106, 107, 109]. Another interesting line of research has focused on characterizing optimal interference management strategies for F-RAN networks [109, 115–120]. The following paper provides an overview and new results on the topic [122]. For instance, while considering single-antenna transmitters, an order-optimal NDT was obtained in [115]. The case where each transmitter is equipped with multiple antennas has then been considered in [118] where, under the assumption of one-shot linear strategies and uncoded placement, an order-optimal NDT has been derived. While all these works have assumed centralized placement, extensions to decentralized setting have been made in [107–109, 119, 120].

The assumption of perfect CSIT has helped to derive optimal cache-aided interference management strategies for all the aforementioned setups, however more challenges arise in the design of robust interference management which takes into consideration imperfect or partial CSIT. The work in [110] was among the first to consider the combination of coded-caching and robust interference management for partial CSIT, in the context of the K-user cache-aided MISO BC. As a main result it was shown that, thanks to benefits of coded-caching, the same NDT as perfect CSIT can be obtained by a relaxed instantaneous CSIT quality β up to a certain threshold.

The fundamental limits of robust cache-aided interference management were first established by assuming partial instantaneous CSIT, modelled by a CSIT quality $\beta \in [0, 1]$, as well as (perfectly accurate) delayed CSIT. In particular, for the K-user cache-aided MISO BC with perfect delayed and partial instantaneous CSIT, optimal robust interference management strategies were established in [32], where an order-optimal sum-DoF (up to 4 factor from the optimal one) was obtained. While [32] assumed the collective cache memory of all users be able to store the entire library, this condition was then relaxed in the follow-up work in [123]. Note that the further assumption of delayed CSIT in [32, 123], in addition to partial instantaneous CSIT, was crucial to obtain the converse, as it allowed to leverage the proofs and the results in [124] to obtain a lower-bound on the optimal NDT. More details will be given in Section 4.5.

The K-user cache-aided MISO BC with partial instantaneous CSIT only was instead first studied in the works in [36, 37], where achievable schemes where proposed without any guarantee of optimality. The fundamental limits of robust interference management for the K-user cache-aided MISO BC with partial instantaneous CSIT were finally obtained in [72], where a robust outer-bound and in turn an order-optimal sum-GDoF were derived. The converse required a novel adaption of the aligned image set (AIS) technique developed in [22] in the context of cache-aided settings, and the details are referred to Section 4.5. The work in [72] will be the main content of Chapter 4.

Extending from the cache-aided MISO BC to cache-aided wireless interference networks, the work in [36] provided achievable GDoF for the no CSIT case. The partial CSIT case was instead considered in the work in [38]. However in this work, differently from all the aforementioned cases, the partial CSIT has been modelled by considering that the transmitters have perfect CSIT over a fraction of the bandwidth and no CSIT in the remaining fraction. By focusing on separable one-shot linear delivery strategies with uncoded placement, the work in [38] has established the order-optimal sum-DoF of this network. Note that the work in [38] will be the main content of Chapter 5.

While order-optimal results are usually obtained by assuming centralized placement, the works [38, 72] have shown, for the considered settings, that decentralized placement still attains order-optimal performance which in turn extends the insights in [79] to cache-aided wireless networks with partial instantaneous CSIT. The details, which include the decentralized achievable schemes as well the order-optimality proofs, will be given in Chapters 4 and 5. To conclude, we want point out that imperfect CSIT, in particular delayed CSIT, has also been considered for F-RAN networks [117].

1.2.3. Main Results of Our Work for Content-Aware Networks

As already anticipated earlier, in Chapters 4 and 5 we make further progress towards the understanding of the fundamental limits of robust cache-aided interference management strategies for cache-aided wireless networks with partial CSIT.

Fundamental Limits of Robust Cache-Aided Interference Management Under Full Transmitter Cooperation

We first consider in Chapter 4 the K-user cache-aided MISO BC, where a K-antenna transmitter serves K single antennas users. The transmitter has a partial instantaneous CSIT of the users. Each of users is equipped with a cache memory where it can pre-store part of the content. The transmitter has access to the entire library and each user requests a specific file in the library during the delivery phase. We assume a symmetric setup where all the cross-links (and the direct-links) have the same strengths. Given the symmetric setup, we utilize as a metric the Generalized Degrees of Freedom (GDoF) framework, which generalizes DoF framework to take into consideration the differences in path-loss between the cross-links and the direct-links, as it will be described in Section 4. Note that in Chapter 2 and 3 we utilizes the DoF framework since, as we will see next, the GDoF analysis for asymmetric setups becomes much more challenging and often intractable given the explosion in the number of system parameters.

In this setup, we first characterize the optimal sum-GDoF up to a constant multiplicative factor, which is independent of all system parameters. This order-optimal sum-GDoF characterization is derived while considering centralized placement. We then show that an order-optimal sum-GDoF is also attained in decentralized setting where no coordination during the placement phase is allowed [72]. These results settle down the problem of (order) optimal robust interference management for cache-aided wireless networks with partial instantaneous CSIT only. The result in this chapter has been published in:

- E. Piovano, H. Joudeh, B. Clerckx, "Generalized Degrees of Freedom of the Symmetric Cache-Aided MISO Broadcast Channel with Partial CSIT", *IEEE Transactions on Information Theory*, 2019.

Robust Interference Management for Cache-Aided Wireless Interference Networks

In Chapter 5 we extend our analysis to a more general setup, by considering a cache-aided wireless interference network where an arbitrary number of single-antenna transmitters serve an arbitrary number of single-antenna receivers. Each transmitter and each receiver is equipped with a cache memory where it can prestore a fraction of the library (hence transmitters cannot access the entire library but only the content in their memories).

In the considered setup, we consider that the communication during the delivery phase takes place over two heterogeneous parallel subchannels: one for which transmitters have access to the instantaneous channel coefficients (i.e. perfect CSIT), and another for which the transmitters have no knowledge of the instantaneous channel coefficients (i.e. no CSIT). This setup models scenarios in which channel state feedback is available only for a fraction of sub-carriers in an OFDMA system. In this context, the partial CSIT can be then interpreted as the fraction of the bandwidth corresponding to the perfect CSIT. We focus here on uncoded placement and on separable one-shot linear delivery schemes where the spreading of channel symbols over time or frequency (i.e. subchannels) is not allowed. Such linear schemes are appealing due to their practicality and their suitability for making theoretical progress on otherwise difficult or intractable information-theoretic problems, like in the considered case.

For the considered setup, we first characterize the optimal one-shot linear sum-DoF up to constant multiplicative factor of 2 for all system parameters, by assuming centralized placement. Next, we extend to the case of decentralized case at the receivers. In particular, we characterize an achievable one-shot linear sum-DoF under decentralized placement which is up to constant multiplicative factor of 3 for all system parameters [38].

The result in this chapter has been accepted for publication in:

- E. Piovano, H. Joudeh, B. Clerckx, "Centralized and Decentralized Cache-Aided Interference Management in Heterogeneous Parallel Channels", *Accepted for publication in IEEE Transactions on Communications*, 2019.

1.3. Thesis Organization

As already aforementioned, this thesis is divided into two parts. In Chapter 2 and 3 we study the fundamental limits of robust interference management for content-oblivious wireless networks. In particular, in Chapter 2 we study the optimal DoF region of the K-user MISO BC with arbitrary CSIT levels. In Chapter 3 we study the overloaded MISO BC, where the number of users is larger than the number of antennas at the transmitter.

On the other hand, in Chapter 4 and 5 we study the fundamental limits of robust interference management for content-aware wireless networks. In Chapter 4 we study the symmetric K-user cache-aided MISO BC under the GDoF metric. In Chapter 5 we study the cache-aided interference channel, where an arbitrary number of cache-aided transmitters serve an arbitrary number of cache-aided users. Finally, 6 concludes the thesis.

1.4. Publications

The work in my PhD has resulted in a number of papers, that have been published, accepted or submitted for publications. Some of the papers have not been included in the content of this thesis.

1.4.1. Fundamental Limits of Content-Oblivious Networks

- E.Piovano, H. Joudeh, B. Clerckx, "Overloaded MU-MISO transmission with imperfect CSIT", *Asilomar Conference on Signals, Systems, and Computers*, Asilomar, 2016.
- E.Piovano, B. Clerckx "Optimal DoF region of the K-User MISO BC with Partial CSIT", *IEEE Communication Letters*, 2017.

1.4.2. Fundamental Limits of Content-Aware Networks

- E. Piovano, H. Joudeh, B. Clerckx, "On coded caching in the overloaded MISO broadcast channel", *International Symposium on Information Theory*, ISIT, 2017.
- E. Piovano, H. Joudeh, B. Clerckx, "Robust cache-aided interference management under full transmitter cooperation", *International Symposium on Information Theory*, ISIT, 2018.
- E. Piovano, H. Joudeh, B. Clerckx, "Generalized Degrees of Freedom of the Symmetric Cache-Aided MISO Broadcast Channel with Partial CSIT", *IEEE Transactions on Information Theory*, 2019.

- E. Piovano, H. Joudeh, B. Clerckx, “Centralized and Decentralized Cache-Aided Interference Management in Heterogeneous Parallel Channels”, *Accepted for publication in IEEE Transactions on Communications*, 2019.

2. DoF Region of the K-user MISO BC with Partial CSIT

2.1. Overview of the Chapter

In this chapter we make progress towards the understanding of the fundamental limits of robust interference management under full transmitter cooperation in the context of the K-user MISO BC, where a K-antenna transmitter serves K single-antenna users. The transmitter has partial instantaneous knowledge of the channels of the users. The main result of this chapter is the characterization of the optimal DoF region under arbitrary CSIT levels for the users. Note that, before this work, only the sum-DoF was known in the literature, while no attempt was made to derive the entire DoF region.

The derivation of the DoF region requires a two-steps approach. We first derive a polyhedral outer-bound region. Then, we characterize all the facets of such region and we show that a rate-splitting strategy with flexible assignment of the DoF of the common codeword and flexible power allocation for the private codewords achieves each of such facet. Note that considering the facets of the polyhedral region is an original and unconventional approach. In fact, most of the achievability proofs existing in the literature rely on characterizing and showing the achievability of the corner points of the region. While this latter approach is feasible for small values of K, the number of corner points dramatically explodes for increasing K, making this impracticable to generalize in the setting of this chapter. Surprisingly, differently from the corner points, the facets of the considered polyhedral region can be more easily characterized. In particular, each of such facets can be rewritten in form of K inequalities which bound the per-user DoF of each individual user, which turns out to be key to show the achievability through rate-splitting.

2.2. Introduction

In this chapter we consider the K-user Multiple-Input-Single-Output (MISO) Broadcast Channel (BC), which consists of a K-antenna transmitter which serves K single-antenna users. Such setup can model a radio cell where a multiple antenna Base Station (BS) is connected to multiple users. However, the transmit antennas are not necessarily physically co-located, and may generally represent K radio heads (or remote antennas) connected through a strong fronthaul.

We assume that the transmitter has a partial instantaneous knowledge of the channels of the users, modelled by the CSIT quality parameter β introduced in Section 1.1. In particular, each user i has a CSIT quality $\beta_i \in [0, 1]$. As seen in Section 1.1, the case $\beta_i = 1$ is equivalent to perfect CSIT from a DoF perspective. In particular, if all users have a CSIT quality equal to 1, a full spatial multiplexing gain of K can be enabled.

The other extreme case $\beta_i = 0$, also denoted as finite precision CSIT, proved instead to be significantly more difficult to characterize. In particular, a nearly one decade old open problem proposed in [49] conjectured the total collapse of the sum-DoF to 1 under finite precision CSIT for all users, i.e. $\beta_i = 0$ for all i. This conjecture was shown to be true by the seminal work of Davoodi and Jafar in [22]. This in turn implies that final precision CSIT is as (un)useful as no CSIT from a DoF perspective. On the other hand, a CSIT quality $\beta_i \in (0, 1)$ helps to save some of the spatial multiplexing gains and to achieve a sum-DoF between 1 and K.

The role of partial instantaneous channel knowledge is best exemplified by the main result in [22] which, by assuming without loss of generality that the CSIT qualities of the users are sorted as $\beta_1 \geq \beta_i$ for all i, proved that an upper-bound of the optimal sum-DoF d_Σ for the K-user MISO BC is given by

$$d_\Sigma \leq 1 + \beta_2 + \cdots + \beta_K.$$

Note that all the observations made earlier about the impact of the CSIT quality, in particular the DoF collapse under finite precision CSIT, are reflected in the above upper-bound. Interestingly, note that the sum-DoF collapse is still verified in case one user has perfect CSIT and all the other users finite precision CSIT. This upper-bound was proved by Davoodi and Jafar in [22] by introducing a novel unconventional combinatorial argument known as aligned image set approach (AIS). We will give some more insights about the AIS in Section 2.4. Moreover, the AIS will be revisited in Chapter 4, where the proof in [22] will be extended to cache-aided settings.

Remarkably, this upper-bound is achievable through a rate-splitting strategy with proper allocation for the private codewords, as shown in the two users case in [13] and then extended to the K users case in [45, 46]. It follows that this upper-bound is tight and corresponds to the optimal sum-DoF of the K-user MISO BC under arbitrary CSIT levels. The rate-splitting scheme as well the achievability proof will be revisited in Section 2.5.

2.2.1. Main Contributions

While the sum-DoF is an important information to know, it does not reveal any information about the per-user DoF achieved by each user. The per-user DoF are instead characterized by the DoF region, which is the set of all achievable DoF tuples $(d_1, \ldots, d_K)$.

The main result of this chapter is the characterization of the optimal DoF region of the K-user MISO BC with arbitrary CSIT levels for the users. The proof involves two steps: a converse argument and an achievability argument. Starting from the converse, on the basis of the upper-bound in [22], a polyhedral outer-bound can be constructed by bounding the sum-DoF of each subset of users while ignoring the remaining users. The main challenge becomes then the achievability argument of the constructed outer-bound, as it requires to show that each point of the region is achievable. For small values of K, a conventional way to prove the achievability of a region is to characterize and show the achievability of the corner points (also denoted as vertices) [76]. Any other point can be then obtain by time-sharing over the corner points. However, for increasing values of K, the characterization of all corner points becomes in general unfeasible.

To overcome this problem, we introduce a novel and original approach: instead of characterizing and showing the achievability of the corner points, we characterize and show the achievability of each facet of the region. The key is to notice that, differently from the corner points, each of the facets of the polyhedral region can be rewritten in form of K inequalities which bound the per-user DoF of each individual user. Such characterization turns out to be suitable to show the achievability of each point of each facet by employing a rate-splitting strategy with flexible power allocation for the private codewords and flexible assignment of the DoF of the common codeword.

To conclude, the characterization of the DoF region of the K-user MISO BC is an essential step towards a deeper understanding of the fundamental limits of robust interference management under full transmitter cooperation, as it is an important achievement along a path of refinements towards deriving capacity limits and capacity regions.

2.2.2. Notation

In order to state the main result of the chapter, we will define $\mathcal{A}$ as the set of all possible non-empty subsets of $\mathcal{K}$ with elements arranged in an ascending order. For instance, in case of $\mathcal{K} = \{1, 2, 3\}$, the set $\mathcal{A}$ is given by $\mathcal{A} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. Throughput the chapter, we denote a generic element of $\mathcal{A}$, which is itself a set, with a calligraphic upper case letter and we denote its elements with the corresponding lower case letters (with numbered subscripts). For instance the subset $\mathcal{S} = \{s_1, s_2, \dots, s_{|\mathcal{S}|}\} \in \mathcal{A}$, where $s_1 < s_2 < \dots < s_{|\mathcal{S}|}$, or the subset $\mathcal{G} = \{g_1, g_2, \dots, g_{|\mathcal{G}|}\} \in \mathcal{A}$, where $g_1 < g_2 < \dots < g_{|\mathcal{G}|}$. Note that in this chapter, without loss of generality, we assume that the users are ordered with respect to their CSIT qualities, i.e. $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_K$.

As for the remainder of the chapter, the organization is as follows. Section 2.3 introduces the system model. Section 2.4 introduce the sum-DoF upper-bound extablished in [22]. Section 2.5 introduces the rate-splitting scheme. Section 2.6 presents the main result and related insights. In Section 2.7, we derive an outer-bound of the optimal DoF region. In Section 2.8 we prove the achievability of the DoF region. Finally, Section 2.9 summarizes and concludes the chapter.

2.3. System Model

Before diving into the details of work of this chapter, we briefly revisit the problem setting in Section 1.1.2 to see how it specializes for the K-user MISO BC considered in this chapter. We consider here a MISO BC comprising of a K-antenna transmitter which serves K single-antenna receivers (or users). The users are indexed by the set $\mathcal{K} = \{1, 2, \dots, K\}$. The input-output relationship at the t-th use of the physical channel, $t \in [T]$ where T is the duration of the communication, for each receiver $i \in \mathcal{K}$, is modelled by

$$Y_i(t) = \sum_{j=1}^{K} G_{ij}(t) X_j(t) + Z_i(t) \tag{2.1}$$

where $Y_i(t) \in \mathbb{C}$ is the received signal, $X_j(t) \in \mathbb{C}$ is the signal transmitted from antenna j, $G_{ij}(t)$ is the fading channel coefficient between transmit j and receiver i, and $Z_i(t) \sim \mathcal{N}_{\mathbb{C}}(0,1)$ is the normalized additive white Gaussian noise (AWGN), which is i.i.d. across all dimensions. The transmitted signal across all transmitting antennas at the t-th channel use is $\mathbf{X}(t) \triangleq [X_1(t) \cdots X_K(t)]^{\mathrm{T}}$. The transmitter is then subject to the power constraint $\frac{1}{T}\sum_{t=1}^{T} |\mathbf{X}(t)|^2 \leq P$. On the other hand, for each user i, we collect all its corresponding channel coefficients from all the transmitting antennas into the vector given by $\mathbf{G}_i(t) \triangleq [G_{i1}(t) \cdots G_{iK}(t)]^{\mathrm{T}}$. The equation in (2.1) can then be rewritten as

$$Y_i(t) = \mathbf{G}_i^{\mathrm{T}}(t)\mathbf{X}(t) + Z_i(t). \tag{2.2}$$

Such compact form will be useful to simplify the notation in the explanation of rate-splitting. To conclude, we point out that the same ergodic-rate assumption explained in Section 1.1.2 holds here.

2.3.1. Partial CSIT

We also briefly revisit here the partial CSIT assumption described in Section 1.1.2. Motivations and more details can be find in Section 1.1.2. Under partial CSIT, the channel coefficient between antenna j and user i is modeled as

$$G_{ij}(t) = \hat{G}_{ij}(t) + \sqrt{P^{-\beta_i}}\tilde{G}_{ij}(t) \tag{2.3}$$

where $\hat{G}_{ij}(t)$ is channel estimate, $\tilde{G}_{ij}(t)$ is the channel estimation error and $\beta_i \in [0,1]$ is the parameter capturing the CSIT quality level. As discussed in Section 1.1.2 and Section 2.2, the parameter $\beta_i \in [0,1]$ captures the whole range of channel knowledge, where $\beta_i = 0$ corresponds to the case of finite precision CSIT (equivalent to absent CSIT), while $\beta_i = 1$ corresponds to the case of perfect CSIT. The difference between $\hat{G}_{ij}(t)$ and $\tilde{G}_{ij}(t)$ is that the former is revealed to the transmitter while the latter is not. As done for the channel coefficients, we collect all the channel estimates and channel error terms from all antennas to user i into the vectors $\hat{\mathbf{G}}_i(t) = [\hat{G}_{i1}(t) \cdots \hat{G}_{iK}(t)]^{\mathrm{T}}$ and $\tilde{\mathbf{G}}_i(t) = [\tilde{G}_{i1}(t) \cdots \tilde{G}_{iK}(t)]^{\mathrm{T}}$, respectively. It follows that $\mathbf{G}_i(t)$ can be written as

$$\mathbf{G}_i(t) = \hat{\mathbf{G}}_i(t) + \sqrt{P^{-\beta_i}}\tilde{\mathbf{G}}_i(t). \tag{2.4}$$

We assume, without loss of generality, that the users are ordered with respect to their CSIT qualities, i.e. $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_K$. The tuple of CSIT levels is collected into the vector $\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K)$. Before we proceed, it is worth highlighting that channel state information at the receivers (CSIR) is assumed to be perfect.

In the next sections we will revisit the two main ingredients for the derivation of the optimal DoF region of the K-user MISO BC under arbitrary CSIT levels. The first ingredient is the sum-DoF upper-bound obtained in [22]. This upper-bound will be utilized to derive an outer-bound of the entire DoF region in Section 2.7. The second ingredient is the rate-splitting scheme, which will be utilized to prove the achievability of such outer-bound.

2.4. Sum-DoF Upper-Bound

In this section, we dive more into the details and insights of the DoF upper-bound derived in [22]. To start, we state the main theorem in [22].

Theorem 2.1. *[22, Th. 1] For the K-user MISO BC the sum-DoF is bounded above by*

$$\sum_{i \in \mathcal{K}} d_i \leq 1 + \sum_{i \in \mathcal{K} \setminus \{1\}} \beta_i. \tag{2.5}$$

The result was shown assuming $\beta_1 = 1$ for the first user. However, since enhancing the CSIT does not harm the sum-DoF, the same upper-bound holds for a generic value of $\beta_1 \in [\beta_2, 1]$. This upper-bound was obtained by Davoodi and Jafar by introducing a new technique denoted as aligned image set (AIS). The key idea behind the AIS argument is to bound the expected number of codewords which can be decoded at their desidered receivers whose images align at the uninteded receivers under finite precision CSIT. The AIS will be revisited and extended to the cache-aided setting in Chapter 4. Importantly, the DoF upper-bound of Theorem 2.1 can be achieved by employing a rate-splitting strategy, where interference is managed by superimposing a common codeword decoded by all users on top of private codewords for the intended users only. It follows that this upper-bound is tight and it corresponds to the optimal sum-DoF of the K-user MISO BC with arbitrary CSIT levels. We want to recall three main insights from this result (the sum DoF is again denoted as d_{Σ}):

1. In case of $\beta_i = 0$ for all $i \in \mathcal{K}$ we obtain $d_{\Sigma} \leq 1$, hence all the benefits of multiple transmitting antennas are lost. This is also denoted as the collapse of the DoF under finite precision and it was a decade long conjecture proposed in [49] and finally solved in [22]. Interestingly, the DoF collapse still happens when perfect CSIT is available for one user and finite precision CSIT for all the others.

2. In case of $\beta_i = 1$ for all $i \in \mathcal{K}$ we obtain $d_{\Sigma} \leq K$. This DoF upper-bound is then achieved by employing a simple zero-forcing strategy, which in turn enables full spatial multiplexing gain when the channel estimation errors of the users decay as $O(P^{-1})$.

3. In case of $\beta_i \in (0, 1)$, some of the spatial multiplexing gains can be saved. In this case, the sum-DoF upper-bound can be achieved by rate-splitting, by properly designing the transmission power of the common and private codewords. This will be revisited in the next section.

2.5. Rate-Splitting

In this section we revisit the rate-splitting scheme, the key strategy for the achievability argument. In rate-splitting two kind of codewords are superimposed in the power domain: a common codeword decoded by all users, on top of private codewords decoded by the respective users only. In particular, the transmitter splits the message W_i intended for each user $i \in \mathcal{K}$ into a common (or public) sub-message $W_i^{(\mathrm{c})}$ and a private sub-message $W_i^{(\mathrm{p})}$, i.e. $W_i = (W_i^{(\mathrm{c})}, W_i^{(\mathrm{p})})$. All the common sub-messages $W_1^{(\mathrm{c})}, \dots, W_K^{(\mathrm{c})}$ are jointly encoded into the common codeword $X^{(\mathrm{c})}$, i.e. $\left(W_1^{(\mathrm{c})}, \dots, W_K^{(\mathrm{c})}\right) \to X^{(\mathrm{c})}$, which has to be decoded by all K users. On the other hand, each private sub-message $W_i^{(\mathrm{p})}$ is encoded into the private codeword $X_i^{(\mathrm{p})}$, i.e. $W_i^{(\mathrm{p})} \to X_i^{(\mathrm{p})}$, which is decoded by user i only. All the codewords are assumed to be drawn from a unitary-power Gaussian codebook. After the encoding, the codewords are linearly precoded and power allocated. The transmitted signal at the t-th channel use, i.e. $\mathbf{X}(t) = [X_1(t) \cdots X_K(t)]^T$, takes then the form

$$\mathbf{X}(t) = \sqrt{P^{(\mathrm{c})}}\mathbf{V}^{(\mathrm{c})}(t)X^{(\mathrm{c})}(t) + \sum_{i \in \mathcal{K}} \sqrt{P_i^{(\mathrm{p})}}\mathbf{V}_i^{(\mathrm{p})}(t)X_i^{(\mathrm{p})}(t) \tag{2.6}$$

where $\mathbf{V}^{(\mathrm{c})}(t) \in \mathbb{C}^{K\times 1}$ and $\mathbf{V}_i^{(\mathrm{p})}(t) \in \mathbb{C}^{K\times 1}$ are unitary precoding vectors, and $P^{(\mathrm{c})}$ and $P_i^{(\mathrm{p})}$ are the corresponding long-term allocated powers to the precoded codewords, which have to satisfy the constraint $P^{(\mathrm{c})} + \sum_{i\in\mathcal{K}} P_i^{(\mathrm{p})} \leq P$. Note that $X^{(\mathrm{c})}(t)$ corresponds to the t-th symbol of the common codeword, while $X_i^{(\mathrm{p})}(t)$ corresponds to the t-th symbol of the private codeword intended for user i.

As the common codeword $X^{(\mathrm{c})}$ has to be decoded by all users, $\mathbf{V}^{(\mathrm{c})}(t)$ are chosen as a generic (random) unit vector. On the other hand, as the private codewords are intended for the respective users only, the precoding vectors $\mathbf{V}_i^{(\mathrm{p})}(t)$ are chosen in order to zero-force interference at the unintended users. It follows that $\mathbf{V}_i^{(\mathrm{p})}(t) \triangleq \left[V_{i1}^{(\mathrm{p})}(t) \ \cdots V_{iK}^{(\mathrm{p})}(t)\right]^{\mathsf{T}}$ is a zero-forcing unit vector designed by the transmitter using the channel estimates $\hat{\mathbf{G}}_j(t)$ such that

$$\hat{\mathbf{G}}_j(t)^{\mathrm{T}}\mathbf{V}_i(t)^{(\mathrm{p})} = 0, \quad j \neq i. \tag{2.7}$$

2.5.1. Power Allocation and Received Signal

The achievable DoF of rate-splitting depends in general on the power allocations for the common and private codewords. However, as we will next, the power allocation for the common codeword $P^{(\mathrm{c})}$ is always let to scale with the SNR P as $O(P)$, which let the power allocations for the private codewords to be the only design variables. We first consider arbitrary power allocations for the private codewords, and we will then specialize to the specific setting considered in this chapter. As already mentioned, rate-splitting considers a power allocation which scales with the SNR P as $O(P)$ for the common codeword, and as $O(P^{a_i})$ for the i-th private codeword, where a_i corresponds to the power level and it is such that $a_i \in [0, 1]$. This can be formalized by $P^{(\mathrm{c})} = O(P)$ for the common codeword and $P_i^{(\mathrm{p})} = O(P^{a_i})$ for the i-th private codeword. Note that the different scaling of the power allocations is used to achieve different DoF values. The use of the big O notation, or Landau's symbol, is used to guarantee that the power constraint $P^{(\mathrm{c})} + \sum_{i\in\mathcal{K}} P_i^{(\mathrm{p})} \leq P$ is not violated. In fact, while we allocate the power for the common and private codewords to grow with an order of P and P^{a_i}, respectively, the power allocations can be adjusted with appropriated scaling or additive factors to guarantee that the power constraint is maintained, without changing the achievable DoF. For instance, we can allocate a power $P - \sum_{i=1}^{K} P^{a_i}$ to the common codeword and P^{a_i} to the i-th private codeword. Another way to maintain the power constraint would be to divide all the allocated powers by a constant scaling factor.

The transmitted signal can be then written as:

$$\mathbf{X}(t) = \underbrace{\sqrt{P^{(\mathrm{c})}}\mathbf{V}^{(\mathrm{c})}(t)X^{(\mathrm{c})}(t)}_{O(P)} + \sum_{i\in\mathcal{K}} \underbrace{\sqrt{P_i^{(\mathrm{p})}}\mathbf{V}_i^{(\mathrm{p})}(t)X_i^{(\mathrm{p})}(t)}_{O(P^{a_i})}. \tag{2.8}$$

where the power constraint $P^{(\mathrm{c})} + \sum_{i\in\mathcal{K}} P_i^{(\mathrm{p})} \leq P$ must be mantained.

The received signal by user $i \in \mathcal{K}$ is given by

$$\begin{aligned} Y_i(t) = \underbrace{\sqrt{P^{(\mathrm{c})}}\mathbf{G}_i^{\mathrm{T}}(t)\mathbf{V}^{(\mathrm{c})}(t)X^{(\mathrm{c})}(t)}_{O(P)} + \underbrace{\sqrt{P_i^{(\mathrm{p})}}\mathbf{G}_i^{\mathrm{T}}(t)\mathbf{V}_i^{(\mathrm{p})}(t)X_i^{(\mathrm{p})}(t)}_{O(P^{a_i})} \\ + \sum_{j\in\mathcal{K}\setminus\{i\}} \underbrace{\sqrt{P_j^{(\mathrm{p})}}\mathbf{G}_i^{\mathrm{T}}(t)\mathbf{V}_j^{(\mathrm{p})}(t)X_j^{(\mathrm{p})}(t)}_{O(P^{a_j-\beta_i})} + Z_i(t) \end{aligned} \tag{2.9}$$

where, in the equation above, we have pointed out that the common codeword $X^{(\mathrm{c})}$ is received with strength $O(P)$, the intended private codeword $X_i^{(\mathrm{p})}$ is received with strength $O(P^{a_i})$, while the codeword $X_j^{(\mathrm{p})}$ intended for users $j \neq i$ generates an interference with strength $O(P^{a_j-\beta_i})$. The reason why the codeword intended for user j generates an interference of $O(P^{a_j-\beta_i})$ can be derived from the utilized definition of partial CSIT in Section 2.3.1. In fact, with the zero-forcing unit vector $\mathbf{V}_i^{(\mathrm{p})}$ designed to satisfy Eq. (2.7), it can be shown that

$$\lim_{P\to\infty} -\frac{\log \mathbb{E}[|\tilde{\mathbf{G}}_i(t)^{\mathrm{T}}\mathbf{V}_j(t)^{(\mathrm{p})}|^2]}{\log(P)} = \beta_i, \tag{2.10}$$

where the expectation is taken over both the channel estimate and the actual channel vector. More details can be found [14, 15] and references therein. From (2.10), and given that the private codeword $X_j^{(\mathrm{p})}(t)$ is delivered with power $O(P^{a_j})$, it follows that the generated interference is $O(P^{a_j-\beta_i})$.

2.5.2. Rate-Splitting Decoding Scheme and Achievable DoF

In rate-splitting, all users decode the common codeword $X^{(\mathrm{c})}$ by treating interference from the private codewords as noise, and they retrieve the common message $W^{(\mathrm{c})}$. From $W^{(\mathrm{c})}$ each user i can then retrieve its own common sub-message $W_i^{(\mathrm{c})}$. Each user i removes then $X^{(\mathrm{c})}$ by performing successive interference cancellation (SIC) and proceeds to decode its own private codeword $X_i^{(\mathrm{p})}$, by treating the interference from all the other private codewords as noise. From $X_i^{(\mathrm{p})}$ each user i can then retrieve its own private message $W_i^{(\mathrm{p})}$.

The achievable DoF of the common and private codewords can be then characterized using the analysis developed in the previous section. In particular, we remind that each user receives the common codeword $X^{(\mathrm{c})}$ with power $O(P)$, its own private codeword with power $O(P^{a_i})$ and the private codeword for user $j \neq i$ with power $O(P^{a_j-\beta_i})$. A simple calculation verifies that, in order to be successfully decode by all users, the common codeword $X^{(\mathrm{c})}$ can support a DoF of

$$d^{(\mathrm{c})} = 1 - \max_{i \in \mathcal{K}} a_i. \tag{2.11}$$

The DoF of the common codeword can be split in all possible ways among users in $\mathcal{K}$. We denote as $d_i^{(\mathrm{c})}$ the DoF of the common codeword assigned to user i. It follows that any non-negative real tuple $\mathbf{d}^{(\mathrm{c})} = (d_1^{(\mathrm{c})}, \dots, d_K^{(\mathrm{c})})$, which satisfies $\sum_{i \in \mathcal{K}} d_i^{(\mathrm{c})} = d^{(\mathrm{c})}$, is an admissible partition of the DoF carried by the common codeword among the users in $\mathcal{K}$.

Next, each user i removes $X^{(\mathrm{c})}$ by performing SIC and proceeds to decode its own private codeword $X_i^{(\mathrm{p})}$. A simple calculation can verify that the private codeword $X_i^{(\mathrm{p})}$ intended for user $i \in \mathcal{K}$ can support a DoF of

$$d_i^{(\mathrm{p})} = \left(a_i - \left(\max_{j \in \mathcal{K} \setminus \{i\}} a_j - \beta_i \right)^+ \right)^+ \tag{2.12}$$

where $(x)^+ = \max\{x, 0\}$. The DoF of all the private codewords are collected into the vector the tuple $\mathbf{d}^{(\mathrm{p})} = (d_1^{(\mathrm{p})}, \dots, d_K^{(\mathrm{p})})$.

To sum up, a per-user DoF tuple $\mathbf{d} = (d_1, \dots, d_K)$ is achievable by rate-splitting with power levels given by $\mathbf{a} = (a_1, \dots, a_K)$ if the following equality holds:

$$(d_1, \dots, d_K) = (d_1^{(\mathrm{p})}, \dots, d_K^{(\mathrm{p})}) + (d_1^{(\mathrm{c})}, \dots, d_K^{(\mathrm{c})}) \tag{2.13}$$

where $d_i^{(\mathrm{p})}$ for any $i \in \mathcal{K}$ is given by (2.12), while $(d_1^{(\mathrm{c})}, \dots, d_K^{(\mathrm{c})})$ indicates an admissible partition of the total DoF carried by the common codeword, where $d^{(\mathrm{c})}$ is given by Eq. (2.11).

2.5.3. Comparison between Rate-Splitting and Zero-Forcing

The main difference between rate-splitting and zero-forcing is the transmission of the common codeword in rate-splitting which is absent in zero-forcing. As in zero-forcing the messages are not split and only private codewords are delivered over the channel, the transmitted signal takes the form

$$\mathbf{X}(t) = \sum_{i \in \mathcal{K}} \sqrt{P_i} \mathbf{V}_i(t) X_i(t) \tag{2.14}$$

where X_i is the codeword intended for receiver i and the precoding vectors $\mathbf{V}_i(t)$ correspond to the zero-forcing precoding vectors $\mathbf{V}_i^{(\mathrm{p})}(t)$ in rate-splitting. It follows that, for a power allocation vector $\mathbf{a} = (a_1, \ldots, a_K)$, zero-forcing achieves the DoF tuple $\mathbf{d} = (d_1, \ldots, d_K)$, where for $j \in \mathcal{K}$

$$d_i = \left(a_i - \left(\max_{j \in \mathcal{K} \setminus \{i\}} a_j - \beta_j \right)^+ \right)^+ . \tag{2.15}$$

This clearly corresponds to the achievable DoF tuple for the private codewords in rate-splitting. It is readily seen that rate-splitting better tackles interference thanks to the transmission of the common codeword which in turn enhances the DoF, as characterized in Eq. (2.11).

2.5.4. Achievability of the Upper-Bound in Theorem 2.1

We prove here that rate-splitting attains the sum-DoF upper-bound in (2.1). To do so, we need to consider a power allocation vector $\mathbf{a}$ with equal power levels for all the users, where this power level can be any value between β_2 and β_1, i.e.

$$\mathbf{a} = (b, b, \ldots, b), \quad \text{with } \beta_2 \leq b \leq \beta_1. \tag{2.16}$$

Applying this to Eq. (2.11), the common codeword can support a DoF of

$$d^{(\mathrm{c})} = 1 - b. \tag{2.17}$$

As we are considering the sum-DoF, the split of the common codeword DoF among users is irrelevant. To calculate the DoF supported by the private codewords, we apply Eq. (2.12) and we obtain

$$d_i^{(\mathrm{p})} = \begin{cases} b & i = 1 \\ \beta_i & i \in \mathcal{K} \setminus \{1\}. \end{cases} \tag{2.18}$$

Hence, the sum-DoF is equal to

$$1 + \beta_2 + \cdots + \beta_K, \tag{2.19}$$

which corresponds to the one in Theorem 2.1. Note that it can be easily verified that zero-forcing achieves a sum-DoF of

$$\beta_1 + \cdots + \beta_K \tag{2.20}$$

hence, it only attains the upper-bound in Theorem 2.1 for $\beta_1 = 1$, while it fails when $\beta_1 < 1$, and in this latter case rate-splitting is needed.

2.6. Optimal DoF Region of MISO BC with Partial CSIT

We state here the main result of this chapter, which is the characterization of the optimal DoF region of the K-user MISO BC under arbitrary CSIT levels for the users. Note that we use the notation in Section 2.2.2.

Theorem 2.2. *The optimal DoF region $\mathcal{D}^*$ of the K-user MISO BC under arbitrary CSIT levels is given by all the real tuples $(d_1, \ldots, d_K)$ which satisfy*

$$d_i \geq 0, \quad \forall i \in \mathcal{K} \tag{2.21}$$

$$\sum_{i \in \mathcal{S}} d_i \leq 1 + \sum_{i \in \mathcal{S} \setminus \{s_1\}} \beta_i, \quad \forall \mathcal{S} \in \mathcal{A}, \tag{2.22}$$

where we consider here the notation in Section 2.2.2. To remind, $\mathcal{A}$ is the set of all possible non-empty subsets of $\mathcal{K}$ with elements arranged in an ascending order, and s_1 indicates the smallest element of the subset $\mathcal{S} \in \mathcal{A}$.

We first denote as $\mathcal{D}$ the above region described by the inequalities (2.21) and (2.22). We show that $\mathcal{D}$ coincides with the optimal DoF region $\mathcal{D}^*$ by showing that $\mathcal{D}$ is simultaneously an outer-bound of $\mathcal{D}^*$ and it is achievable. We will show in Section 2.7 that $\mathcal{D}$ is an outer-bound of $\mathcal{D}^*$ as a direct extension of Theorem 2.1, as already anticipated in the previous section. On the other hand, the proof that $\mathcal{D}$ is achievable is significantly more involved and it will be presented in Section 2.8.

2.6.1. Main Insights and Connections with Other Works

In this section we provide the main insights which can be derived from Theorem 2.2.

1. First we notice that the DoF region does not change while considering any value of $\beta_1 \in [\beta_2, 1]$. This consideration extends to the DoF region what was already observed for the sum-DoF in Theorem 2.1. Hence, from Theorem 2.2 we obtain that, regardless of the policy or scheduling utilized to serve the users in the network, the performance is not deteriorated when the CSIT quality of the user with the highest CSIT quality is alleviated to β_2.

2. It is readily seen that, in case of $\beta_i = 1$ for all users $i \in \mathcal{K}$, the DoF region boils down to $d_i \in [0, 1]$ for any $i \in \mathcal{K}$. Any point of this region can be then achieved by zero-forcing with flexible power allocation. In particular, any point $(d_1, d_2, \ldots, d_k)$ can be achieved by zero-forcing with power allocation $\mathbf{a} = (d_1, d_2, \ldots, d_k)$.

3. In case of $\beta_1 = 1$ we have seen in Section 2.5.4 that zero-forcing achieves the optimal sum-DoF. However, it fails to achieve the entire DoF region, except for the case where all the users excluding the last one have perfect CSIT quality, i.e. $\beta_i = 1$ for all users $i \in \mathcal{K} \setminus \{K\}$. To better illustrate this, let us consider the case $K = 3$ where $\beta_1 = 1$ and $\beta_2, \beta_3 < 1$. We consider the points of the region $\mathcal{D}^*$ which satisfy $d_2 + d_3 = 1 + \beta_3$. These points cannot be achieved by zero-forcing as zero-forcing only attains $d_2 + d_3 \leq \beta_2 + \beta_3$, as seen in Section 2.5.4.

4. As a corollary of Theorem 2.2 we can deduce that the maximum symmetric per-user DoF, denoted as d^*_{sym}, is given by

$$d^*_{\text{sym}} = \min_{j=\{2,\ldots,K\}} \frac{1 + \sum_{i=K-j+1}^{K} \beta_i}{J} \tag{2.23}$$

which is in agreement with the previous result obtained in [46].

2.7. Construction of the Outer-Bound

In this section we show that the region $\mathcal{D}$ described by Eq. (2.21) and (2.22) is an outer-bound of the optimal DoF region $\mathcal{D}^*$. First, let us consider a subset of users $\mathcal{S} \in \mathcal{A}$. We start by proving that

$$\sum_{i \in \mathcal{S}} d_i \leq 1 + \sum_{i \in \mathcal{S} \backslash \{s_1\}} \beta_i, \quad \forall \mathcal{S} \in \mathcal{A}. \tag{2.24}$$

In fact, the sum-DoF of the subset $\mathcal{S} \in \mathcal{A}$ is bounded above by the sum-DoF obtained while adding $K - |\mathcal{S}|$ users with CSIT quality equal to 0, which we denote as $\mathcal{K}_0$ (this is obvious as adding new users can never hurt the sum-DoF). This latter setting corresponds to the K-user MISO BC and we can consequently apply Theorem 2.1. It follows that

$$\sum_{i \in \mathcal{S}} d_i \leq \sum_{i \in \mathcal{S} \bigcup \mathcal{K}_0} d_i \leq 1 + \sum_{i \in \mathcal{S} \backslash \{s_1\}} \beta_i + \sum_{i \in \mathcal{K}_0} 0 = 1 + \sum_{i \in \mathcal{S} \backslash \{s_1\}} \beta_i. \tag{2.25}$$

Hence, the inequality in (2.24) holds. While considering all possible subsets of users $\mathcal{S} \in \mathcal{A}$, we obtain (2.22). Moreover, the DoF of each user is a non negative value, from which we obtain (2.21). It follows that the region $\mathcal{D}$ described by (2.21) and (2.22) is an outer-bound of the optimal DoF region $\mathcal{D}^*$.

2.8. Proof of the Achievability of $\mathcal{D}$

In this section we prove the achievability of the region $\mathcal{D}$ characterized in Theorem 2.2. The region $\mathcal{D}$ is the K-dimensional polyhedral region given by the intersection of the half-spaces described by (2.21) and (2.22). Each inequality in (2.21) and (2.22) denotes an half-space delimited by the hyperplane obtained while substituting the half-space' inequality with an equality. Any of these hyperplanes contains a facet of the polyhedral region $\mathcal{D}$ and the set of all the facets corresponds to the boundary of $\mathcal{D}$. Showing the achievability of such region turns out to be particularly challenging as the most common approach, which consists of showing the achievability of the corner points of the polyhedral region, cannot be easily generalized here as we will describe in the next section.

2.8.1. Problem in Finding the Corner Points

The most common way to prove the achievability of a DoF region, it is to characterize and show the achievability of its corner points. For instance, while considering $K = 2$, the region boils down to a polygon and the corner points were characterized and shown to be achievable by rate-splitting in [13]. Every other point of the region can be then obtained by time-sharing over the corner points. However, in our case, characterize the corner points is only feasible for small K.

From a combinatorial perspective, it can be obtained that there are $2^K + K - 1$ inequalities in

(2.21) and (2.22), hence $2^K + K - 1$ hyperplanes delimiting the polyhedral region. Any corner point of the K-dimensional polyhedral region is given by the intersection of K hyperplanes. However, K hyperplanes may not intersect or the intersection may not be a point included in the region $\mathcal{D}$. For instance, while considering the case $K = 2$, the lines $d_1 = 0$ and $d_2 = 0$ never intersect, while the lines $d_1 = 1$ and $d_1 + d_2 = 1 + \beta_2$ intersect in the point $(0, 1 + \beta_2)$ which is not contained in the region $\mathcal{D}$. Hence, in order to characterize all the corner points, all the $\binom{2^K+K-1}{K}$ subsets of K hyperplanes have to be analyzed to see if they intersect in a point. In case a subset of K hyperplanes intersect in a point, it is needed to further verify if such a point belongs to the region $\mathcal{D}$, hence satisfies all the inequalities in (2.21) and (2.22). In case the point belongs to the outer-bound, it is a corner point. It is readily seen that such procedure is unfeasible for large K.

Moreover, the DoF region in Theorem 2.2 is not a polymatroid as shown in [125]. Hence, the polymatroid properties to simplify the characterization of the corner points [126, 127] cannot be utilized here. To overcome these problems a new approach is introduced, where the facets, instead of the corner points, are shown to be achievable.

2.8.2. A new Approach to show the Achievability of the Region $\mathcal{D}$

In this section we introduce the main technical novelty of the chapter, which is the new technique developed to show the achievability of the DoF region $\mathcal{D}$ described in Eq. (2.21) and (2.22). The main idea of the proof is to characterize and show the achievability of each facet of the polyhedral region $\mathcal{D}$. The facet contained in a specific hyperplane is characterized by the set of points which satisfy the equation of the hyperplane and all the other inequalities of the DoF region. Interestingly, given the structure of the inequalities in (2.21) and (2.22), the facets of the polyhedral region are much easier to characterize than the corner points. Please note that the corner points are contained into the facets, hence while showing the achievability of the facets, the achievability of the corner points is automatically guaranteed.

We proceed to show the achievability of $\mathcal{D}$ by induction over the number of users K. The hypothesis is clearly true for $K = 1$. In fact, in this case, the region (2.21) and (2.22) becomes $d_1 \geq 0$ and $d_1 \leq 1$, which is the range of DoF values achieved in a single-user scenario. We assume that the hypothesis is valid for $K = 1, \ldots, k-1$ and we consider the case $K = k$. The key idea of the proof is to show that the facets from (2.22) are achievable by rate-splitting with flexible power allocation for the private codewords and flexible split of the DoF carried by the common codeword, while the facets from (2.21) are achievable by induction hypothesis.

2.8.3. Proof of the Achievability of the Facets delimited by the Half-Spaces in Eq. (2.22)

In this section, we show the achievability of the facets contained in the hyperplanes which delimit the half-spaces in (2.22). Any of these hyperplane is given by $\sum_{i\in\mathcal{S}} d_i = 1 + \sum_{i\in\mathcal{S}\setminus\{s_1\}} \beta_i$, for a subset $\mathcal{S} \in \mathcal{A}$. We denote the facet contained in such an hyperplane as $\mathcal{F}_{\mathcal{S}}$. The facet $\mathcal{F}_{\mathcal{S}}$ can be analytically characterized as the set of all the points contained in the hyperplane which satisfy all the other inequalities of the polyhedral region in (2.21) and (2.22). Hence, $\mathcal{F}_{\mathcal{S}}$ is the set of all non-negative real tuples $(d_1, \ldots, d_k)$ such that

$$\sum_{i \in \mathcal{G}} d_i \leq 1 + \sum_{i \in \mathcal{G} \setminus \{g_1\}} \beta_i, \quad \forall \mathcal{G} \in \mathcal{A},\ \mathcal{G} \neq \mathcal{S} \tag{2.26}$$

$$\sum_{i \in \mathcal{S}} d_i = 1 + \sum_{i \in \mathcal{S} \setminus \{s_1\}} \beta_i \tag{2.27}$$

where the elements of any subset $\mathcal{G}$ of $\mathcal{A}$, which are arranged in an increasing order (i.e. $g_i < g_j$ for $i \neq q$) according to the notation in Section 2.2.2, are indicated as $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$. While the inequalities in (2.21) are satisfied by considering non-negative real tuples, Eq. (2.27) identifies the hyperplane containing $\mathcal{F}_{\mathcal{S}}$ and the inequalities in (2.26) identify all the other inequalities of $\mathcal{D}$ in (2.22).

Showing directly the achievability of $\mathcal{F}_{\mathcal{S}}$ by (2.26) and (2.27) seems a difficult task. To overcome this problem, we first rewrite $\mathcal{F}_{\mathcal{S}}$ in a equivalent form where we bound through inequalities the per-user DoF values achieved by each user $i \in \mathcal{K}$. This is obtained, for each $i \in \mathcal{K}$, by comparing an inequality in (2.26), for a specific $\mathcal{G}$ which is a function of the value i as it will be detailed later, with the equality in (2.27). The new parametrized form of the facet $\mathcal{F}_{\mathcal{S}}$ is then suitable to prove the achievability by rate-splitting. For the proof, we separately consider the subsets $\mathcal{S}$ such that $|\mathcal{S}| \geq 2$ and the subsets $\mathcal{S}$ such that $|\mathcal{S}| = 1$. We start with the case $|\mathcal{S}| \geq 2$.

Rewriting the Parametrized Form of the Facets $\mathcal{F}_{\mathcal{S}}$ for $|\mathcal{S}| \geq 2$

We consider here the case $|\mathcal{S}| \geq 2$. The first step is to rewrite the facet $\mathcal{F}_{\mathcal{S}}$ by bounding the per-user DoF values of each user $i \in \mathcal{K}$. We first bound the per-user DoF of the elements in the subset $\mathcal{S}$. Next, we bound the per-user DoF of elements in the subset $\bar{\mathcal{S}} = \mathcal{K} \setminus \mathcal{S}$, i.e. the set of users which not belong to $\mathcal{S}$.

1. Per-user DoF bounding of the users $i \in \mathcal{S}$. First, we consider the user $i = s_1$. We consider the inequality in (2.26) for the specific subset $\mathcal{G} = \mathcal{S} \setminus \{s_1\}$ and the equality in (2.27), i.e.

$$\sum_{j \in \mathcal{S} \setminus \{s_1\}} d_j \leq 1 + \sum_{j \in \mathcal{S} \setminus \{s_1, s_2\}} \beta_j \tag{2.28a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.28b}$$

By comparing the inequality with the equality, it follows that $d_{s_1} \geq \beta_{s_2}$. We then move to the case $i \in \mathcal{S} \setminus \{s_1\}$. Here, we consider the inequality in (2.26) for the specific subset

$\mathcal{G} = \mathcal{S} \setminus \{i\}$ and the equality in (2.27). By comparing the two we obtain

$$\sum_{j \in \mathcal{S} \setminus \{i\}} d_j \leq 1 + \sum_{j \in \mathcal{S} \setminus \{s_1, i\}} \beta_j \tag{2.29a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.29b}$$

Hence, it follows that $d_i \geq \beta_i$. To summarize, we have obtained that for $i \in \mathcal{S}$ we have $d_{s_1} \geq \beta_{s_2}$, and that for $i \in \mathcal{S} \setminus \{s_1\}$ we have $d_i \geq \beta_i$.

2. Per-user DoF bounding of the users $i \in \bar{\mathcal{S}}$, where $\bar{\mathcal{S}} = \mathcal{K} \setminus \mathcal{S}$. The set $\bar{\mathcal{S}}$ is partitioned into three subsets, denoted as $\bar{\mathcal{S}}_1$, $\bar{\mathcal{S}}_2$ and $\bar{\mathcal{S}}_3$. The subset $\bar{\mathcal{S}}_1$ contains all users listed before s_1, hence $\bar{\mathcal{S}}_1 = \{\, i \in \bar{\mathcal{S}} \mid i < s_1 \,\}$. The subset $\bar{\mathcal{S}}_2$ contains all users listed between s_1 and s_2, hence $\bar{\mathcal{S}}_2 = \{\, i \in \bar{\mathcal{S}} \mid s_1 < i < s_2 \,\}$. Finally, the subset $\bar{\mathcal{S}}_3$ contains all users listed after s_2, hence it is given by $\bar{\mathcal{S}}_3 = \{\, i \in \bar{\mathcal{S}} \mid i > s_2 \,\}$.

   We analyse the subset $i \in \bar{\mathcal{S}}_1$ first. By taking any user $i \in \bar{\mathcal{S}}_1$, we first compare the inequality in (2.26) for the case $\mathcal{G} = \mathcal{S} \cup \{i\}$ and the equality in (2.27), i.e.

$$\sum_{j \in \mathcal{S} \cup \{i\}} d_j \leq 1 + \sum_{j \in \mathcal{S}} \beta_j \tag{2.30a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.30b}$$

   It follows that $d_i \leq \beta_{s_1}$. We then compare the inequality (2.26) for the subset $\mathcal{G} = (\mathcal{S} \cup \{i\}) \setminus \{s_1\}$ and the equality in (2.27), i.e.

$$\sum_{j \in (\mathcal{S} \cup \{i\}) \setminus \{s_1\}} d_j \leq 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j \tag{2.31a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.31b}$$

   It follows that $d_i \leq d_{s_1}$. Hence, $d_i \leq \min(\beta_{s_1}, d_{s_1})$ for $i \in \bar{\mathcal{S}}_1$.

   We then move to the case $i \in \bar{\mathcal{S}}_2$. Proceeding as above, by comparing (2.26) for the case $\mathcal{G} = \mathcal{S} \cup \{i\}$ and (2.27), we obtain

$$\sum_{j \in \mathcal{S} \cup \{i\}} d_j \leq 1 + \beta_i + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j \tag{2.32a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.32b}$$

It follows that $d_i \leq \beta_i$. Also, from (2.26) for $\mathcal{G} = (\mathcal{S} \cup \{i\}) \setminus \{s_1\}$ and (2.27), we obtain

$$\sum_{j \in (\mathcal{S} \cup \{i\}) \setminus \{s_1\}} d_j \leq 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j \tag{2.33a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.33b}$$

It follows that $d_i \leq d_{s_1}$. To summarize, $d_i \leq \min(\beta_i, d_{s_1})$ for $i \in \bar{\mathcal{S}}_2$.

As a last case, we consider $i \in \bar{\mathcal{S}}_3$. By simply comparing (2.26) for $\mathcal{G} = \mathcal{S} \cup \{i\}$ with (2.27), we obtain that

$$\sum_{j \in \mathcal{S} \cup \{i\}} d_j \leq 1 + \beta_i + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j \tag{2.34a}$$

$$\sum_{j \in \mathcal{S}} d_j = 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j. \tag{2.34b}$$

It follows that $d_i \leq \beta_i$ for $i \in \bar{\mathcal{S}}_3$.

By summarizing the analysis above, we can conclude that the facet $\mathcal{F}_\mathcal{S}$ is included in the set of all the non-negative real tuples $(d_1, \ldots, d_k)$ characterized by

$$\begin{cases} d_{s_1} \geq \beta_{s_2} & \\ d_i \geq \beta_i, & i \in \mathcal{S} \setminus \{s_1\} \\ d_i \leq \min(\beta_{s_1}, d_{s_1}), & i \in \bar{\mathcal{S}}_1 \\ d_i \leq \min(\beta_i, d_{s_1}), & i \in \bar{\mathcal{S}}_2 \\ d_i \leq \beta_i, & i \in \bar{\mathcal{S}}_3 \\ \sum_{i \in \mathcal{S}} d_i = 1 + \sum_{i \in \mathcal{S} \setminus \{s_1\}} \beta_i. & \end{cases} \tag{2.35}$$

Moreover, simple calculations also verify that each tuple $(d_1, \ldots, d_k)$ in (2.35) satisfies the conditions in (2.26) and (2.27). Hence, (2.35) is equivalent to (2.26) and (2.27). It follows that $\mathcal{F}_\mathcal{S}$ coincides with the set of tuples described by the inequalities in (2.35). We have consequently rewritten the parametrization form of the facet $\mathcal{F}_\mathcal{S}$ from equations (2.26) and (2.27) to the form in (2.35), where the latter bounds the per-user DoF of each user. We next show that this latter form is suitable to show the achievability by rate-splitting.

Showing the Achievability of the Facets $\mathcal{F}_\mathcal{S}$ for $|\mathcal{S}| \geq 2$

After having re-parametrized the facet $\mathcal{F}_\mathcal{S}$ in a suitable form, we can show its achievability. We split $\mathcal{F}_\mathcal{S}$ into two subsets, denoted by $\mathcal{F}_{\mathcal{S},1}$ and $\mathcal{F}_{\mathcal{S},2}$, on the basis of the value of d_{s_1}, i.e. the per-user DoF achieved by user s_1. The subset $\mathcal{F}_{\mathcal{S},1}$ contains all the tuples of $\mathcal{F}_\mathcal{S}$ such that $\beta_{s_1} \geq d_{s_1} \geq \beta_{s_2}$, while $\mathcal{F}_{\mathcal{S},2}$ contains all the tuples of $\mathcal{F}_\mathcal{S}$ such that $d_{s_1} > \beta_{s_1}$. We start by showing the achievability of $\mathcal{F}_{\mathcal{S},1}$. We have that $\mathcal{F}_{\mathcal{S},1}$ is given by

$$\begin{cases} \beta_{s_1} \geq d_{s_1} \geq \beta_{s_2} & \\ d_i \geq \beta_i, & i \in \mathcal{S} \setminus \{s_1\} \\ d_i \leq d_{s_1}, & i \in \bar{\mathcal{S}}_1 \\ d_i \leq d_{s_1}, & i \in \bar{\mathcal{S}}_{21} \\ d_i \leq \beta_i, & i \in \bar{\mathcal{S}}_{22} \\ d_i \leq \beta_i, & i \in \bar{\mathcal{S}}_3 \\ \sum_{i \in \mathcal{S}} d_i = 1 + \sum_{i \in \mathcal{S} \setminus \{s_1\}} \beta_i & \end{cases} \tag{2.36}$$

where, for any value of d_{s_1}, the subsets $\bar{\mathcal{S}}_{21}$ and $\bar{\mathcal{S}}_{22}$ are defined as $\bar{\mathcal{S}}_{21} = \{\, i \in \bar{\mathcal{S}}_2 \mid \beta_i \geq d_{s_1} \,\}$ and $\bar{\mathcal{S}}_{22} = \{\, i \in \bar{\mathcal{S}}_2 \mid \beta_i < d_{s_1} \,\}$ and they correspond to a partition of $\bar{\mathcal{S}}_2$ on the basis of the value of β_i compared to d_{s_1}. In practice this means that, for any tuple in the facet with a specific value of d_{s_1}, we have that $\bar{\mathcal{S}}_{21}$ is the subset of users of $\bar{\mathcal{S}}_2$ with a CSIT quality larger than or equal to d_{s_1}, while $\bar{\mathcal{S}}_{22}$ is the subset of users of $\bar{\mathcal{S}}_2$ with a CSIT quality lower than d_{s_1}.

Each admissible tuple $(d_1, \ldots, d_k)$ above of $\mathcal{F}_{\mathcal{S},1}$ is then achieved by rate-splitting considering a power allocation $\mathbf{a} = (a_1, \ldots, a_k)$ given by

$$a_i = \begin{cases} d_{s_1}, & j \in \mathcal{S} \\ d_i, & j \in \bar{\mathcal{S}}_1 \\ d_i, & j \in \bar{\mathcal{S}}_{21} \\ d_i + d_{s_1} - \beta_i, & i \in \bar{\mathcal{S}}_{22} \\ d_i + d_{s_1} - \beta_i, & i \in \bar{\mathcal{S}}_3. \end{cases} \tag{2.37}$$

In fact, with theis power allocation, the DoF $(d_1^{(\mathrm{p})}, \ldots, d_k^{(\mathrm{p})})$ carried by each private codeword can be computed from Eq. (2.12) and it is given by

$$d_i^{(\mathrm{p})} = \begin{cases} d_{s_1}, & i = s_1 \\ \beta_i, & i \in \mathcal{S} \setminus \{s_1\} \\ d_i, & i \in \bar{\mathcal{S}}. \end{cases} \tag{2.38}$$

The common codeword's DoF, which can be calculated to be equal to $d^{(\mathrm{c})} = 1 - d_{s_1}$ from Eq. (2.11), is partitioned in the following way

$$d_i^{(\mathrm{c})} = \begin{cases} 0, & i = s_1 \\ d_i - \beta_i, & i \in \mathcal{S} \setminus \{s_1\} \\ 0, & i \in \bar{\mathcal{S}}. \end{cases} \tag{2.39}$$

With such power allocation and split of the common codeword, equality in (2.13) is satisfied for the tuple $(d_1, \ldots, d_k)$ of $\mathcal{F}_{\mathcal{S},1}$ and the achievability of the tuple $(d_1, \ldots, d_k)$ directly follows. Hence, $\mathcal{F}_{\mathcal{S},1}$ is achievable.

We can now prooced with the achievability proof for $\mathcal{F}_{\mathcal{S},2}$, which is given by $\mathcal{F}_{\mathcal{S}} \setminus \mathcal{F}_{\mathcal{S},1}$. We have that $\mathcal{F}_{\mathcal{S},2}$ is characterized by all non-negative real tuples $(d_1, \ldots, d_k)$ such that

$$\begin{cases} d_{s_1} > \beta_{s_1} & \\ d_i \geq \beta_i, & i \in \mathcal{S} \setminus \{s_1\} \\ d_i \leq \beta_{s_1}, & i \in \bar{\mathcal{S}}_1 \\ d_i \leq \beta_i, & i \in \bar{\mathcal{S}}_2 \\ d_i \leq \beta_i, & i \in \bar{\mathcal{S}}_3 \\ \sum_{i \in \mathcal{S}} d_i = 1 + \sum_{i \in \mathcal{S} \setminus \{s_1\}} \beta_i. & \end{cases} \tag{2.40}$$

Each tuple $(d_1, \ldots, d_k)$ of $\mathcal{F}_{\mathcal{S},2}$ is achieved by rate-splitting considering a power allocation $\mathbf{a} = (a_1, \ldots, a_k)$ given by

$$a_i = \begin{cases} \beta_{s_1}, & j \in \mathcal{S} \\ d_i, & i \in \bar{\mathcal{S}}_1 \\ d_i + \beta_{s_1} - \beta_i, & i \in \bar{\mathcal{S}}_2 \\ d_i + \beta_{s_1} - \beta_i, & i \in \bar{\mathcal{S}}_3. \end{cases} \tag{2.41}$$

With such power allocation, the DoF $(d_1^{(\mathrm{p})}, \ldots, d_k^{(\mathrm{p})})$ of each private codeword, from (2.12), is then given by

$$d_i^{(\mathrm{p})} = \begin{cases} \beta_i, & i \in \mathcal{S} \\ d_i, & i \in \bar{\mathcal{S}}. \end{cases} \tag{2.42}$$

The DoF carried by the common codeword, which can be computed to be equal to $d^{(\mathrm{c})} = 1 - \beta_{s_1}$ from (2.11), is partitioned in the following way

$$d_i^{(\mathrm{c})} = \begin{cases} d_i - \beta_i, & i \in \mathcal{S} \\ 0, & i \in \bar{\mathcal{S}}. \end{cases} \tag{2.43}$$

Equation (2.13) is satisfied and the tuple $(d_1, \ldots, d_k)$ of $\mathcal{F}_{\mathcal{S},2}$ is achievable. To summarize, as both $\mathcal{F}_{\mathcal{S}_1}$ and $\mathcal{F}_{\mathcal{S}_2}$ are achievable, the entire facet $\mathcal{F}_{\mathcal{S}}$, with $|\mathcal{S}| \geq 2$, is achievable by rate-splitting.

Showing the Achievability of the Facets $\mathcal{F}_{\mathcal{S}}$ for $|\mathcal{S}| = 1$

Next, we move to the case $|\mathcal{S}| = 1$, i.e. $\mathcal{S} = \{s_1\}$. The set $\bar{\mathcal{S}} = \mathcal{K} \setminus \mathcal{S}$ is partitioned into two subsets, denoted as $\bar{\mathcal{S}}_1$ and $\bar{\mathcal{S}}_2$, such that $\bar{\mathcal{S}}_1 = \{\, i \in \bar{\mathcal{S}} \mid i < s_1 \,\}$ and $\bar{\mathcal{S}}_2 = \{\, i \in \bar{\mathcal{S}} \mid i > s_1 \,\}$. Hence, $\bar{\mathcal{S}}_1$ contains all the users which precede s_1 while $\bar{\mathcal{S}}_2$ contains all users which proceed s_1. We start by considering $\bar{\mathcal{S}}_1$ and, for $i \in \bar{\mathcal{S}}_1$, by comparing (2.26) for $\mathcal{G} = \{i, s_1\}$ and (2.27), we obtain

$$d_i + d_{s_1} \leq 1 + \beta_{s_1} \tag{2.44a}$$

$$d_{s_1} = 1. \tag{2.44b}$$

We deduce that $d_i \leq \beta_{s_1}$. Similarly, in case of $i \in \bar{\mathcal{S}}_2$, by comparing (2.26) for $\mathcal{G} = \{s_1, i\}$ and (2.27), we obtain

$$d_{s_1} + d_i \leq 1 + \beta_i \tag{2.45a}$$

$$d_{s_1} = 1. \tag{2.45b}$$

Hence, we deduce that $d_i \leq \beta_i$.

It follows that $\mathcal{F}_{\mathcal{S}}$ can be rewritten as the set of all the non-negative real tuples $(d_1, \dots, d_k)$ given by

$$\begin{cases} d_{s_1} = 1 \\ d_i \leq \beta_{s_1}, & i \in \bar{\mathcal{S}}_1 \\ d_i \leq \beta_i, & i \in \bar{\mathcal{S}}_2. \end{cases} \tag{2.46}$$

Each $(d_1, \dots, d_k)$ is achieved by rate-splitting with power allocation $\mathbf{a} = (a_1, \dots, a_k)$ given by

$$a_i = \begin{cases} \beta_{s_1}, & i = s_1 \\ d_i, & i \in \bar{\mathcal{S}}_1 \\ d_i + \beta_{s_1} - \beta_i, & i \in \bar{\mathcal{S}}_2. \end{cases} \tag{2.47}$$

The common codeword's DoF, which is equal to $d^{(\mathrm{c})} = 1 - \beta_{s_1}$, is given to user s_1 only, i.e. the partition is such that $d_{s_1}^{(\mathrm{c})} = d^{(\mathrm{c})}$ and $d_i^{(\mathrm{c})} = 0$ for $i \in \mathcal{K} \setminus \{s_1\}$ and this guarantees that $d_{s_1} = 1$. Equation (2.13) is satisfied and the tuple $(d_1, \dots, d_k)$ of $\mathcal{F}_{\mathcal{S}}$ is achievable. It follows that $\mathcal{F}_{\mathcal{S}}$, with $|\mathcal{S}| = 1$, is achievable.

2.8.4. Proof of the Achievability of the Facets delimited by the Half-Spaces in Eq. (2.21)

We finally consider the facets contained in the hyperplanes which delimit the half-spaces in (2.21). Taking any $i \in \mathcal{K}$, we denote the facet contained in the hyperplane $d_i = 0$ as $\mathcal{F}_i^{(0)}$. After removing the redundant inequalities, $\mathcal{F}_i^{(0)}$ is given by all the non-negative real tuples $(d_1, \dots, d_k)$ which satisfy

$$d_i = 0 \tag{2.48a}$$

$$\sum_{j \in \mathcal{S}} d_j \leq 1 + \sum_{j \in \mathcal{S} \setminus \{s_1\}} \beta_j, \quad \forall \mathcal{S} \in \bar{\mathcal{A}}_i \tag{2.48b}$$

where $\bar{\mathcal{A}}_i$ is the set of all possible non-empty subsets of $\mathcal{K}\backslash\{i\}$ with elements arranged in an ascending order. For instance, in case of $\mathcal{K} = \{1, 2, 3\}$ and $i = 1$, we have that $\bar{\mathcal{A}}_i = \{\{2\}, \{3\}, \{2, 3\}\}$. While $d_i = 0$ (so user i is not considered), the set of admissible tuples $(d_j)_{j\in\mathcal{K}\backslash\{i\}}$ corresponds to the region in (2.21) and (2.22) when considering the $k-1$ users $\mathcal{K} \setminus \{i\}$. Since we have k antennas and $k - 1$ users, the facet $\mathcal{F}_i^{(0)}$ is achievable by induction hypothesis. For instance, this can be shown by shutting down one of the transmitting antennas and reducing the setting to the case of $k - 1$ transmitting antennas and $k - 1$ users, which is achievable by induction hypothesis.

To conclude, since all facets of the polyhedral region are achievable, all the remaining points of the polyhedral region are achievable by time-sharing. Hence, the outer-bound $\mathcal{D}$ for $K = k$ is achievable and it coincides with the optimal DoF region $\mathcal{D}$.

2.8.5. Alternative Proof for the Achievability

A alternative proof for the achievability argument has been recently proposed in the works in [125, 128]. In our approach, we build the polyhedral outer-bound first and we then exhaustively characterize and show the achievability of all its facets. In their work, the authors of [125, 128] considered the opposite approach. Instead of starting from the outer-bound and showing its achievability, they described the rate-splitting achievable region first and they then showed that it coincides with the outer-bound. This latter result was obtained through a mathematical procedure called inductive Fourier-Motzkin elimination scheme. Both our and their approaches have advantages and disadvantages. Our approach dives deeper into the design of variables such as power allocation but it requires an exhaustive characterization of all cases, on the other hand the approach in [125, 128] does not offer this more deeper understanding but it avoids an explicit construction of specific rate-splitting strategies for all cases.

2.9. Summary of the Chapter

In this chapter we have made progress in the characterization of the fundamental limits of robust interference management under full transmitter cooperation. By considering the K-user MISO BC with arbitrary CSIT levels, building upon previous works which have derived the optimal sum-DoF, we have characterized the optimal DoF region. Moreover, we have shown that rate-splitting is the key strategy to achieve this region. The essence of rate-splitting, compared to conventional transmission techniques as zero-forcing which rely on the transmission of private codewords only, is the transmission of a common codeword on top of the private codewords. The presence of the common codeword allows to tackle the multi-user interference originating from the partial CSIT more efficiently and, considering a flexible power allocation for the private codewords and flexible split of the DoF of the common codeword, to achieve the entire DoF region. Rate-splitting boils down to zero-forcing in case of perfect CSIT for all users, where the common message becomes unnecessary and zero-forcing is sufficient to achieve the whole DoF region.

3. Overloaded Multiuser MISO Transmission with Imperfect CSIT

3.1. Overview of the Chapter

In this chapter we make progress towards the understanding of the fundamental limits of robust interference management under full transmitter cooperation in the context of an overloaded MISO BC, where the number of users is larger than the number of antennas at the transmitter. This problem, which extends the setup in Chapter 2 where an equal number of transmitting antennas and users was assumed, is motivated by the fact that a required feature for the next generation of wireless communication networks will be the capability to serve simultaneously a large number of devices with heterogeneous CSIT qualities and demands.

In particular, we consider an overloaded MISO BC with two groups of CSIT qualities. One group has a CSIT quality $\beta > 0$, while the other group has a CSIT quality $\beta = 0$. The main contribution of this chapter is two-fold: 1) we first propose a transmission scheme where no CSIT codewords are superimposed on top of spatially-multiplexed codewords. The developed strategy allows to serve all users in a non-orthogonal manner and the analysis shows an enhanced perfomance compared to existing schemes. 2) We then characterize the optimal DoF region in the considered setting, by employing a dual argument based on converse and achievability similar to the one in Chapter 2.

3.2. Introduction

The ability to simultaneously support a tremendous number of devices with heterogeneous demands and capabilities is amongst the various features envisioned for future wireless networks [1]. Hence, it is expected that many networks will operate in overloaded regimes, roughly described as scenarios where the number of messages exceeds the number of transmitting antennas. One fundamental example is captured by the Single-Input-Single-Output (SISO) Broadcast Channel (BC), widely studied in literature. However, insights drawn from such studies are deemed insufficient when considering multiple antennas, as the SISO BC is robust against CSIT inaccuracies due to its degraded nature. On the other hand, the study of overloaded multiantenna channels is uncommon, e.g. works on the Multiple-Input-Single-Output (MISO) BC with imperfect CSIT consider a number of users less or equal to the number of transmitting antennas, as assumed in Chapter 2 or in the works in [13, 19, 21, 22].

3.2.1. An Overloaded MISO BC with Heterogeneous Partial CSIT

In this chapter, we extend the results in Chapter 2 and we make progress towards understanding the fundamental limits of overloaded multiantenna networks with heterogeneous partial CSIT. We consider a MISO BC comprising a transmitter equipped with K_{T} antennas, and $K_{\mathrm{R}} > K_{\mathrm{T}}$ single-antenna receivers (or users) indexed by $\mathcal{K}_{\mathrm{R}} = \{1, \ldots, K_{\mathrm{R}}\}$. While a general heterogeneous setup would consider arbitrary CSIT qualities, we restrict the analysis to the case where partial CSIT for K_{T} of the K_{R} users is available ($\beta_k > 0$), while no CSIT is available for the remaining $K_{\mathrm{R}} - K_{\mathrm{T}}$ users ($\beta_k = 0$)[1]. This assumption is introduced in order to make the problem analytically tractable, in particular to simplify the derivation of the sum-DoF upper-bound as well as the DoF region outer-bound. We further simplify the analysis by considering a symmetric scenario where all users with partial CSIT have the same quality β. It is implicitly understood that the CSIT quality is defined as in Section 1.1.2 and Section 2.3.1. Such setup is sufficient to gain some insights into the structure of the DoF-optimal transmission scheme and the influence of heterogeneous partial CSIT. Before we proceed, let us denote the groups of receivers by $\mathcal{K}_{\beta}$ and $\mathcal{K}_{0}$, where the subscript indicates the CSIT quality.

3.2.2. Main Contributions: Time Partitioning versus Power Partitioning

In the presence of only one of the two groups $\mathcal{K}_{\beta}$ and $\mathcal{K}_{0}$, DoF-optimal schemes are known. As widely discussed Chapter 2, the optimal sum-DoF for group $\mathcal{K}_{\beta}$ is achieved through rate-splitting, which relies on the transmission of a degraded common codeword on the top of the classical zero-forced private codewords [45]. On the other hand, the absence of CSIT results in a collapse of the sum-DoF to unity [22], and the degraded layer becomes sufficient to achieve the DoF of group $\mathcal{K}_{0}$. As a baseline, we consider the case where the two groups are served independently through orthogonal time partitioning (or sharing). We show that such strategy is in fact suboptimal in a DoF sense by proposing a superior strategy.

We propose a transmission scheme where the signals carrying the messages of groups $\mathcal{K}_{0}$ and $\mathcal{K}_{\beta}$ are superimposed and separated in the power domain. Users in $\mathcal{K}_{0}$ decode their codewords by treating the interference caused by the signals intended to $\mathcal{K}_{\beta}$ as noise. On the other hand, users in $\mathcal{K}_{\beta}$ first decode the codewords intended to $\mathcal{K}_{0}$ (without hurting their DoF!), and then proceed to decode their own codewords. Contrary to the orthogonal time partitioning, this leads to a non-orthogonal power partitioning. First, we show that such strategy achieves a strict DoF gain over time partitioning when users in each group achieve a per-user symmetric-DoF. Second, we show that this strategy in fact achieves the optimal DoF region for the considered overloaded MISO BC. Third, we show using simulations that the DoF gains achieved through power partitioning over time partitioning manifest in the finite SNR regime as significant achievable rate gains.

[1] In this chapter, as in Chapter 2, no CSIT implies that the transmitter has no (or finite precision [22]) information about the channel direction. However, the channel gain (or long term SNR) is known to guarantee reliable communication.

3.3. A Time Partitioning approach

Since group $\mathcal{K}_\beta$ has (partial) CSIT and group $\mathcal{K}_0$ has no CSIT, it seems natural to partition the time resource and carry out the transmission over two phases. In particular, the first phase occupies a fraction $\lambda \in [0, 1]$ of the time in which group $\mathcal{K}_\beta$ is served using a multiuser scheme that leverages partial CSIT and achieves spatial-multiplexing gains. On the other hand, the second phase occupies the remaining $1-\lambda$ fraction of the time in which group $\mathcal{K}_0$ is served with no multiplexing gains due to to the absence of CSIT. This time partitioning scheme acts as a baseline for the scheme proposed in the following section. Moreover, the two phases are in fact used as basic building blocks to construct the proposed scheme. Next, we describe the two phases in more detail.

Phase 1

For the first phase where users with CSIT are served, we adopt the rate-splitting scheme described in Section 2.5, as in fact optimal for this scenario. For completeness, we briefly revisit rate-splitting here. Users $k \in \mathcal{K}_\beta$ split their respective messages into $\big(W_k^{(\mathrm{p})}, W_k^{(\mathrm{c})}\big)$, where $W_k^{(\mathrm{p})}$ is a private sub-message and $W_k^{(\mathrm{c})}$ is a common (or public) sub-message. The sub-message $W_k^{(\mathrm{p})}$ is encoded into the private codeword $X_k^{(p)}$ decoded only by user k, while $W_1^{(c)}, \ldots, W_{K_\mathrm{R}}^{(c)}$ are jointly encoded into the common codeword $X^{(c)}$ decoded by all users in $\mathcal{K}_\beta$. It is assumed that all codewords are drawn from Gaussian codebooks with unitary powers. All codewords are linearly precoded and power allocated and the transmitted signal at the t-th channel use is given by

$$\mathbf{X}(t) = \sqrt{P^{(\mathrm{c})}}\mathbf{V}^{(\mathrm{c})}(t)X^{(\mathrm{c})}(t) + \sum_{k\in\mathcal{K}_\beta}\sqrt{P_k^{(\mathrm{p})}}\mathbf{V}_k^{(\mathrm{p})}(t)X_k^{(\mathrm{p})}(t) \tag{3.1}$$

where $\mathbf{V}^{(\mathrm{c})}(t) \in \mathbb{C}^{K_\mathrm{T}\times 1}$ and $\mathbf{V}_k^{(\mathrm{p})}(t) \in \mathbb{C}^{K_\mathrm{T}\times 1}$ are unitary precoding vectors, and $P^{(\mathrm{c})}$ and $P_k^{(\mathrm{p})}$ are the corresponding allocated powers with $P^{(\mathrm{c})}+\sum_{k\in\mathcal{K}_\beta} P_k^{(\mathrm{p})} \leq P$. Since the common codeword is decoded by all users, $\mathbf{V}^{(\mathrm{c})}(t)$ is chosen as a random (or generic) precoding vector. On the other hand, the private codewords are precoded by zero-forcing over the channel estimate, i.e. $\mathbf{V}_k^{(\mathrm{p})}(t) \perp \big\{\hat{\mathbf{G}}_l(t)\big\}_{l\in\mathcal{K}_\beta\setminus k}$. The power allocation is set such that $P^{(\mathrm{c})} = O(P)$ and $P_k^{(\mathrm{p})} = O(P^\beta)$.

Following the steps explained in Section 2.5, all users decode the common codeword by treating the interference from all private codewords as noise, from which the Signal to Interference plus Noise Ratio (SINR) scales as $O(P^{1-\beta})$. This is followed by removing the common codeword, and then each receiver decodes its private codeword with SINR of $O(P^\beta)$. Normalized by the time partitioning factor λ, the DoF achieved by the common codeword is given by $1-\beta$, while each private codeword achieves a DoF of β [45]. Hence, the per-user symmetric normalized DoF achieved by evenly sharing the common codeword is given by $\frac{1+(K_\mathrm{T}-1)\beta}{K_\mathrm{T}}$.

Phase 2

In the second phase, users $k \in \mathcal{K}_0$ are served. Since all users have no CSIT, after normalizing by the time partition $1-\lambda$, the sum-DoF collapses to 1 [22], as widely seen in Chapter 2. This single normalized DoF can be shared in an orthogonal fashion using time-sharing or in a non-orthogonal fashion using superposition coding and SIC. From a DoF perspective, these two strategies achieve the same performance. Assuming superposition coding, messages are encoded into codewords and then precoded such that

$$\mathbf{X}(t) = \sum_{k \in \mathcal{K}_0} \sqrt{P_k} \mathbf{V}_k(t) X_k(t) \tag{3.2}$$

where X_k is an encoded codeword, $\mathbf{V}_k(t)$ is a random unitary precoding vector and P_k is the power allocation. Using an appropriate power allocation, it can be shown that the single normalized DoF can be split evenly amongst users such that each user achieves a normalized DoF of $\frac{1}{K_{\mathrm{R}} - K_{\mathrm{T}}}$.

Achievable DoF

It can be seen that within each phase (or group), power allocation is carried out such that users achieve symmetric normalized per-user DoF. By incorporating the time partitioning factor $\lambda \in [0, 1]$, the actual (non-normalized) per-user DoF achieved by the k-th user is given by

$$d_k = \begin{cases} \lambda \frac{1 + (K_{\mathrm{T}} - 1)\beta}{K_{\mathrm{T}}}, & k \in \mathcal{K}_\beta \\ (1 - \lambda) \frac{1}{K_{\mathrm{R}} - K_{\mathrm{T}}}, & k \in \mathcal{K}_0. \end{cases} \tag{3.3}$$

The time partitioning factor can be further optimized to achieve a symmetric-DoF amongst all users in the system, or any other tradeoff depending on the design objective.

3.4. A Power Partitioning Approach

In contrast to the time partitioning approach in the previous section, we propose a scheme based on power partitioning. For some partitioning factor $\Lambda \in [0, 1]$, the bottom Λ power levels are reserved for the transmission to $\mathcal{K}_\beta$ with partial CSIT, while the top $1 - \Lambda$ power levels are occupied by the transmission to $\mathcal{K}_0$ with no CSIT. It can be seen that power partition Λ in this scheme is reminiscent to the time partition λ in the previous scheme. Moreover, the transmitted signal is in fact a superposition of the signals in (2.6) and (3.2) such that

$$\begin{aligned} \mathbf{X}(t) = & \sqrt{P_0} \sum_{i \in \mathcal{K}_0} \sqrt{q_i} \mathbf{V}_i(t) X_i(t) + \sqrt{P^{(\mathrm{c})}} \mathbf{V}^{(\mathrm{c})}(t) X^{(\mathrm{c})}(t) \\ & + \sum_{k \in \mathcal{K}_\beta} \sqrt{P_k^{(\mathrm{p})}} \mathbf{V}_k^{(\mathrm{p})}(t) X_k^{(\mathrm{p})}(t) \end{aligned} \tag{3.4}$$

where codewords, precoding vectors and powers are as defined in the previous section. To highlight the power partitioning, we introduce P_0 which denotes the total power allocated to the signal intended to all users in $\mathcal{K}_0$. It follows that $q_i = P_i / P_0$ is the normalized power allocated user $i \in \mathcal{K}_0$. An example that illustrates the two scheme is given in Fig. 3.1.

Before we proceed to take a closer look at the power partitioning, it is useful to highlight that the

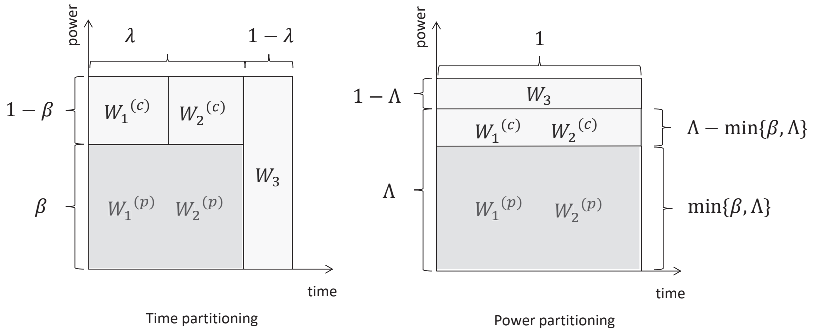


Figure 3.1.: Time partitioning and power partitioning for $K_{\mathrm{T}} = 2$ and $K_{\mathrm{R}} = 3$. Define the normalized spatial-multiplexing as the sum-DoF normalized by both the time partition and power partition. The normalized spatial-multiplexing gain in rectangles with light and dark shadings is 1 and 2 respectively.

signal received by the k users is expressed by

$$\begin{aligned} Y_k(t) =& \sqrt{P_0} \sum_{i \in \mathcal{K}_0} \sqrt{q_i} \mathbf{G}_k^{\mathrm{T}}(t) \mathbf{V}_i(t) X_i(t) + \sqrt{P^{(\mathrm{c})}} \mathbf{G}_k^{\mathrm{T}}(t) \mathbf{V}^{(\mathrm{c})}(t) X^{(\mathrm{c})}(t) \\ &+ \sum_{i \in \mathcal{K}_\beta} \sqrt{P_i^{(\mathrm{p})}} \mathbf{G}_k^{\mathrm{T}}(t) \mathbf{V}_i^{(\mathrm{p})}(t) X_i(t)^{(\mathrm{p})} + n_k(t), \end{aligned} \tag{3.5}$$

in which all different desired and interference components can be seen. In order to partition the signal-space through the power domain, the power allocation is carried out such that

$$\begin{cases} P_0 = O(P) \\ P^{(\mathrm{c})} + \sum_{k \in \mathcal{K}_\beta} P_k^{(\mathrm{p})} = O(P^{\Lambda}). \end{cases} \tag{3.6}$$

No CSIT Receivers

Users in $\mathcal{K}_0$ decode their messages by treating the interference (consisting of signals intended to users in $\mathcal{K}_\beta$) as noise. This is equivalent to raising the noise floor to P^{Λ} in Phase 2 of the previous section. Hence, the sum-DoF achieved by users in $\mathcal{K}_0$ is given by $1 - \Lambda$. Through an appropriate allocation of $\{q_i\}_{i \in \mathcal{K}_0}$, this DoF can be split evenly amongst users in $\mathcal{K}_0$.

Partial CSIT Receivers

As for users in $\mathcal{K}_\beta$, the same rate-splitting strategy of Phase 1 in the previous section is carried out where the power of $O(P^{\Lambda})$ is further split between the common codeword and the private codewords. In particular, the common codeword is allocated a power of $O(P^{\Lambda})$, while private codewords are allocated a power of $O(P^{a})$ where $a \leq \Lambda$. Before decoding their codewords, receivers first decode all codewords intended to users in $\mathcal{K}_0$ and remove them from the received signal. Since such messages are degraded already, the DoF achieved by users in $\mathcal{K}_0$ remains uninfluenced by this step. On the other hand, users in $\mathcal{K}_\beta$ now fully occupy the bottom Λ power levels.

Users in $\mathcal{K}_\beta$ now proceed to decode the common codeword as in Phase 1 of the previous section. This is received with a SINR of $O(P^{\Lambda-a})$, hence achieves a DoF of $\Lambda - a$. After removing the common codeword, each receiver decodes its private codeword with SINR of $O(P^a)$, achieving a DoF of a.

It remains to highlight that since the channel estimation error scales as $O(P^{-\beta})$, and due to zero-forcing, each receiver in $\mathcal{K}_\beta$ experiences an interference from the other private codewords that scales as $O(P^{a-\beta})$. This is drowned by noise if $a \leq \beta$. Knowing that $a \leq \Lambda$, we may set $a = \min\{\beta, \Lambda\}$. In other words, as long as the partition Λ satisfies $\Lambda \leq \beta$, users in $\mathcal{K}_\beta$ only need to rely on private messages using zero-forcing as interference can be drown by noise and rate-splitting is unnecessary. For partitions with $\Lambda > \beta$, zero-forcing is insufficient to neutralize interference, and rate-splitting becomes useful for users in $\mathcal{K}_\beta$. It follows that each private codeword achieves a DoF of $\min\{\beta, \Lambda\}$, while the common codeword achieves a DoF of $\Lambda - \min\{\beta, \Lambda\}$.

Remark 3.1. *The proposed scheme is a superposition of layers. The top layers consists of no CSIT codewords coming from $\mathcal{K}_0$ and the common rate-splitting codeword in $\mathcal{K}_\beta$, decoded by treating the bottom layer as noise, and removed using SIC. The bottom layer consist of spatially-multiplexed codewords carrying the remaining information for $\mathcal{K}_\beta$, which see no interference due to SIC of the top layers and zero-forcing up to the β-th power level.*

Achievable DoF

As in the previous section, we consider the case where users in each group achieve a symmetric-DoF. It follows that the DoF achieved by the k-th user is given by

$$d_k = \begin{cases} \frac{\Lambda + (K_\mathrm{T}-1)\cdot\min\{\beta,\Lambda\}}{K_\mathrm{T}}, & k \in \mathcal{K}_\beta \\ (1-\Lambda)\frac{1}{K_\mathrm{R}-K_\mathrm{T}}, & k \in \mathcal{K}_0. \end{cases} \tag{3.7}$$

Moreover, Λ can be optimized to achieve different tradeoffs.

Gain over Time Partitioning

Here we demonstrate that the power partitioning scheme achieves a DoF gain over the time partitioning scheme. Let $d_k^{(\mathrm{tp})}$ be the DoF achieved by the k-th user through time partitioning as in the previous section, i.e. obtained using (3.3) for some partition λ.

To highlight the DoF gains, let us consider the symmetric-DoF achieved by users in $\mathcal{K}_\beta$ through power partitioning given that users in $\mathcal{K}_0$ maintain the same per-user DoF as in time partitioning, i.e. $d_k = d_k^{(\mathrm{tp})}$ for all $k \in \mathcal{K}_0$. To achieve this, we need to set $\Lambda = \lambda$ in the power partitioning scheme. It follows from (3.7) that the per-user DoF of the remaining users is given by

$$d_k = \begin{cases} \frac{\lambda + (K_\mathrm{T}-1)\cdot\min\{\beta,\lambda\}}{K_\mathrm{T}}, & \beta \leq \lambda \\ \lambda, & \beta > \lambda \end{cases}, \text{ for all } k \in \mathcal{K}_\beta. \tag{3.8}$$

It can be seen that $d_k \geq d_k^{(\mathrm{tp})}$ for all $k \in \mathcal{K}$. For $k \in \mathcal{K}_0$, this follows directly from the design criteria. For the remaining users $k \in \mathcal{K}_\beta$, this follows by noting that for all $\beta, \lambda \leq 1$, we have $\frac{\lambda + (K_\mathrm{T}-1)\cdot\min\{\beta,\lambda\}}{K_\mathrm{T}} \geq \frac{\lambda + (K_\mathrm{T}-1)\cdot\beta\lambda}{K_\mathrm{T}}$. Moreover, this inequality is strict whenever $0 < \beta, \lambda < 1$, i.e. partial CSIT for $\mathcal{K}_\beta$ and non-zero (or unity) partitioning. Under such conditions, power partitioning achieves a strict improvement in the DoF of users in $\mathcal{K}_\beta$ over time partitioning.

To gain more insight into the DoF gain, consider the example shown in Fig. 3.1. It can be seen that the DoF achieved in each rectangle (a time-power resource block) is given by the rectangle's area times the normalized spatial-multiplexing gain (2 for zero-forcing and 1 for degraded). First, assume that user-3 is switched off. The sum-DoF achieved by the remaining two users through RS is given by $1 + \beta$. Now, introducing user-3 through time partitioning reduces the sum-DoF to $\lambda(1+\beta) + (1-\lambda) = 1 + \lambda\beta$. On the other hand, user-3 is introduced through power partitioning without harming the sum-DoF as long as $\Lambda = \lambda \geq \beta$. Keeping in mind that user-3 achieves the same DoF in both cases, it follows that user-1 and user-2 achieve higher DoF in the latter. For $\Lambda = \lambda < \beta$, introducing user-3 through power partitioning reduces the sum-DoF to $1+\lambda$. However, this is still higher than the sum-DoF of $1 + \lambda\beta$ achieved through time partitioning.

3.5. Optimal DoF Region

In the previous section, we considered the case where DoF tuples of the form $(d_\beta, \ldots, d_\beta, d_0, \ldots, d_0)$ are achieved, i.e. users in $\mathcal{K}_\beta$ achieve the per-user symmetric-DoF of d_β while users in $\mathcal{K}_0$ achieve the per-user symmetric-DoF of d_0. This gave some insight into the gains achieved through power partitioning as opposed to time partitioning. However, in more general scenarios, achievable DoF tuples assume a wide variety of tradeoffs characterized by achievable and optimal DoF regions, as we have seen in Chapter 2. Interestingly, the optimal DoF region for the considered setup is achieved through variants of the power partitioning scheme proposed in the previous section. This region is characterized in the following result.

Theorem 3.1. *For the overloaded MISO BC described in this chapter, the optimal DoF region $\mathcal{D}^*$ is given by*

$$d_k \geq 0, \quad \forall k \in \mathcal{K}_\mathrm{R} \tag{3.9}$$

$$\sum_{k \in \mathcal{S}} d_k + \sum_{k \in \mathcal{K}_0} d_k \leq 1 + (|\mathcal{S}| - 1)\beta, \quad \forall \mathcal{S} \subseteq \mathcal{K}_\beta, |\mathcal{S}| \geq 1. \tag{3.10}$$

The achievability of the DoF region is based on generalizing the power partitioning scheme of Section 3.4 by allowing arbitrary power allocations and splits of the common message. The achievability follows a similar philosophy than the one in Section 2.8. On the other hand, the converse is also based on the sum-DoF upper-bound in [22] as the converse in Section 2.7. The complete proof of Theorem 3.1 is given in the Appendix.

Note that from (3.10), we have $d_k \leq 1$ for all $k \in \mathcal{K}$ which is a trivial upper-bound for the per-user DoF, and $\sum_{k \in \mathcal{K}_0} d_k \leq 1$ which limits the sum-DoF of the no CSIT users to unity. To better visualize the optimal DoF region, an example is given in Fig. 3.2 (left) for a channel with

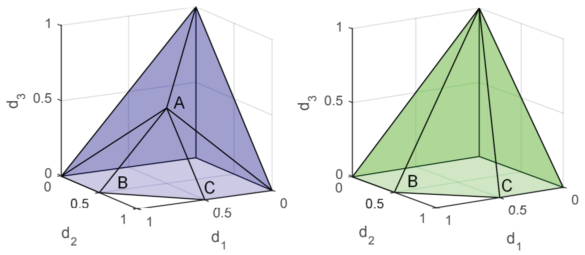


Figure 3.2.: DoF region achieved by power partitioning (left) and time partitioning (right) for $M = 2$ and $K = 3$, and CSIT quality $\beta = 0.5$ for the first two users. The points are $A = (\beta, \beta, 1 - \beta)$, $B = (1, \beta, 0)$ and $C = (\beta, 1, 0)$. It can be seen that A cannot be achieved through time partitioning.

$K_{\mathrm{T}} = 2$ and $K_{\mathrm{R}} = 3$, where the CSIT quality of the first two users is $\beta = 0.5$. Moreover, for the sake of comparison, the DoF region achieved through time partitioning is shown in Fig. 3.2 (right). The time partitioning region is obtained by time-sharing the DoF of 1 achieved by user-3 with the DoF region of the two remaining users achieved through rate-splitting (see [23]). For the power partitioning region, the facet given by $A - B - C$ is in fact sum-DoF optimal. Hence, user-3 can be served with non-zero DoF without influencing the Sum-DoF (e.g. point A). On the other hand, serving user-3 with non-zero DoF through time partitioning is not possible without decreasing the sum-DoF as it requires moving away from the segment $B - C$.

3.6. Numerical Results

In this section, we show that the obtained DoF gains translate into enhanced rate performances. We consider a MU-MISO scenario with $K_{\mathrm{T}} = 2$ antennas and $K_{\mathrm{R}} = 3$ users. Uncorrelated channels are assumed with entries drawn from $\mathcal{CN}(0, 1)$. Users 1 and 2 have CSIT qualities β, where channel estimation errors have entries drawn from $\mathcal{CN}(0, \sigma^2)$ with $\sigma^2 = P^{-\beta}$. On the other hand, the instantaneous CSIT of user 3 is unknown. In agreement to Sections 3.3 and 3.4, the precoding vector of the codeword intendend for user 3 as well as the precoding vector of the common rate-splitting codeword are chosen as unitary random vectors. On the other hand, the private codewords for users 1 and 2 are precoded by zero-forcing.

We numerically evaluate the ergodic sum-rate of the first two users achieved by power partitioning and time partitioning, while maintaining the ergodic rate of the third user to be the same in both cases. This is obtained by properly tuning, in the power partitioning approach, the power P_0 allocated to the codeword of the third user, while considering a RS strategy for the first two users. Fig. 3.3 shows the sum rate of user 1 and user 2 with respect to their long-term SNR for both power partitioning and time partitioning. We assume a scenario with $\beta = 0.5$ and we set the parameter $\lambda = 0.5$. We consider two cases where the SNR of users 1 and 2 is taken to be 10 dB, and then 20 dB, larger than the SNR of user 3. Since in time partitioning users 1 and 2 are scheduled separately from user 3, the difference in SNR only affects their sum rate performance in power partitioning. In the legend, we include this difference by brackets (only for power partitioning). From Fig. 3.3,

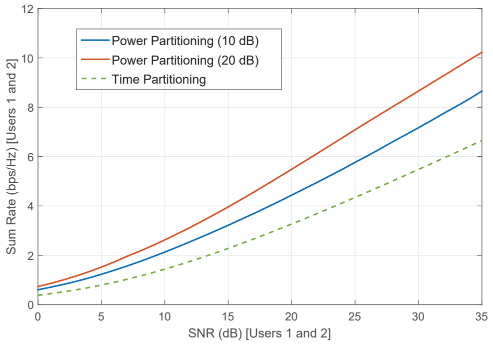


Figure 3.3.: Sum rate of user-1 and user-2 while maintaining the same rate for user-3 for the cases when the long-term SNR of user-3 is 10 dB and 20 dB lower than users-1,2. The parameters are set as $\beta = 0.5$ and $\lambda = 0.5$.

it is evident that our proposed power partitioning based approach significantly outperforms time partitioning in both cases. Furthermore, as the difference between the SNR of users 1 and 2 and the SNR of user 3 becomes larger, the rate gain increases which cannot be seen from DoF analysis.

3.7. Summary of the Chapter

In this chapter, we have considered an overloaded MISO BC where the transmitter has partial CSI for K_{T} users (equal to the number of antennas) and no-CSI for the remaining $K_{\mathrm{R}} - K_{\mathrm{T}}$, in a DoF sense. We have proposed a transmission scheme based on power partitioning and showed that it achieves strict DoF gains compared to a scheme where the two sets of users are independently served over orthogonal time slots. Moreover, we have showed that the optimal DoF region for such channel is in fact achieved by generalizing the proposed power partitioning scheme. This in turn extends the results in Chapter 2 to the overloaded regime. The finite SNR rate performance of the proposed DoF-motivated scheme is evaluated through simulations in which significant gains over time partitioning are demonstrated. This shows that such DoF-motivated design and analysis can be indeed very useful in guiding the design and optimization of more efficient practical transmission strategies.

4. Generalized Degrees of Freedom of the Symmetric Cache-Aided MISO BC with Partial CSIT

4.1. Overview of the Chapter

In this chapter we make progress towards the understanding of the fundamental limits of robust cache-aided interference management under full transmitter cooperation. We consider the K-user cache-aided MISO BC, where a K-antenna transmitter serves K single-antenna receivers, and each receiver is equipped with a cache memory. The transmitter has partial instantaneous knowledge of the channels of the users. It is readily seen that this setup is the cache-aided counterpart of the classical K-user MISO BC studied in Chapter 2, while taking into consideration the presence of caches across the network.

The performance metric utilized in this chapter is the Generalized Degrees of Freedom (GDoF), an extension of the DoF metric which takes into consideration the difference in channel strengths between the cross-links and the direct-links. Note that the GDoF framework is mostly understood for symmetric settings, and for this reason the DoF metric was utilized in Chapter 2 and 3.

The main result of this chapter, for a symmetric setting in terms of channel strength levels, partial channel knowledge levels and cache sizes, is the characterization of the sum-GDoF of the considered network up to a constant multiplicative factor. The achievability scheme exploits the interplay between spatial multiplexing gains and coded-multicasting gain. On the other hand, a cut-set-based argument in conjunction with a sum-GDoF upper-bound for a parallel MISO BC under channel uncertainty are used for the converse. We further show that the characterized order-optimal sum-GDoF is also attained in a decentralized setting, where no coordination is required for content placement in the caches.

4.2. Introduction and Main Contributions

In this chapter we consider the K-user cache-aided multiple-input-single-output broadcast channel (MISO BC), which is the extension of the classical K-user MISO BC to cache-aided setting. In particular, the K-user cache-aided MISO BC consists of a K-antenna transmitter which serves K single-antenna receivers, where each receiver is equipped with a cache memory. As already pointed out in Chapter 2 note that, even though such setup can model a radio cell where a multiple antenna BS is connected to multiple users, the transmit antennas in the considered setup are not necessarily physically co-located, and may generally represent K radio heads (or remote antennas) connected through a strong fronthaul. When CSIT is available with high accuracy, parallel non-interfering links can be created through zero-forcing. In this case, interference is completely managed through spatial pre-processing, and the usefulness of caches is restricted to local caching gains. However, this is not the case when only partial or imperfect CSIT is available as observed in [32, 36, 110].

From Chapter 2, we know that studying the classical MISO BC (with no caches) reveals that spatial multiplexing gains (i.e. DoF) of this channel suffer losses under imperfect CSIT. As we seen, the extreme case of finite precision CSIT causes a total collapse of the sum-DoF to 1, where all (DoF) benefits of multiple transmitting antennas are lost [22]. The availability of partial instantaneous CSIT can help salvage some of the lost gains, achieving sum-DoF between 1 and K depending on the CSIT quality. The complementary role of coded-caching in such scenarios was first observed in [110]. In particular, while the primary role of CSIT is to facilitate interference management (e.g. through zero-forcing), coded-caching reduces interference all together by creating multicasting opportunities. Hence, it was shown in [110] that coded-caching can offset the loss due to partial CSIT, up to a certain CSIT quality given the cache size.

The DoF metric, which we utilized in Chapter 2 and 3, however can be very pessimistic, as best exemplified by the sum-DoF collapse in [22]. This is mainly due the limitations of the DoF framework, assigning equal strengths to every link (with non-zero gain) in the wireless network. In a way, the DoF metric fails to capture one of the wireless channel's most important features: the propagation loss. This limitation is countered by the GDoF framework, which largely inherits the tractability of the DoF framework while capturing the diversity in channel strengths [41, 43, 66]. The GDoF framework has been mostly considered in the literature for symmetric settings, where all users have the same CSIT quality and/or channel strenghts. Extending to asymmetric settings poses many challenges and only limited results are known, and most of these results center around the two-user case only [41]. For this reason, the DoF metric was utilized in Chapter 2 and 3.

For a symmetric setup in terms of channel strenght, the cache-aided MISO BC was studied under the GDoF framework in [36], while limiting to completely absent CSIT and considering only achievability, with no guarantees on optimality[1]. In a different line of work, the cache-aided MISO BC under partial CSIT was considered while focusing on the massive MIMO regime [112]. In particular, [112] studies the delivery rate scaling laws, as the number of transmitting antennas grows arbitrarily large, using off-the-shelf caching strategies. While no guarantees on information-theoretic optimality are provided in the above work, the emphasis on the interplay between spatial multiplexing gains and coded-multicasting gains is very interesting. It turns out that this interplay, which was first noticed in [110] and then further investigated in [37, 112, 113], plays a central role in achieving and interpreting the order-optimal sum-GDoF of the cache-aided MISO BC under partial CSIT as we show through our results. Next, we highlight the main contributions of this chapter.

[1] The same can be said about [110], where the DoF under partial CSIT can be equivalently interpreted as the GDoF under no CSIT (see Section 4.4.1). No converse is given in [110], except for the trivial case where perfect CSIT is available.

4.2.1. Main Contributions and Organization

In this chapter, we consider the K-user cache-aided MISO BC within the (symmetric) GDoF framework, where the channel strength of cross-links is captured through the famous $\alpha \in [0, 1]$ parameter [41, 43, 66], which will be described in detail over the next sections. In addition, we capture the entire range of (symmetric) partial CSIT levels through the quality parameter $\beta \in [0, \alpha]$, where $\beta = 0$ and $\beta = \alpha$ correspond to essentially absent and perfect CSIT, respectively [43]. For this setting, the main contributions are twofold, as stated below:

1. We characterize the optimal sum-GDoF up to a constant multiplicative factor, which is independent of all system parameters. This order-optimal sum-GDoF characterization is derived while allowing central coordination during the placement phase of the achievability scheme.
2. We show that the order-optimal sum-GDoF, characterized under centralized placement, is also attained in decentralized settings where no coordination during the placement phase is allowed.

It is worthwhile highlighting that the order optimal schemes for the considered cache-aided MISO BC, for both the centralized and decentralized cases, abide by the *separation* principle [34]. In particular, the placement and generation of coded-multicasting messages are independent of the physical channel parameters (e.g. link strengths or topology), and follow the placement and message generation of the original shared-link Maddah-Ali and Niesen schemes explained in Section 1.2.1, first developed in the works in [24, 79]. On the other hand, the delivery of the coded-multicasting messages over the physical channel uses the principle of rate-splitting with common and private signalling, extensively explained and utilized in Chapters 2 and 3, which essentially operates the physical channel at some point of its multiple multicast GDoF region.

One of the technical challenges in characterizing the optimal sum-GDoF for the above setting is the converse, i.e. deriving an upper-bound which is within a constant multiplicative factor from the achievable sum-GDoF. Under partial CSIT, the conventional cut-set-based argument in [24] fails when employed on its own (see also [34, 106, 115] for variants of such argument). Alternatively, we derive an upper-bound by marrying the approach in [24] with a robust sum-GDoF upper-bound for a parallel MISO BC under partial CSIT, which in turn employs results from recent works by Davoodi and Jafar on classical networks (with no caches) under channel uncertainty [22, 41, 43], already mentioned in Chapter 2 and 3 when deriving robust outer-bounds of the DoF regions. Specifically, in this novel adaptation of the approach in [22, 41, 43] to cache-aided networks, caches at receivers are replaced with equivalent parallel side links, and then an upper-bound on the sum-GDoF of the resulting parallel sub-channels is derived.

Another technical challenge arises when dealing with the decentralized setting, particularly due to the intractable form of the sum-GDoF achieved under decentralized placement. This intractability is circumvented by observing that the decentralized achievable sum-GDoF is bounded below by a centralized-like achievable sum-GDoF, yet with a smaller coded-multicasting gain compared to the one achieved in a true centralized setting. This key observation enables us to prove order-optimality in the decentralized setting.

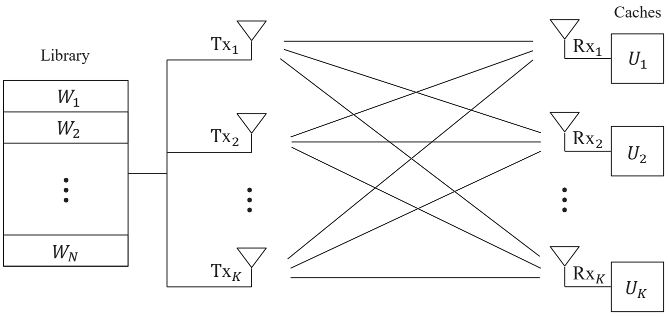


Figure 4.1.: A wireless network in which a transmitter with K antennas, $\mathrm{Tx}_1, \ldots, \mathrm{Tx}_K$, serves K single-antenna receivers, $\mathrm{Rx}_1, \ldots, \mathrm{Rx}_K$. The transmitter has access to a library of N files, while each receiver Rx_i is equipped with a cache memory $\mathcal{U}_i$.

In addition to the contributions highlighted above, we derive several insights from the optimal sum-GDoF characterization, which generalize former observations obtained in special cases of the considered setting [22, 36, 43, 110]. Such insights, and how they relate to previous observations, can be found in Section 4.4.1. As for the remainder of the chapter, the organization is as follows. Section 4.3 introduces the considered setting and problem. Section 4.4 presents the two main results and related insights. In Section 4.5, we derive an upper-bound which is employed in the following two sections to show order optimality. In Section 4.6 and Section 4.7, we prove the two main results, the centralized setting result and the decentralized setting result respectively. Section 4.8 summarizes and concludes the chapter.

4.3. Problem Setting

In this section we extend the setup in Chapter 2 to the cache-aided setting and to the GDoF framework. For completeness, we repeat some of the explanations in Section 2.3. We consider the MISO BC consisting of a K-antenna transmitter serving K receivers (or users), where users are equipped with a single-antenna each. Users are indexed by the set $[K] \triangleq \{1, 2, \ldots, K\}$. In a communication session, each user requests one file from a content library $\mathcal{W} \triangleq \{W_1, \ldots, W_N\}$ consisting of $N \geq K$ files, each of size F bits. We assume that the transmitter has access to the entire library (this applies to each radio head, or remote antenna, in physically distributed settings).

At the receiving end of the channel, each user i is equipped with a cache memory $\mathcal{U}_i$ of size MF bits, where $M \in [0, N]$. We define the *normalized cache size* as

$$\mu \triangleq \frac{M}{N} \tag{4.1}$$

which is interpreted as the fraction of the content library each user is able to store locally. An illustration of the setup is given in Fig. 4.1. It is readily seen that $\mu = 0$ reduces the setup to the classical MISO BC, while no communication needs to take place under $\mu = 1$. We refer to the j-th transmit antenna (or radio head) as the *j-th transmitter* to emphatize the different channel gains between each antenna and the different users, while *transmitter* refers to the K transmit antennas jointly.

As previously described in Section 1.2.1, the network operates in two phases, *a placement phase* and a *delivery phase* [24]. The placement phase takes place during the off-peak times before knowing the future demands of different users. During this phase, the cache memories of the users are filled as an arbitrary function of the N files, where such function is denoted as $\mathcal{U}_i = \phi_i(\mathcal{W})$. The delivery phase takes place during peak times where each user requests one of the N files. For example, user i requests file W_{d_i} for some $d_i \in [N]$, where $\mathbf{d} = (d_1, \ldots, d_K)$ is the tuple of all user demands. Upon receiving the requests, each transmitter j sends a codeword $X_j^T = X_j(1), \ldots, X_j(T)$ over $T \in \mathbb{N}$ uses of the *physical channel*. At the other end, each user i receives the sequence $Y_i^T = Y_i(1), \ldots, Y_i(T)$, a noisy linear combination of the K transmitted codewords. The user then decodes for its requested file from Y_i^T and the content of its own cache memory $\mathcal{U}_i$. This is described in more detail below.

4.3.1. Physical Channel

The input-output relationship at the t-th use of the physical channel, $t \in [T]$, is modeled by

$$Y_i(t) = \sum_{j=1}^{K} \sqrt{a_{ij}} G_{ij}(t) X_j(t) + Z_i(t) \tag{4.2}$$

where $Y_i(t) \in \mathbb{C}$ is the signal received by the i-th user, $X_j(t) \in \mathbb{C}$ is the j-th transmitter's normalized signal with power constraint $\mathbb{E}\left(|X_j(t)|^2\right) \leq 1$ and $Z_i(t) \sim \mathcal{N}_{\mathbb{C}}(0,1)$ is the normalized additive white Gaussian noise (AWGN), which is i.i.d. across all dimensions. $a_{ij} \in \mathbb{R}_+, \forall j, i \in [K]$, captures the long-term constant gain of the *link* between the j-th transmitter and the i-th receiver, while $G_{ij}(t) \in \mathbb{C}$ is the corresponding time-varying fading channel coefficient. Note that the terms a_{ij} are omitted in the DoF framework as the DoF metric assumes equal values of all the direct and cross links. To avoid degenerate situations, we assume that the instantaneous value $|G_{ij}(t)|$ is bounded away from zero and infinity for all $i, j \in [K]$ and $t \in [T]$.

GDoF Framework

For any $i, j \in [K]$ and $i \neq j$, we refer to the link between transmitter i and receiver i as a *direct-link*, while the link from transmitter j to receiver i is referred to as a *cross-link*. We consider a symmetric setup in which all direct-links (or cross-links) have similar long-term gains. For GDoF purposes, we introduce the nominal SNR value $P \in \mathbb{R}_+$, simply referred to as the SNR henceforth. Following the GDoF framework [41, 66], channel gains are expressed in terms of the SNR as

$$a_{ii} = P \ \text{ and } \ a_{ij} = P^{\alpha}, \ \forall i, j \in [K], \ i \neq j \tag{4.3}$$

where the parameter $\alpha \geq 0$ quantifies the strength of cross-links. The exponents of P in (4.3), i.e. 1 and α, are known as the channel strength parameters or levels. The channel model in (4.2) is rewritten as

$$Y_i(t) = \sqrt{P} G_{ii}(t) X_i(t) + \sum_{j=1, j \neq i}^{K} \sqrt{P^{\alpha}} G_{ij}(t) X_j(t) + Z_i(t) \tag{4.4}$$

which is the model used throughout the chapter. Note here the contrast with the DoF framework where all the direct and cross links are considered equal. The results in this chapter are restricted to the regime $\alpha \in [0, 1]$, i.e. scenarios in which the cross-link strength level is at most as strong as the direct-link strength level. This is the most practically relevant regime, since each receiver associates with a transmitter (i.e. radio head or remote antenna) from which it receives the strongest signal. Moreover, as highlighted in [43], the regime $\alpha > 1$ poses new challenges both in terms of achievability and upper-bounds and remains an open problem even for the classical MISO BC (with no caches) under partial CSIT. In the following paragraph, we remark the main difference between the GDoF and the DoF framework.

Remark 4.1. *As pointed out in [41], the scaling of P in the GDoF framework does not correspond to a physical scaling of transmitting powers in a given channel (or network). The correct interpretation is that each value of P defines a new channel. A class of channels parameterized by α belong together because the point-to-point capacity of any link (direct or cross) normalized by* $\log(P)$ *is approximately the same across all such channels belonging to the same class. Hence, unlike the DoF framework, the GDoF framework preserves the diversity in link strengths as $P \to \infty$. Moreover, DoF results are recovered from GDoF results by setting $\alpha = 1$, i.e, the special case in which all links are equally strong.*

Partial CSIT

The partial CSIT is modelled as for the DoF case, with the only exception that the quality parameter β is assumed to be in the range $[0, \alpha]$ instead of the range $[0, 1]$. Let $\mathcal{G} \triangleq \{G_{ij}(t) : i, j \in [K],\ t \in [T]\}$ be the set of all channel coefficient variables. Under partial CSIT, such channel coefficients may be represented as for the DoF framework

$$G_{ij}(t) = \hat{G}_{ij}(t) + \sqrt{P^{-\beta}} \tilde{G}_{ij}(t) \tag{4.5}$$

where $\hat{\mathcal{G}} \triangleq \{\hat{G}_{ij}(t) : i, j \in [K],\ t \in [T]\}$ are channel estimates, $\tilde{\mathcal{G}} \triangleq \{\tilde{G}_{ij}(t) : i, j \in [K],\ t \in [T]\}$ are estimation error terms and $\beta \in \mathbb{R}$ is a parameter capturing the CSIT quality level. The channel knowledge available to the transmitters includes the coarse channel strength level α, the CSIT quality level β and the estimates in $\hat{\mathcal{G}}$, but does not include the error terms in $\tilde{\mathcal{G}}$.

All variables in $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$ are subject to the bounded density assumption as explained in [41, 43]. The difference between $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$, as pointed out earlier, is that the former is revealed to the transmitters while the latter is not. Hence, given the estimates $\hat{\mathcal{G}}$, the variance of each channel coefficient in $\mathcal{G}$ behaves as $\sim P^{-\beta}$ and the peak of the probability density function behaves as $\sim \sqrt{P^{\beta}}$. We assume throughout this chapter that $\beta \in [0, \alpha]$. In particular, $\beta = 0$ and $\beta = \alpha$ capture the two extremes where channel knowledge at the transmitters is absent and perfectly available, respectively, and a value $\beta > \alpha$ would not lead to any improvement in the GDoF compared to

$\beta = \alpha$ [43]. Note here the difference with respect the DoF framework where cross-links and direct-links have the same channel strengths, hence $\beta = 1$ is needed to obtain full spatial multiplexing gains. However, in the GDoF framework, the difference in channel strength between the direct and cross links helps to obtain full spatial multiplexing gains already when $\beta = \alpha$. Before we proceed, it is worth highlighting that channel state information at the receivers (CSIR) is assumed here to be perfect. Moreover, in a slight abuse of notation, also here we henceforth use $\hat{\mathcal{G}}$ to denote the entire channel knowledge available to the transmitters.

4.3.2. Performance Measures

Once transmitters are informed of the demands $\mathbf{d}$ in the delivery phase, each transmitter j generates a sequence of T channel inputs $X_j^T = \psi_j^{(T)}(\mathcal{W}, \mathbf{d}, \mathcal{U}_1, \ldots, \mathcal{U}_K, \hat{\mathcal{G}})$, where $\psi_j^{(T)}$ is an encoding function. Note that the availability of partial CSIT is reflected in the argument $\hat{\mathcal{G}}$ of $\psi_j^{(T)}$. Once the transmission is complete, each user i maps its received signal, local cache content, user demands and perfect channel knowledge to an estimate of the requested file W_{d_i} denoted as $\hat{W}_i = \eta_i^{(T)}(Y_i^T, \mathcal{U}_i, \mathbf{d}, \mathcal{G})$, where η_i is the decoding function. The information theoretic limits of the system are studied by fixing N, K, M, P, and $\hat{\mathcal{G}}$, referred to as system parameters, while allowing F and T to grow arbitrarily large.

For fixed system parameters, a code which takes files of size F bits and transmits codewords of block-length T channel uses is defined as $\mathcal{C}^{(T)} \triangleq \left\{\phi_i, \psi_i^{(T)}, \eta_i^{(T)} : i \in [K]\right\}$. It is evident that a code is characterized by its corresponding caching, encoding and decoding functions defined earlier. The performance of a code is governed by its worst-case probability of error defined as

$$P_e^{(T)} \triangleq \max_{\mathcal{G}|\hat{\mathcal{G}}} \max_{\mathbf{d} \in [N]^K} \max_{i \in [K]} \mathbb{P}\left(\hat{W}_i \neq W_{d_i}\right), \tag{4.6}$$

which is taken over all possible users, for all possible demands, under all possible realizations of the channel coefficients given the available CSIT. The (sum) rate of such code is defined as

$$R \triangleq \frac{KF}{T}. \tag{4.7}$$

For given system parameters, we say that the rate R is achievable if there exists a coding scheme, consisting of a sequence of codes $\left\{\mathcal{C}^{(T)} : T \in \mathbb{N}\right\}$ of rate R each, with a vanishing probability of error as the block-length grows arbitrarily large, i.e. $P_e^{(T)} \to 0$ as $T \to \infty$. Note that a strictly positive rate $R > 0$ requires $F \to \infty$ as $T \to \infty$. The (sum) capacity C is defined as the supremum of all achievable rates taken over all feasible coding schemes.

Note that this corresponds to the definition of capacity in the classical sense, i.e. the rate is still defined in bits for channel use. The only thing here is that we have taken the caches into consideration when calculating the capacity, i.e. the delivered bits include both the bits transmitted over the air and the ones brought from the local caches.

GDoF

By highlighting the dependency on the SNR P, it can be seen that each P defines a new channel (or network) with capacity $C(P)$. The optimal sum-GDoF is hence defined as

$$\mathsf{GDoF} \triangleq \lim_{P\to\infty} \frac{C(P)}{\log(P)}. \tag{4.8}$$

Being an asymptotic (high-SNR) measure, it is well understood that the GDoF, as the DoF, does not depend on P. On the other hand, while fixing the number of users K, we often write $\mathsf{GDoF}(\mu, \alpha, \beta)$ to highlight the dependency on the system parameters μ, α and β. In particular, it turns out that our GDoF characterization is expressed in terms of the normalized cache size $\mu = M/N$ instead of the exact N and M, and the cross-link strength level α and partial CSIT level β instead of the entire CSIT $\hat{\mathcal{G}}$. These observations are consistent with existing DoF results for cache-aided networks on one hand [31, 32, 34], and GDoF studies in classical networks under finite precision and partial CSIT on the other hand [41, 43].

Generalized Normalized Delivery Time

Instead of working directly with the GDoF, it will be seen later that it is easier to derive the results in terms of a function of the reciprocal[2] $1/\mathsf{GDoF}$. Hence, we introduce the *generalized normalized delivery time* (GNDT), where the optimal GNDT is defined as

$$\mathsf{GNDT}(\mu, \alpha, \beta) \triangleq \frac{K}{\mathsf{GDoF}(\mu, \alpha, \beta)}. \tag{4.9}$$

The GNDT (or the *delivery time* as we refer to it throughout the chapter) is measured in *time-slots*. One time-slot is the optimal amount of time required to communicate a single file to a single user over a direct-link (with strength level 1) under no caching and no interference as $P \to \infty$. In particular, since a single user direct-link with no interference and no caching has a capacity of $\log(P) + o\big(\log(P)\big)$, i.e. $\mathsf{GDoF} = 1$, it is readily seen that $\mathsf{GNDT} = 1$ time-slot for such setting. For any given μ, α and β, as $\mathsf{GNDT}(\mu, \alpha, \beta)$ corresponds to the optimal delivery time, it follows that a delivery time $\mathsf{GNDT}'(\mu, \alpha, \beta)$ is achievable if and only if $\mathsf{GNDT}'(\mu, \alpha, \beta) \geq \mathsf{GNDT}(\mu, \alpha, \beta)$.

The GNDT generalizes the *normalized delivery time* (NDT) metric in [115] to suit the GDoF framework. Hence, it is not surprising to observe that the GNDT-GDoF relationship resembles (and generalizes) the NDT-DoF relationship. Moreover, it is readily seen from (4.9) that the GDoF can be interpreted as the capacity in files per time-slot. We would also like to highlight that in the approach followed in this chapter we first define the GDoF and then the *delivery time* as a function of the GDoF. However, in many works in the the literature (see, for instance, [32, 38, 107, 109, 115, 129] and references therein) the opposite approach is adopted, where the *delivery time* is defined first and then the DoF as a function of the *delivery time*.

Before we proceed, we remark that in this chapter, as in [24, 25, 30–32, 34, 36, 79, 106, 115], we

[2] This has been observed when dealing with the DoF in many works including [25, 30, 32, 34, 106].

adopt a worst-case definition of performance measures with respect to user requests. As a result, it is always assumed that each user requests a different file.

4.3.3. Centralized Placement vs. Decentralized Placement

Although the placement phase does not depend on the actual user demands $\mathbf{d}$ in the delivery phase, placement strategies may still depend on the identity and number of active users during the delivery phase. Such coordination in the placement phase is known as centralized placement, which has been described in Section 1.2.1. Since the identity, or even the number, of active users may not be known several hours before the delivery phase takes place, it is also important to consider strategies in which placement is not allowed to depend on such information. In particular, this is even more crucial in wireless networks where users enjoy an high-degree of mobility. This lack of coordination is known as decentralized placement [79], also explained in Section 1.2.1. Decentralization during the placement phase can be realized by allowing randomized placement schemes. For instance, each user i independently draws a caching function $\phi_i(\mathcal{W}; D)$ from an ensemble of randomized caching functions parameterized by an arbitrary random variable D, independent of i and K.

4.4. Main Results and Insights

The main results of this chapter are: 1) the sum-GDoF characterization of the symmetric cache-aided MISO BC under partial CSIT, described in Section 4.3, to within a constant multiplicative gap, and 2) showing that such sum-GDoF characterization is robust to decentralization. We start by presenting the first result and deriving useful insights assuming a centralized setting, then we extend to the decentralized setting.

4.4.1. Centralized placement

In order to state the sum-GDoF result, we define the centralized GNDT function $\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta)$, where

$$\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta) \triangleq \frac{K(1-\mu)}{K(1-(\alpha-\beta)) + (1+K\mu)(\alpha-\beta)} \tag{4.10}$$

for any $\alpha \in [0,1]$, $\beta \in [0,\alpha]$ and $\mu \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\}$, and the lower convex envelope of these points for all other $\mu \in [0,1]$.

Theorem 4.1. *For the symmetric K-user cache-aided MISO BC under partial CSIT described in Section 4.3, under centralized placement we achieve the sum-GDoF given by*

$$\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta) = \frac{K}{\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta)}. \tag{4.11}$$

Moreover, the achievable sum-GDoF in (4.11) *satisfies*

$$\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta) \leq \mathsf{GDoF}(\mu, \alpha, \beta) \leq 12 \cdot \mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta). \tag{4.12}$$

The proof of Theorem 4.1 is presented in Section 4.6. As in [25, 34], the somewhat loose multiplicative gap of 12 in Theorem 4.1 is due to the analytical bounding techniques used in the converse. Numerical simulations suggest that such factor is no more than 3.5 for $K \leq 100$ and $N \leq 500$.

To gain some insights into the sum-GDoF characterized in Theorem 4.1, we restrict the following discussion to $\mu \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}\}$, for which the achievable sum-GDoF in (4.11) is expressed as

$$\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta) = (1 - (\alpha - \beta))\frac{K}{1-\mu} + (\alpha - \beta)\frac{1 + K\mu}{1 - \mu}. \tag{4.13}$$

It is easily seen that $\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta)$ in (4.13) reduces to its classical counterpart in [43] under $\mu = 0$, i.e. where no caches are available. In this case, the multiplicative factor of 12 can be reduced to 1. However, more significantly, the form taken by the sum-GDoF in (4.13), for any μ (in the set above), is analogous to the form of the classical sum-GDoF in [43]. This is explained in more details next, where we use the terminology of signal power levels measured in terms of the exponent of P [130]. We start by looking at specialized cases from which we build our way towards the general case.

DoF Under Partial CSIT

Recall that sum-DoF characterization under partial CSIT is obtained by setting $\alpha = 1$. Defining $\mathsf{DoF}_{\mathrm{C}}(\mu, \beta) \triangleq \mathsf{GDoF}_{\mathrm{C}}(\mu, 1, \beta)$ and applying such specialization to (4.13), we obtain

$$\mathsf{DoF}_{\mathrm{C}}(\mu, \beta) = \beta\frac{K}{1-\mu} + (1 - \beta)\frac{1 + K\mu}{1 - \mu}. \tag{4.14}$$

Under perfect CSIT ($\beta = 1$), zero-forcing over the physical channel enables a spatial multiplexing gain of K. By incorporating caches into the picture, we obtain a further local caching gain of $\frac{1}{1-\mu}$, which is the only relevant caching gain here as zero-forcing creates parallel (non-interfering) single-user links. Under the other extreme, i.e. finite precision CSIT ($\beta = 0$), all spatial multiplexing gains in the physical channel are lost and the sum-DoF collapses to the one obtained in the original setting with a shared link [24]. In this case, the network relies on the local caching gain of $\frac{1}{1-\mu}$ and the global caching gain of $1 + K\mu$, where the latter is enabled by creating coded-multicasting opportunities.

It is readily seen that finite precision CSIT is as (un)useful as no CSIT from a DoF perspective[3]. This is reminiscent of the sum-DoF collapse in the classical MISO BC seen in Chapter 2, proved for the first time in [22]. Moreover, it is worth noting that since the sum-DoF of the cache-aided MISO BC is an upper-bound for the sum-DoF of cache-aided interference networks, this collapse under finite precision CSIT also holds for the networks in [31, 34, 106].

For partial CSIT ($0 < \beta < 1$), the sum-DoF takes the form $\beta\mathsf{DoF}_{\mathrm{C}}(\mu, 1) + (1 - \beta)\mathsf{DoF}_{\mathrm{C}}(\mu, 0)$, laying on the line connecting the two extremes. In this case, partial CSIT of level β allows (power-

[3] It is implicitly understood that such statements hold in an order-optimal sense. This applies to all similar observations herein.

controlled) zero-forcing transmission in the bottom β signal power levels without leaking any interference above the noise floor at undesired users, as extensively seen in Chapter 2 and 3 . This utilization of only a fraction of power levels yields the factor β in the DoF. The remaining signal power levels are used for a shared-link-type transmission requiring no CSIT. In particular, this transmission sees interference from the zero-forcing layer, hence is left with the top $(1-\beta)$ power levels as reflected in the DoF. Since all users can decode (and remove) all codewords in the shared-link layer without influencing its achievable DoF, the zero-forcing layer remains unaffected. This is again in agreement with what already extensively seen in Chapter 2 and 3. To facilitate such partitioned transmission, messages (or files in this case) are split into private and common parts delivered through the zero-forcing and shared link layers, respectively, in the way will be described in detail later.

The scheme described above expands upon, and inherits the main features of, the rate-splitting scheme used for the classical MISO BC with partial CSIT (alongside other networks) described in Section 2.5. Hence, it is not surprising to see that the cache-aided sum-DoF takes the same weighted-sum form of the classical sum-DoF $\beta K + (1-\beta)$, recovered from the above by setting $\mu = 0$.

GDoF Under Finite Precision CSIT

This is recovered from (4.13) by setting $\beta = 0$ and corresponds to the achievable sum-GDoF in [36]. It is easily checked that the sum-GDoF in this case takes the form of the DoF in (4.13), after replacing β with $1-\alpha$. This is inline with the observation that DoF results under partial CSIT translate to GDoF results under finite precision CSIT [131]. This also highlights that unlike the DoF metric, the GDoF metric captures spatial multiplexing gains under finite precision (or even absent) CSIT. Such multiplexing gains, however, are achieved by exploiting the signal power levels only.

The General Case

For arbitrary levels of β and α, the insights derived in [43] for the GDoF of the classical MISO BC extend to the cache-aided counterpart. In particular, the cross-link strength level α and the CSIT quality level β equally counter each other and hence only their difference $(\alpha - \beta)$ matters. The bottom $1 - (\alpha - \beta)$ power levels are reserved for parallel-link-type transmission through zero-forcing and power control, while the shared-link-type transmission rises above, essentially occupying the top $(\alpha - \beta)$ power levels. Therefore, it is readily seen that as $(\alpha - \beta)$ increases, the network starts relying more on the global caching gain and less on spatial multiplexing gains as reflected in (4.13).

4.4.2. Decentralized placement

In this part we consider the decentralized setting where centrally coordinated placement is not allowed during the placement phase. Before we state the following result, we define the decentralized

GNDT function $\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)$, where

$$\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta) \triangleq K \sum_{m=0}^{K-1} \frac{\binom{K-1}{m}\mu^{m}\,(1-\mu)^{K-m}}{K(1-(\alpha-\beta))+(1+m)(\alpha-\beta)} \tag{4.15}$$

for any $\alpha \in [0,1]$, $\beta \in [0,\alpha]$ and $\mu \in [0,1]$.

Theorem 4.2. *For the symmetric K-user cache-aided MISO BC under partial CSIT described in Section 4.3, under decentralized placement we achieve the sum-GDoF given by*

$$\mathsf{GDoF}_{\mathrm{D}}(\mu,\alpha,\beta) = \frac{K}{\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)}. \tag{4.16}$$

Moreover, the achievable sum-GDoF in (4.16) *satisfies*

$$\mathsf{GDoF}_{\mathrm{D}}(\mu,\alpha,\beta) \leq \mathsf{GDoF}(\mu,\alpha,\beta) \leq 12 \cdot \mathsf{GDoF}_{\mathrm{D}}(\mu,\alpha,\beta). \tag{4.17}$$

The proof of Theorem 4.2 is presented in Section 4.7. The most significant consequence of Theorem 4.2 is that centralized placement leads to at most a constant-factor improvement of the GDoF over decentralized placement. Through a straightforward inspection, this constant-factor improvement is bounded above by $\mathsf{GDoF}_{\mathrm{C}}(\mu,\alpha,\beta) \leq 12 \cdot \mathsf{GDoF}_{\mathrm{D}}(\mu,\alpha,\beta)$, obtained from (4.12) and (4.17). In Section 4.7.3, this multiplicative gap between the centralized GDoF and decentralized GDoF is tightened to 1.5.

In Section 4.7.2, we show that an upper-bound on $\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)$ takes the form of the centralized delivery time in (4.10), yet with a lower coded-multicasting gain. It follows that the insights that follow Theorem 4.1, derived in the light of the centralized achievable sum-GDoF, extend to the decentralized setting.

4.5. Upper-Bound

In this section, we obtain an upper-bound for the sum-GDoF. Since it is more convenient to work with the GNDT in (4.9), the upper-bound is derived in terms of a lower-bound on $\mathsf{GNDT}(\mu,\alpha,\beta)$.

Theorem 4.3. *For the symmetric cache-aided MISO BC under partial CSIT described in Section 4.3, a lower-bound on the optimal GNDT is given by*

$$\mathsf{GNDT}(\mu,\alpha,\beta) \geq \max_{s\in\{1,2,\ldots,K\}} \mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\alpha,\beta), \tag{4.18}$$

where $\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\alpha,\beta)$ *is defined as*[4]

$$\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\alpha,\beta) \triangleq \left(\frac{s}{1+(s-1)(1-(\alpha-\beta))}\left(1-\frac{M}{\lfloor\frac{N}{s}\rfloor}\right)\right)^{+}. \tag{4.19}$$

[4] For any $x \in \mathbb{R}$, we define $(x)^{+} \triangleq \max\{0,x\}$.

In the above, for any subset of $s \leq K$ users, the corresponding $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)$ in (4.19) is a lower-bound on the optimal delivery time $\mathsf{GNDT}(\mu, \alpha, \beta)$. It follows that the tightest of such lower-bounds is obtained by maximizing $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)$ over s. We also observe that $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)$ depends on the parameters of the physical channel through the difference $(\alpha - \beta)$. In particular, for a fixed number of users s, library size N and cache size M, $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)$ decreases when $(\alpha - \beta)$ decreases. This is intuitively explained by the fact that decreasing $(\alpha - \beta)$ corresponds to higher (relative) CSIT quality, enabling larger spatial multiplexing gains which in turn reduce the delivery time.

From Theorem 4.3 and (4.9), it is easily seen that an upper-bound for the sum-GDoF is given by

$$\mathsf{GDoF}(\mu, \alpha, \beta) \leq \min_{s \in \{1,2,\ldots,K\}} \frac{K}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)}. \tag{4.20}$$

The upper-bound in Theorem 4.3 is employed to prove the converse parts of Theorem 4.1 and Theorem 4.2 in the following sections. In the remainder of this section, we present a proof for Theorem 4.3. The proof relies on two main ingredients summarized as follows.

1. A lower-bound on $\mathsf{GNDT}(\mu, \alpha, \beta)$ is obtained by considering a subset of $s \leq K$ users and a multi-demand communication, in which each user requests multiple distinct files.

2. Each cache memory is replaced with a parallel side link of capacity that can convey the information content of the cache to the user by the end of the multi-demand communication. By bounding the sum-GDoF of this new channel, we bound the delivery time of the multi-demand communication.

Similarities and differences between this proof and previous works are discussed at the end of this section.

4.5.1. Multi-Demand Communication

Consider a subset of $s \leq K$ users and a multi-demand communication over the cache-aided channel, in which each user requests a set of $\lfloor \frac{N}{s} \rfloor$ distinct files and no file is requested by two different users. We denote the $\lfloor \frac{N}{s} \rfloor$ files requested by user i as $W_{d_i^1}, \ldots, W_{d_i^{\lfloor N/s \rfloor}}$. By the end of the communication, each user is able to recover the $\lfloor \frac{N}{s} \rfloor$ requested files from the received signals and the local cache content. The optimal delivery time for this multi-demand communication is denoted by $\mathsf{GNDT}_{\mathrm{md}}$, which is also defined in the worst-case sense, i.e. for the worst-case amongst all possible multi-demands of $\lfloor \frac{N}{s} \rfloor$ files. It is readily seen that $\mathsf{GNDT}_{\mathrm{md}}$ satisfies

$$\mathsf{GNDT}_{\mathrm{md}} \leq \left\lfloor \frac{N}{s} \right\rfloor \mathsf{GNDT}(\mu, \alpha, \beta) \tag{4.21}$$

since we are ignoring $K - s$ users and it is always feasible to treat each demand of s files separately in a consecutive manner. Next, we transfer to an equivalent setup with no caches.

4.5.2. Cache Replacement and Delivery Time Lower-Bound

Now consider a new MISO BC consisting of the same K transmitters, with access to the same library of N files, and the $s \leq K$ users served in the multi-demand communication above. However, users in this new channel are not equipped with caches. Alternatively, communication is carried out over two parallel sub-channels. The input-output relationship is given by

$$Y_i(t) = \sqrt{P} G_{ii}(t) X_i(t) + \sum_{j=1, j \neq i}^{K} \sqrt{P^{\alpha}} G_{ij}(t) X_j(t) + Z_i(t) \tag{4.22}$$

$$B_i(t) = \sqrt{P^{\gamma}} A_i(t) + C_i(t) \tag{4.23}$$

where (4.22) and (4.23) describe the first and second sub-channels, respectively. All physical properties of (4.4), described in Section 4.3.1, are inherited by the first sub-channel in (4.22). For the second sub-channel, $A_i(t) \in \mathbb{C}$ is the signal transmitted to the i-th user with a power constraint $\mathbb{E}\left(|A_i(t)|^2\right) \leq 1$, $B_i(t) \in \mathbb{C}$ is the signal received by the i-th user and $C_i(t) \sim \mathcal{N}_{\mathbb{C}}(0,1)$ is the i.i.d. AWGN. Each link in the second sub-channel remains constant over t and has channel strength level $\gamma \geq 0$, hence supports a transmission at rate $\gamma \log(P) + o\big(\log(P)\big)$ without influencing the rate over the first sub-channel. Equivalently, γ is the GDoF (or capacity in files per time-slot) of each individual link in the second sub-channel.

In this new MISO BC with parallel sub-channels, each user i requests the same $\left\lfloor \frac{N}{s} \right\rfloor$ files requested by the corresponding user in the multi-demand communication, i.e. $W_{d_i^1}, \ldots, W_{d_i^{\lfloor N/s \rfloor}}$. Each transmitter j then generates the codewords X_j^n and A_j^n, sent over $n \in \mathbb{N}$ channel uses through the sub-channels in (4.22) and (4.23) respectively. By the end of the communication, user i recovers the $\lfloor N/s \rfloor$ requested files from the signals Y_i^n and B_i^n, received through the sub-channels in (4.22) and (4.23) respectively. The optimal sum-GDoF of this new MISO BC, denoted by $\mathsf{GDoF}_{\mathrm{P}}(\alpha, \beta, \gamma)$, is bounded above as follows.

Lemma 4.1. *For the s-user MISO BC, consisting of two parallel sub-channels, described in* (4.22) *and* (4.23)*, the optimal sum-GDoF is bounded above as*

$$\mathsf{GDoF}_{\mathrm{P}}(\alpha, \beta, \gamma) \leq (\alpha - \beta) + s\big(1 - (\alpha - \beta)\big) + s\gamma. \tag{4.24}$$

It is evident that the bound on $\mathsf{GDoF}_{\mathrm{P}}(\alpha, \beta, \gamma)$ in (4.24) depends on α and β through their difference $(\alpha - \beta)$. For the extreme case of $(\alpha - \beta) = 0$, the parallel MISO BC enjoys full spatial multiplexing gains over the first sub-channel. On the other hand, for the other extreme of $(\alpha - \beta) = 1$, all spatial multiplexing gains are annihilated and the sum-GDoF of the first sub-channel collapses to 1. Note that the contribution from the second sub-channel is unaffected since it consists of non-interfering links. The proof of Lemma 4.1 is relegated to Appendix B.1. Next, we argue that by setting γ such that

$$\gamma \cdot \mathsf{GNDT}_{\mathrm{md}} = M \tag{4.25}$$

the corresponding optimal delivery time of the new channel is a lower-bound on the optimal total delivery time of the cache-aided multi-demand communication, i.e.

$$\frac{s\left\lfloor\frac{N}{s}\right\rfloor}{\mathsf{GDoF}_{\mathrm{P}}(\alpha,\beta,\gamma)} \leq \mathsf{GNDT}_{\mathrm{md}}. \tag{4.26}$$

This follows by observing that (4.25) guarantees that for each user i, the content of the cache $\mathcal{U}_i$ in the original channel can be delivered over the second sub-channel in (4.23) using at most $\mathsf{GNDT}_{\mathrm{md}}$ time-slots. Since this does not influence the GDoF achieved over the first sub-channel in (4.22), any placement and delivery strategy implemented for the cache-aided multi-demand communication is feasible in the new channel and will take at most $\mathsf{GNDT}_{\mathrm{md}}$ time-slots. We proceed while assuming that (4.25) holds.

By combining (4.26) with Lemma 4.1 and (4.25), followed by invoking (4.21), we obtain

$$\left\lfloor\frac{N}{s}\right\rfloor s \leq \mathsf{GNDT}_{\mathrm{md}}\big(1+(s-1)(1-(\alpha-\beta))+s\gamma\big) \tag{4.27}$$

$$= \mathsf{GNDT}_{\mathrm{md}}\big(1+(s-1)(1-(\alpha-\beta))\big)+sM \tag{4.28}$$

$$\leq \mathsf{GNDT}(\mu,\alpha,\beta)\left\lfloor\frac{N}{s}\right\rfloor\big(1+(s-1)(1-(\alpha-\beta))\big)+sM. \tag{4.29}$$

After some rearrangement and by considering that the delivery time is non-negative, we obtain

$$\mathsf{GNDT}(\mu,\alpha,\beta) \geq \left(\frac{s}{1+(s-1)(1-(\alpha-\beta))}\left(1-\frac{M}{\left\lfloor\frac{N}{s}\right\rfloor}\right)\right)^{+}. \tag{4.30}$$

The lower-bound in (4.30) is further tightened by maximizing over all possible sizes of user subsets, i.e. $s \in [K]$, from which the result in (4.18) directly follows.

4.5.3. Insights and Relation to Prior Works

The multi-demand communication to a subset of users corresponds to the cut-set-based bound in [24], while the cache replacement is inspired by [32]. However, it is worthwhile highlighting that bounding the DoF under partial current and perfect delayed CSIT and side links (after cache replacement) in [32] is very different from bounding the sum-GDoF under only partial current CSIT and side links in Lemma 4.1. In particular, the DoF upper-bound in [32] follows the footsteps of [124], and is essentially based on a genie-aided argument. Such argument does not work for the DoF/GDoF with only partial current CSIT and is known to give a loose bound in general. The proof of Lemma 4.1 is hence based on the outer-bounds in [22, 41, 43], which rely on the aligned image sets approach under channel uncertainty, already mentioned in Chapters 2 and 3.

It is also worthwhile highlighting that the sum-GDoF upper-bound in Lemma 4.1 is achievable through separate coding over the two sub-channels, i.e. there are no synergistic gains to be exploited through joint coding. This comes in contrast to the setting in [32], where jointly coding over the parallel sub-channels (after cache replacement) can strictly outperform separate coding. The influence of this synergy (or the lack of it) is clear when we revert back to the cache-aided channels. In particular, we saw in Theorem 4.1 that the considered cache-aided MISO BC collapses to the shared-link setting in [24] when $(\alpha - \beta) = 1$. However, even when current CSIT is completely absent in [32], the synergy between caches and delayed CSIT leads to an improved performance compared to the shared-link setting.

4.6. Centralized Placement

In this section, we treat the centralized setting and prove Theorem 4.1. We start with the achievability and then we prove order-optimality using the upper-bound in Theorem 4.3.

4.6.1. Achievability scheme

Here we present a centralized scheme which achieves the delivery time given by $\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta)$ in (4.10), and hence the sum-GDoF given by $\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta)$ in Theorem 4.1. This scheme builds upon and generalizes the one proposed for the cache-aided MISO BC in [36]. The key difference is that the scheme in [36] is tuned to a special case in which only finite precision CSIT (i.e. $\beta = 0$) is available, while the one proposed here bridges the gap by considering all relevant levels of partial CSIT, i.e. $\beta \in [0, \alpha]$.

A key ingredient of the achievability scheme is the transmission of common and private codewords during the delivery phase. We start by treating this physical-layer aspect through the following result.

Lemma 4.2. *Consider the K-user MISO BC with signal model given by* (4.4) *and properties described in Section 4.3.1. Further assume that the transmitter has a common message $W^{(\mathrm{c})}$, intended to all user, and private messages $W_1^{(\mathrm{p})}, \dots, W_K^{(\mathrm{p})}$, where $W_i^{(\mathrm{p})}$ is intended only to user i. We achieve the GDoF*

$$\mathsf{GDoF}^{(\mathrm{c})} = \alpha - \beta \tag{4.31}$$

$$\mathsf{GDoF}_i^{(\mathrm{p})} = 1 - (\alpha - \beta), \ \forall i \in [K] \tag{4.32}$$

where $\mathsf{GDoF}^{(\mathrm{c})}$ is the GDoF achieved by the common message and $\mathsf{GDoF}_i^{(\mathrm{p})}$ is the GDoF achieved by the i-th private message.

The GDoF in (4.31) and (4.32) is achieved using rate-splitting applied to the GDoF framework. Using the terminology of signal power levels to explain the power-level partitioning, the upper $(\alpha - \beta)$ power levels are occupied by the common message while the bottom $1 - (\alpha - \beta)$ power levels are reserved for the private messages. Note that the transmission of the common message requires no CSIT, while the transmission of the private messages is carried out using zero-forcing and power control, and hence may rely on the available partial CSIT. Therefore, in the extreme case of $(\alpha - \beta) = 1$ (i.e. finite precision CSIT and equal strength paths), spatial multiplexing gains achieved through zero-forcing and power control collapse and the corresponding private messages will have a GDoF of zero. Note that this is inline with the DoF analysis in Section 2.5. The full proof of Lemma 4.2 is relegated to Appendix B.2.

In the following, we focus on $\mu \in \{\frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}\}$, such that $K\mu$ is an integer. For $\mu = 0$, no caching is possible and the GDoF-optimal transmission strategy is given in [43], which is an extension of the DoF-optimal strategy in Section 2.5 to the DoF framework. For the other extreme of $\mu = 1$, we have $\mathsf{GNDT}_{\mathrm{C}}(1, \alpha, \beta) = 0$ as each user can store the entire library. For the remaining μ, where $K\mu$ is not necessarily an integer, $\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta)$ is obtained by memory-sharing over the schemes corresponding to $\mu \in \big\{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\big\}$, as pointed out in [24].

Placement phase

The placement is analogous to [24] and does not depend on the parameters specific to the considered channel, e.g. transmitting antennas, α and β. We use $m_{\mathrm{C}} \triangleq \mu K$ for notational briefness and to facilitate reusing some parts in the following section for the decentralized case. Let $\Omega = \{\mathcal{T} \subseteq [K] : |\mathcal{T}| = m_{\mathrm{C}}\}$ be the family of all subsets of users with cardinality m_{C}. Each file $W_l \in \mathcal{W}$ is split into $\binom{K}{m_{\mathrm{C}}}$ non overlapping, equal size, subfiles labeled as $W_{l,\mathcal{T}}$, for all $\mathcal{T} \in \Omega$, where each subfile consists of $F/\binom{K}{m_{\mathrm{C}}}$ bits. User i caches all the subfiles $W_{l,\mathcal{T}}$ such that $i \in \mathcal{T}$ and $l \in [N]$. Hence, the corresponding cache memory is filled as $\mathcal{U}_i = \{W_{l,\mathcal{T}} : \ \mathcal{T} \in \Omega, \ i \in \mathcal{T}, \ l \in [N]\}$. Each user stores $N\binom{K-1}{m_{\mathrm{C}}-1}$ subfiles which corresponds to a total of MF bits, hence satisfying the memory constraint.

Delivery phase

During the delivery phase, the tuple $\mathbf{d}$ of all user demands is revealed, where each user i makes a request for file W_{d_i}. Since user i has all subfiles $W_{d_i,\mathcal{T}}$ such that $i \in \mathcal{T}$, the transmitter has to deliver all subfiles $W_{d_i,\mathcal{T}}$ such that $i \notin \mathcal{T}$, for all users $i \in [K]$. This corresponds to a total of $K(1-\mu)F$ bits to be delivered over the wireless channel.

The transmitter splits each subfile $W_{d_i,\mathcal{T}}$, with $i \notin \mathcal{T}$, into a common mini-subfile $W^{(\mathrm{c})}_{d_i,\mathcal{T}}$ and a private mini-subfile $W^{(\mathrm{p})}_{d_i,\mathcal{T}}$ such that $W_{d_i,\mathcal{T}} = \big(W^{(\mathrm{c})}_{d_i,\mathcal{T}}, W^{(\mathrm{p})}_{d_i,\mathcal{T}}\big)$. The two mini-subfiles $W^{(\mathrm{c})}_{d_i,\mathcal{T}}$ and $W^{(\mathrm{p})}_{d_i,\mathcal{T}}$ have sizes $q|W_{d_i,\mathcal{T}}|$ bits and $(1-q)|W_{d_i,\mathcal{T}}|$ bits respectively, where $|W_{d_i,\mathcal{T}}|$ is the size of file $W_{d_i,\mathcal{T}}$ and q is the file splitting ratio given by

$$q = \frac{(1+m_{\mathrm{C}})(\alpha-\beta)}{K(1-(\alpha-\beta)) + (1+m_{\mathrm{C}})(\alpha-\beta)}. \tag{4.33}$$

All common mini-subfiles are coded using the techniques in the original coded-multicasting scheme in [24]. In particular, subsets of $1+m_{\mathrm{C}}$ common mini-subfiles $W^{(\mathrm{c})}_{d_i,\mathcal{T}}$ are combined together using a bitwise XOR operation to generate multicasting messages intended for subsets of $1+m_{\mathrm{C}}$ users as follows

$$W^{(\mathrm{c})}_{\mathcal{S}} = \oplus_{i \in \mathcal{S}} W^{(\mathrm{c})}_{d_i,\mathcal{S}\setminus\{i\}} \tag{4.34}$$

for all $\mathcal{S} \in \Theta$, where $\Theta = \{\mathcal{S} \subseteq [K] : |\mathcal{S}| = 1+m_{\mathrm{C}}\}$. All multicasting messages $W^{(\mathrm{c})}_{\mathcal{S}}$ are encoded into a common codeword $X^{(\mathrm{c})}$, while all private mini-subfiles $W^{(\mathrm{p})}_{d_i,\mathcal{T}}$ intended to user i

are encoded into the private codeword $X_i^{(\mathrm{p})}$. Next, the transmission of the common and private codewords over the wireless channel is carried out as described in Appendix B.2.

By decoding $X^{(\mathrm{c})}$, each user i retrieves the multicasting messages $W_{\mathcal{S}}^{(\mathrm{c})}$ for all $\mathcal{S} \in \Theta$. Hence, user i recovers all missing common mini-subfiles by combining with the content of its local cache as in [24]. For example, for some $\mathcal{T}$ such that $i \notin \mathcal{T}$, user i solves for the missing $W_{d_i,\mathcal{T}}^{(\mathrm{c})}$ using XOR combining of $W_{\mathcal{S}}^{(\mathrm{c})}$, where $\mathcal{S} = \mathcal{T} \cup \{i\}$, with the pre-stored m_{C} common mini-subfiles $W_{d_k,\mathcal{S}\setminus\{k\}}^{(\mathrm{c})}$ with $k \in \mathcal{T}$. After decoding $X^{(\mathrm{c})}$, and removing its contribution from the received signal as explained in Appendix B.2, user i decodes the private codeword $X_i^{(\mathrm{p})}$, from which the missing private mini-subfiles $W_{d_i,\mathcal{T}}^{(\mathrm{p})}$, with $\mathcal{T}$ such that $i \notin \mathcal{T}$, are retrieved. At this stage, the entire requested file W_{d_i} is recovered.

Achievable Delivery Time

The shared-link-type transmission, taking place over $X^{(\mathrm{c})}$, delivers a total of $qK(1-\mu)$ files (by excluding the parts already cached) at rate $(\alpha-\beta)(1+m_{\mathrm{C}})$ files per time slot, where $(\alpha-\beta)$ is the GDoF of the physical channel as seen from Lemma 4.2 and $(1+m_{\mathrm{C}})$ is the gain due to coded-multicasting. Hence, the delivery time for the shared-link layer is

$$\frac{Kq(1-\mu)}{(\alpha-\beta)(1+m_{\mathrm{C}})} = \frac{K\,(1-\mu)}{K\,(1-(\alpha-\beta))+(1+m_{\mathrm{C}})\,(\alpha-\beta)}. \tag{4.35}$$

On the other hand, each $X_i^{(\mathrm{p})}$ in the zero-forcing layer delivers a total of $(1-q)(1-\mu)$ files at rate $1-(\alpha-\beta)$ files per time slot, as seen from Lemma 4.2. Hence, the delivery time for this layer is

$$\frac{K(1-q)(1-\mu)}{K\big(1-(\alpha-\beta)\big)} = \frac{K\,(1-\mu)}{K\,(1-(\alpha-\beta))+(1+m_{\mathrm{C}})\,(\alpha-\beta)}. \tag{4.36}$$

Since the two layers take place in parallel, the total delivery time is also given by

$$\mathsf{GNDT}_{\mathrm{C}}(\mu,\alpha,\beta) = \frac{K\,(1-\mu)}{K\,(1-(\alpha-\beta))+(1+m_{\mathrm{C}})\,(\alpha-\beta)}. \tag{4.37}$$

As $\mathsf{GNDT}_{\mathrm{C}}(\mu,\alpha,\beta)$ is achievable, then the corresponding sum-GDoF given by $\mathsf{GDoF}_{\mathrm{C}}(\mu,\alpha,\beta)$ is achievable.

4.6.2. Converse

Here we prove the converse in (4.12), which is equivalent to showing order-optimality of $\mathsf{GNDT}_{\mathrm{C}}(\mu,\alpha,\beta)$, i.e. $\mathsf{GNDT}_{\mathrm{C}}(\mu,\alpha,\beta)/\mathsf{GNDT}(\mu,\alpha,\beta) \leq 12$. Since $\mathsf{GNDT}_{\mathrm{C}}(\mu,\alpha,\beta)$ and $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\alpha,\beta)$ only depend on the difference $(\alpha-\beta)$, with a slight abuse of notation we define

$$\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta) \triangleq \frac{K\,(1-\mu)}{K\,(1-\delta)+(1+K\mu)\,\delta} \tag{4.38}$$

and

$$\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta) \triangleq \left(\frac{s}{1+(s-1)(1-\delta)}\left(1-\frac{M}{\lfloor\frac{N}{s}\rfloor}\right)\right)^{+}. \tag{4.39}$$

where $\delta \in [0,1]$, $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta=\alpha-\beta)=\mathsf{GNDT}_{\mathrm{C}}(\mu,\alpha,\beta)$ and $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta=\alpha-\beta)=\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\alpha,\beta)$. Note $\mu \in \{0,\frac{1}{K},\frac{2}{K},\ldots,\frac{K-1}{K},1\}$ is assumed in (4.38), where the lower convex envelope is taken for the remaining points in $\mu \in [0,1]$. From the above, the lower-bound in (4.18) is rewritten as

$$\mathsf{GNDT}(\mu,\alpha,\beta) \geq \max_{s\in\{1,2,\ldots,K\}} \mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta=\alpha-\beta) \tag{4.40}$$

In the remaining part, we work with $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)$ and $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta)$ for convenience. We show in Appendix B.3.1 that for any μ, there exists a particular $s \in [K]$ such that $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)/\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta) \leq 12$ for all $\delta \in [0,1]$. Since the right-hand-side of (4.40) is bounded below by $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta)$ for any $s \in [K]$, the order-optimality within a factor of 12 follows. This concludes the proof of the converse.

4.7. Decentralized Placement

In this section, we prove Theorem 4.2 which considers the decentralized setting. As in Section 4.6, we start with the achievability and then proceed to prove order-optimality.

4.7.1. Achievability Scheme

Here we propose a decentralized scheme which achieves the delivery time given by $\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)$ in (4.15), and hence the sum-GDoF given by $\mathsf{GDoF}_{\mathrm{D}}(\mu,\alpha,\beta)$ in Theorem 4.2. We start with the placement phase.

Placement Phase

This is similar to the procedure in the original decentralized coded-caching paper [79], and hence does not depend on the wireless channel parameters. Each user i stores a subset of μF bits from each file, chosen uniformly at random. Therefore, each bit of each file is stored in some subset of users[5] $\tilde{\mathcal{T}} \in 2^{[K]}$, where $|\tilde{\mathcal{T}}| \in \{0,1,\ldots,K\}$. For some $l \in [N]$, we use $W_{l,\tilde{\mathcal{T}}}$ to denote the bits of file W_l which are stored by all users in $\tilde{\mathcal{T}}$, where each $W_{l,\tilde{\mathcal{T}}}$ is referred to as a subfile henceforth. It is readily seen that W_l can be reconstructed from $\left\{W_{l,\tilde{\mathcal{T}}} : \tilde{\mathcal{T}} \in 2^{[K]}\right\}$.

Delivery Phase

User i requires all subfiles $W_{d_i,\tilde{\mathcal{T}}}$, such that $i \notin \tilde{\mathcal{T}}$, in order to recover the requested file W_{d_i}. The delivery phase takes place over K sub-phases indexed by $m \in \{0,1,\ldots,K-1\}$. In the m-th sub-phase, the transmitter delivers all subfiles $W_{d_i,\tilde{\mathcal{T}}}$, such that $i \in [K]$ and $i \notin \tilde{\mathcal{T}}$, with $|\tilde{\mathcal{T}}| = m$.

[5] For a set $\mathcal{S}$, the power set $2^{\mathcal{S}}$ consists of all subsets of $\mathcal{S}$ (including $\mathcal{S}$ itself) and the empty set $\emptyset$. Note that we consider finite $[K]$, i.e. K does not go to infinity. This guarantees that the power set is not an uncountable set.

Note that m goes up to $K-1$ since for $|\tilde{\mathcal{T}}| = K$, the corresponding subfiles are pre-stored by all users.

Focusing on the m-th delivery sub-phase, delivery is carried out as described in Section 4.6.1 for the centralized setting, while replacing m_{C} in Section 4.6.1 by m. This is due to the fact that each subfile to be delivered during the m-th decentralized delivery sub-phase is pre-stored by m users instead of m_{C} users in centralized delivery. It follows that coded-multicasting messages have order $1+m$ in the m-th decentralized delivery sub-phase compared to $1+m_{\mathrm{C}}$ in centralized delivery, which is due to the random decentralized placement. Note that when performing the XOR operation in (4.34) for the decentralized setting, all subfiles are assumed to be zero-padded to the length of the longest subfile [79]. By the end of the K delivery sub-phases, the entire requested files are recovered by the users.

Note that in sub-phase $m = 0$, there are no coded-multicasting opportunities as this sub-phase delivers parts which are not pre-stored by any user. Hence, the transmission here is similar to the centralized setting with $\mu = 0$, which corresponds to transmission in the classical MISO BC with no caches [43].

Achievable Delivery Time

Consider the m-th sub-phase and an arbitrary subset of users $\tilde{\mathcal{T}}$ with size m. For each file W_l, $l \in [N]$, the probability of any of its bits to be stored in the cache of some user in $\tilde{\mathcal{T}}$ is given by μ. Hence, the probability of this bit to be stored by exactly the m users of $\tilde{\mathcal{T}}$ is given by $\mu^m(1-\mu)^{K-m}$, from which the expected number of bits stored by each of such users is given by $\mu^m(1-\mu)^{K-m}F$. It follows that, as $F \to \infty$, the expected size of $W_{l,\tilde{\mathcal{T}}}$ is given by

$$\mu^m(1-\mu)^{K-m}F + o(F) \tag{4.41}$$

where the term $o(F)$ is omitted in the following calculations. Since there is a total of $\binom{K}{m}$ subsets of m users, we have $\binom{K}{m}\mu^m(1-\mu)^{K-m}F$ bits of each file which are cached by exactly m users.

Now we proceed to calculated the number of bits of the file W_{d_i}, which are stored by exactly m users, which have to be delivered to user i. Recall that user i already has all subfiles $W_{d_i,\tilde{\mathcal{T}}}$, with $|\tilde{\mathcal{T}}| = m$ and $i \in \tilde{\mathcal{T}}$, pre-stored. Hence, user i already has $\binom{K-1}{m-1}\mu^m(1-\mu)^{K-m}F$ bits of W_{d_i} which are cached in exactly m users. Hence, the number of unavailable bits, contained in all subfiles $W_{d_i,\tilde{\mathcal{T}}}$ with $|\tilde{\mathcal{T}}| = m$ and $i \notin \tilde{\mathcal{T}}$, is given by $\binom{K-1}{m}\mu^m(1-\mu)^{K-m}F$. Since there are K users in total, the total number of files (obtained after normalizing by F) which have to be delivered during the m-th sub-phase is given by

$$K\binom{K-1}{m}\mu^m(1-\mu)^{K-m}. \tag{4.42}$$

A portion $q(m) = \frac{(1+m)(\alpha-\beta)}{(1+m)(\alpha-\beta)+K(1-(\alpha-\beta))}$ of such files are delivered with coded-multicasting gain $1+m$ (i.e. simultaneously useful for $1+m$ users) over the common codeword with GDoF $(\alpha-\beta)$ files per time-slot. On the other hand, the remaining portion of $1-q(m)$ is delivered over the private codewords with GDoF $K\,(1-(\alpha-\beta))$ files per time-slot. Hence, the delivery time of the m-th sub-phase is

$$\frac{K\binom{K-1}{m}\mu^m\,(1-\mu)^{K-m}}{K(1-(\alpha-\beta))+(1+m)(\alpha-\beta)}. \tag{4.43}$$

By summing over all K sub-phases, the total delivery time is given by

$$\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta) = K\sum_{m=0}^{K-1}\frac{\binom{K-1}{m}\mu^m\,(1-\mu)^{K-m}}{K(1-(\alpha-\beta))+(1+m)(\alpha-\beta)}. \tag{4.44}$$

It follows that the corresponding GDoF given by $\mathsf{GDoF}_{\mathrm{D}}(\mu,\alpha,\beta)$ is achievable.

4.7.2. Converse

In this part, we prove the converse in (4.17), which is equivalent to showing order-optimality of $\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)$, i.e. $\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)/\mathsf{GNDT}(\mu,\alpha,\beta) \leq 12$. As in the centralized setting, $\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)$ only depends on the difference $\delta=(\alpha-\beta)$. Therefore, we work with

$$\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta) \triangleq K\sum_{m=0}^{K-1}\frac{\binom{K-1}{m}\mu^m\,(1-\mu)^{K-m}}{K(1-\delta)+(1+m)\delta} \tag{4.45}$$

where $\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta=\alpha-\beta)=\mathsf{GNDT}_{\mathrm{D}}(\mu,\alpha,\beta)$. Unlike $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)$ in (4.38), $\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta)$ does not have the desirable form which allows comparing it to the bound in (4.40) directly. Hence, the first (key) step of the converse is to derive an upper-bound on $\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta)$, denoted by $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)$, which takes the form of the centralized achievable delivery time in (4.38). This is given in the following result.

Lemma 4.3. *The decentralized delivery time $\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta)$ is bounded above as*

$$\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta) \leq \mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta) = \frac{K\,(1-\mu)}{K(1-\delta)+(1+u)\delta} \tag{4.46}$$

where u is given by

$$u = \frac{K\,(1-\mu)}{\mathsf{GNDT}_{\mathrm{D}}(\mu,1)} - 1. \tag{4.47}$$

The proof of Lemma 4.3 is given in Appendix B.4. One important consequence of Lemma 4.3 is that the expression in (4.46) allows us to show order-optimality of $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)$ to within a factor of 12 using similar techniques to the ones used for the centralized setting. The details are relegated to Appendix B.3.2. The order-optimality of $\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta)$ to within a factor of 12 follows, which concludes the converse.

4.7.3. Gap Between Decentralized and Centralized Schemes

From a straightforward inspection of (4.38) and (4.46), it can be seen that for integer values of $K\mu$ (for which a close form of $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)$ is obtained), we have

$$\frac{\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)}{\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)} = \frac{\mathsf{GNDT}_{\mathrm{D}}(\mu,\delta)}{\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)} \leq \frac{K(1-\delta)+(K\mu+1)\delta}{K(1-\delta)+(u+1)\delta}. \tag{4.48}$$

We know that when $\delta = 1$ (i.e. $\alpha = 1$ and $\beta = 0$), all spatial multiplexing gains are lost and the achievable delivery times collapse to the ones in [24, 79]. Hence, it follows from the observations in [79] (and then the proof in [76]) that for $\delta = 1$, there is a small price to pay due to decentralization, making the ratio in (4.48) small. By further examining the bound on the right-most side of (4.48), it can be seen that it decreases when δ decreases, hence further reducing the price of decentralization. For example, such price is minimal when $\delta = 0$ (i.e. $\alpha = \beta$), where both the centralized and decentralized strategies achieve the optimal delivery $\mathsf{GNDT}(\mu,\alpha,\beta=\alpha) = 1-\mu$. This is intuitive as with a decreased δ, the system starts to rely more on spatial multiplexing gains and local caching gains and less on global caching gains, which are affected by decentralization. Concretely, the gap in (4.48) is bounded above as follows.

Corollary 4.1. *For any $\delta \in [0,1]$ and $\mu \in [0,1]$, we have*

$$\frac{\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)}{\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)} \leq 1.5. \tag{4.49}$$

The above corollary is obtained by employing the results in Theorem 4.1, Lemma 4.3 and [76]. The full proof is relegated to Appendix B.5.

4.8. Summary of the Chapter

In this chapter, we have study the fundamental limits of cache-aided interference management under full transmitter cooperation. In particular, we characterized the optimal sum-GDoF of the K-user symmetric cache-aided MISO BC under partial CSIT up to a constant multiplicative factor. Moreover, we showed that such sum-GDoF characterization is robust to decentralization, i.e. we proposed a decentralized caching strategy which attains an order-optimal sum-GDoF performance. In order to derive the sum-GDoF results, we introduced the generalized normalized delivery time (GNDT) metric, which extends the normalized delivery time (NDT) metric in the same way the GDoF extends the DoF. The GNDT is related to the reciprocal of the sum-GDoF, and is generally easier to deal with when characterizing achievable and optimal performances.

At the heart of our converse proof is a sum-GDoF upper-bound for a parallel MISO BC with partial CSIT, which extends a family of robust outer-bounds based on the aligned image sets approach, initially developed in the context of classical networks with no caches, to cache-aided networks. On the other hand, we showed that the order optimal sum-GDoF takes a familiar weighted-sum form, often observed in classical networks (with no caches) under partial CSIT. Achieving such sum-GDoF relies on a key interplay between spatial multiplexing and coded-multicasting gains.

5. Centralized and Decentralized Cache-Aided Interference Management in Heterogeneous Parallel Channels

5.1. Overview of the Chapter

In Chapter 4 we have characterized the information-theoretic limits of robust cache-aided interference management under full transmitter cooperation. In particular, we considered a cache-aided MISO BC setting, where multiple transmitters access the entire library and fully cooperate to serve the users. Moreover, we assumed the same number of transmitters and receivers. In this chapter we generalize those results by considering the more general problem of cache-aided interference management in a network which consists of K_{T} single-antenna transmitters and K_{R} single-antenna receivers, where each node is equipped with a cache memory. In particular, this setting generalizes the one in Chapter 4 in two directions: 1) each transmitter does not have access to the entire library but only to the content stored in its cache memory, 2) an arbitrary number of transmitters and receivers is considered.

In this chapter we assume that the transmitters communicate with the receivers over two heterogeneous parallel subchannels: the P-subchannel for which transmitters have perfect instantaneous knowledge of the channel state, and the N-subchannel for which the transmitters have no knowledge of the instantaneous channel state. This is reminiscent of the partial CSIT setting considered in the previous chapters, as the two subchannels can be interpreted as the fractions of the bandwidth where the transmitters have perfect and no CSIT, respectively. As we will see later, this can also be linked to recent results which has shown the equivalence between wireless networks where partial CSIT is reported for all the bandwidth and wireless networks where CSIT is only reported for a fraction of the bandwidth.

In this chapter we focus on one-shot linear delivery strategies [31, 118, 132], where channel symbols cannot be spread over time and frequency. One-shot linear schemes have been widely used as they are practical appealing and allows to tackle otherwise intractable information-theoretic problems. The first result of this chapter, under the assumptions of uncoded placement and separable one-shot linear delivery over the two subchannels, is the characterization of the optimal sum-DoF to within a constant multiplicative factor of 2. Next, and this proves to be technically very challenging, we extend the result to decentralized placement in which no coordination is required for content placement at the receivers. In this case, we characterize the optimal one-shot linear sum-DoF to within a factor of 3.

5.2. Introduction and Main Contributions

In this chapter, we consider a setup comprising a content library of N files and a cache-aided wireless network consisting of K_{T} transmitters and K_{R} receivers, each equipped with a single antenna and a cache memory. The normalized sizes of transmitter and receiver cache memories are given by $\mu_{\mathrm{T}} \in [0, 1]$ and $\mu_{\mathrm{R}} \in [0, 1]$, respectively. As known from the previous chapters, the network operates in two phases: 1) a *placement phase* which takes place before user demands are revealed and in which all the nodes (both transmitters and receivers) store arbitrary parts of the library according to a certain caching strategy, and 2) a *delivery phase* in which users are actively making demands for different files of the library and in which demands are satisfied through a combination of transmissions and the locally stored content from the placement phase. Note that, differently from the cache-aided MISO BC setup in Chapter 4 where only the caches of the receivers were filled during the placement phase, in the setup considered in this chapter both the caches at the transmitters and receivers are filled during the placement phase. In the delivery phase, transmitters can then only access the content in their own cache memory.

In the considered setup, communication during the delivery phase takes place over two heterogeneous parallel subchannels: one for which transmitters have access to the instantaneous channel coefficients (i.e. perfect CSIT), and another for which the transmitters have no knowledge of the instantaneous channel coefficients (i.e. no CSIT). The two subchannels are referred to as the P-subchannel and the N-subchannel, respectively. For the sake of generality, we assume that the two subchannels occupy arbitrary fractions of the bandwidth given by $\beta \in [0, 1]$ and $\bar{\beta} = 1 - \beta$, respectively. Different variants of this hybrid PN-parallel channel model have been widely adopted in information-theoretic studies focusing on capacity and DoF limits of wireless networks under CSIT imperfections (see e.g. [47, 125, 133, 134] and references therein). This wide adoption may be attributed to the fact that the PN-parallel channel model abstracts practically relevant scenarios in which channel state feedback is available only for a fraction of signalling dimensions, e.g. sub-carriers in OFDMA systems, due to limited feedback capabilities.

Moreover, recent results in [73, 125] have made the link between wireless networks where CSI is only reported for a fraction of the bandwidth and wireless networks where CSI is reported for the entire bandwidth but with a certain quality (for instance, receivers feedback the CSI over a certain number of bits). In fact, it was shown in [73] (the paper [125] is the extended journal version of [73]), in the context of a MISO BC with multiple parallel subchannels, that reporting partial CSIT over all subchannels allows to achieve the same sum-DoF than reporting perfect CSIT over a fraction of the subchannels, and no CSIT over the remaining subchannels. In particular, the result shows that that reporting perfect CSIT for a fraction β of the subchannels is equivalent to report a partial CSIT with quality β for all the subchannels (where the partial CSIT quality is defined as in the previous chapters). Furthermore, this setup and the results we obtained may also be linked to other related wireless and wired scenarios with mixed multicast and unicast capabilities as explained further on in Section 5.4.4, making it all the more relevant.

We would like to highlight that the main reason why we do not consider in this chapter the same partial CSIT definition as in the previous chapters, but a simplified version with the introduction of parallel subchannels, is to make the problem analytically more tractable. Hence, the work in this chapter is a first step towards the characterization of the fundamental limits of interfence management for cache-aided interference networks with partial CSIT.

In the same spirit of [31], we focus on separable one-shot linear delivery schemes where the spreading of channel symbols over time or frequency (i.e. subchannels) is not allowed. This is also known as linear precoding with no symbol extension [135]. Such linear schemes are appealing due to their practicality and their suitability for making theoretical progress on otherwise difficult or intractable information-theoretic problems.

5.2.1. Main Results and Contributions

Centralized Setting

For the above described setup, we first characterize an achievable one-shot linear sum-DoF under centralized placement and show that it is within a factor 2 from the optimal one-shot linear sum-DoF for all system parameters. This achievable one-shot linear sum-DoF is given by

$$\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) = \beta \cdot \min\{K_{\mathrm{T}}\mu_{\mathrm{T}} + K_{\mathrm{R}}\mu_{\mathrm{R}}, K_{\mathrm{R}}\} + \bar{\beta} \cdot \min\{1 + K_{\mathrm{R}}\mu_{\mathrm{R}}, K_{\mathrm{R}}\}.$$

From the separable nature of the proposed scheme, $\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ takes a weighted-sum form of $\beta \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)$, and is hence achieved by employing the scheme in [31] over the P-subchannel and the scheme in [24], with a slight modification, over the N-subchannel.

To prove the order-optimality, we derive an upper-bound for the one-shot linear sum-DoF by building upon the converse proof in [31], where an integer optimization problem is formulated and then a worst-case to average demands relaxation is employed. Further to the proof in [31] however, obtaining the upper-bound for the considered setup requires two more judicious steps, namely: a decoupling of the two subchannels and then a careful optimization over a delivery rate splitting ratio. This yields an upper-bound, denoted by $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$, which also takes a weighted-sum form of $\beta \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)$, hence reducing the task of proving order optimality to comparing $\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ and $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ at the two extreme points of $\beta = 0$ and $\beta = 1$ (see Sections 5.5 and 5.6).

Decentralized Setting

The insights gained from addressing the centralized setting are then employed to tackle a decentralized variant of the considered setup, which proves more technically challenging. In the considered decentralized setting, placement at the receivers is randomized and requires no central coordination. On the other hand, centralized placement at the transmitters is still allowed, as transmitters are assumed to be fixed nodes in the network, e.g. base stations, access points or servers. For this decentralized setting, we show that an achievable one-shot linear sum-DoF, which is within a factor of 3 from the optimal one-shot linear sum-DoF for all system parameters, is characterized by

$$\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) = \beta \cdot \frac{1}{\sum_{l=0}^{K_{\mathrm{R}}-1} \frac{\binom{K_{\mathrm{R}}-1}{l} \mu_{\mathrm{R}}^{l} (1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l-1}}{\min\{K_{\mathrm{T}}\mu_{\mathrm{T}}+l, K_{\mathrm{R}}\}}} + \bar{\beta} \cdot \frac{K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}}}$$

which evidently takes the weighted-sum form of $\beta \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)$.

Once again, order-optimality is shown by comparing $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ and $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ at the two extreme points $\beta = 0$ and $\beta = 1$. While the case $\beta = 0$ follows by a direct comparison of $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)$ and $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)$, the intricate form of $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)$ does not easily lend itself to such direct approach. Alternatively, we prove that $\frac{\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)}{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)} \leq \frac{\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)}{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)}$, which serves the same purpose. Showing that this last inequality holds true turns out to be particularly challenging and involves first reformulating it as a inequality involving a polynomial, and then proving a key quasiconcavity property for such polynomial from which the inequality follows (see Section 5.7).

As highlighted above, the main technical challenge for the decentralized setting, and in general of this chapter, is the proof that $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)$ is to within a factor 3 from the optimal one-shot linear sum-DoF. Moreover, as $\beta = 1$ corresponds to decentralized placement for the setting in [31], this is an important result on its own. Hence, a major outcome of this chapter is the proof that decentralized placement attains an order-optimal one-shot linear sum-DoF (to within a factor 3) in the setup in [31].

Related Works

We conclude this part by highlighting the connection to other works that consider related setups. It is evident that for $\beta = 1$, the considered setup reduces to the one in [31,34,106], where only centralized placement was considered. Since we adopt one-shot linear delivery schemes, our work is most related to [31] and expands upon it in two main directions: 1) the consideration of parallel heterogenous subchannels, and 2) the consideration of decentralized placement at the receivers. Note that already in [132,136] a decentralized variant of the setting in [31] was considered, with additional assumptions of partial connectivity and asymptotically large networks. The latter assumption allows for a considerable simplification of the achievable sum-DoF, which in turn, allows for a direct comparison with the corresponding upper-bound to show order-optimality[1]. This approach, however, does not work for the setting with finite transmitters and receivers considered here.

Regarding the point 1) above, the incorporation of parallel heterogeneous subchannels with the β parameter into cache-aided interference networks reveals a tradeoff between CSIT feedback budget and cache sizes as it will be described in Section 5.4.3. This tradeoff extends previous observations that were made for the cache-aided multi-antenna broadcast channel [32,110]. Regarding the point 2) above, decentralized scenarios, which are somewhat related the setting of this work, were considered in Chapter 4 and the works in [108,109,119,120]. In Chapter 4, the multi-antenna broadcast

[1] In particular, the achievable sum-DoF in [132] is approximated by moving a summation over the delivery time and the corresponding multicasting gains from the denominator into the numerator (see the expression of $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ for $\beta = 1$).

channel with partial CSIT was considered. As already pointed out earlier, the partial CSIT setting in Chapter 4 can be translated into the parallel subchannels setting of this chapter by considering the partial CSIT as the fraction of the bandwidth in which perfect CSIT is available. However, the assumption of full transmitter cooperation (i.e. $\mu_{\mathrm{T}} = 1$) as well as the assumption of the same number of trasmitters and receivers in Chapter 4 limits the applicability of the results in Chapter 4 to the setting of this chapter. We want to remark that, in case of full transmitter cooperation, already the work in [37] considered the case of a different number of transmitters and receivers. In particular, an overloaded cache-aided MISO BC where the number of receivers is larger than the number of antennas at the transmitter was considered, which can be seen as the translation of the setup in Chapter 3 to the cache-aided setting, while considering the same CSIT quality for all users.

The work in [108] introduced a cache-aided interference network with a setting with similar placement to the one considered in this chapter, i.e. centralized at the transmitters and decentralized at the receivers, and provided an achievable sum-DoF for an arbitrary number of transmitters and receivers, by focusing on achievable schemes with no proofs of order-optimality. On the other hand, [119,120] consider an F-RAN setting with randomized decentralized placement at both transmitters and receivers. However, decentralization at both ends necessitates cloud transmission through the front-haul in [119, 120], and the results are also not applicable to the setting considered in this work. Finally, [109] provides achievable schemes for a F-RAN setting with similar placement to the one considered here, i.e. centralized at the transmitters and decentralized at the receivers, with no proofs of optimality.

To conclude, we want to remark that, while establishing the information-theoretic limits of cache-aided interference management under partial CSIT and partial cooperation is a very challenging problem, the work in this chapter is a first step forward towards the solution of this broader problem. Moreover, while considering the P-subchannel only ($\beta = 1$), which corresponds to the setup in [31], an important consequence of this chapter is that decentralized placement attains order-optimal one-shot linear sum-DoF in the cache-aided interference channel with perfect CSIT considered in [31]. As for the remainder of the chapter, the organization is as follows. Section 5.3 introduces the considered setting and problem. Section 5.4 presents the two main results and related insights. In Section 5.5, we derive an outer-bound which is employed in the following two sections to show order optimality. In Section 5.6 and Section 5.7, we prove the two main results, the centralized setting result and the decentralized setting result respectively. Section 5.8 summarizes and concludes the chapter.

5.3. Problem Setting

The considered wireless network consists of K_{T} transmitters, denoted by $\{\mathrm{Tx}_i\}_{i=1}^{K_T}$, and K_{R} receivers (or users), denoted by $\{\mathrm{Rx}_i\}_{i=1}^{K_R}$. The wireless channel comprises two parallel subchannels: 1) the P-subchannel for which the transmitters have perfect CSIT, and 2) the N-subchannel for which the transmitters have no CSIT[2]. We assume that the capacities of single links in the

[2] Also here, the CSIR is assumed to be perfectly available at all receivers.

P-subchannel and the N-subchannel are given by $\beta \log P + o(\log P)$ and $\bar{\beta} \log P + o(\log P)$ respectively, where $\beta \in [0,1]$ and $\bar{\beta} \triangleq 1-\beta$ are the corresponding normalized single link capacities (or DoF) and P is the SNR. Note that under the normalization $0 \leq \beta \leq 1$, the parameters β and $\bar{\beta}$ can be interpreted as the fractions of the total bandwidth for which CSIT is perfect and not available respectively, in a DoF sense.

Communication over the two subchannels at time (or channel use) t is modeled by

$$Y_j^{(\mathrm{p})}(t) = \sqrt{P^{\beta}} \sum_{i=1}^{K_{\mathrm{T}}} G_{ji}^{(\mathrm{p})}(t) X_i^{(\mathrm{p})}(t) + Z_j^{(\mathrm{p})}(t) \tag{5.1}$$

$$Y_j^{(\mathrm{n})}(t) = \sqrt{P^{\bar{\beta}}} \sum_{i=1}^{K_{\mathrm{T}}} G_{ji}^{(\mathrm{n})}(t) X_i^{(\mathrm{n})}(t) + Z_j^{(\mathrm{n})}(t) \tag{5.2}$$

where for the P-subchannel and the N-subchannel respectively, $X_i^{(\mathrm{p})}(t)$ and $X_i^{(\mathrm{n})}(t)$ denote the signals transmitted by Tx$_i$, $i \in [K_{\mathrm{T}}] \triangleq \{1,\ldots,K_{\mathrm{T}}\}$, while $Y_j^{(\mathrm{p})}(t)$ and $Y_j^{(\mathrm{n})}(t)$ denote the signals received by Rx$_j$, $j \in [K_{\mathrm{R}}]$. Moreover, $Z_j^{(\mathrm{p})}(t)$ and $Z_j^{(\mathrm{n})}(t)$ denote the corresponding additive white Gaussian noise signals at Rx$_j$, distributed as $\mathcal{N}_{\mathbb{C}}(0,1)$. $G_{ji}^{(\mathrm{p})}(t)$ and $G_{ji}^{(\mathrm{n})}(t)$ denote the fading channel coefficients from Tx$_i$ to Rx$_j$, drawn from continuous stationary ergodic processes such that $G_{ji}^{(\mathrm{p})}(t)$, $\forall i,j,t$, are perfectly known to the transmitters (perfect CSIT), while $G_{ji}^{(\mathrm{n})}(t)$, $\forall i,j,t$, are not known to the transmitters (no CSIT). The transmit signals at Tx$_i$, $i \in [K_{\mathrm{T}}]$, are subject to the power constraints $\mathbb{E}\big[|X_i^{(\mathrm{p})}(t)|^2\big] \leq 1$ and $\mathbb{E}\big[|X_i^{(\mathrm{n})}(t)|^2\big] \leq 1$. Note that P is a nominal power (or SNR) value, borrowed from the GDoF framework in Chapter 4 and in [66], which alongside β and $\bar{\beta}$ is used to distinguish the strengths of the two subchannels.

In any communication session, each user requests an arbitrary file out of a content library of N files given by $\mathcal{W} \triangleq \{W_1,\ldots,W_N\}$. Following the same model in [31], each file W_n consists of F packets, denoted by $\{\mathbf{w}_{n,f}\}_{f=1}^{F}$, where each packet is a vector of B bits, i.e. $\mathbf{w}_{n,f} \in \mathbb{F}_2^B$. Furthermore, each transmitter Tx$_i$, $i \in [K_{\mathrm{T}}]$, is equipped with a cache memory $\mathcal{P}_i$ of size $M_{\mathrm{T}}F$ packets, while each receiver Rx$_j$, $j \in [K_{\mathrm{R}}]$, is equipped with a cache memory $\mathcal{U}_j$ of size $M_{\mathrm{R}}F$ packets. We assume that each cache memory, whether at transmitters or receivers, can be used to cache arbitrary contents from the library before communication sessions begin. Moreover, we assume that $K_{\mathrm{T}}M_{\mathrm{T}} \geq N$, which ensures that the entire library $\mathcal{W}$ can be cached across the collective memory of all transmitters.

We define the *normalized transmitter cache size* and the *normalized receiver cache size* as $\mu_{\mathrm{T}} = \frac{M_{\mathrm{T}}}{N}$ and $\mu_{\mathrm{R}} = \frac{M_{\mathrm{R}}}{N}$, respectively. For the sake of convenience, we assume that $K_{\mathrm{T}}\mu_{\mathrm{T}}$ and $K_{\mathrm{R}}\mu_{\mathrm{R}}$ have integer values whenever we deal with the centralized case, while only $K_{\mathrm{T}}\mu_{\mathrm{T}}$ is assumed to be integer for the decentralized case. This is not a major restriction as schemes that correspond to the remaining values are realized through memory-sharing. We next describe the *placement phase* and the *delivery phase*.

5.3.1. Placement Phase

Following the assumptions in [31], placement is done at the packet level, i.e. each memory is filled with an arbitrary subset of the NF packets in the library where the breaking of packets into smaller subpackets is not allowed. Moreover, *uncoded placement* is assumed [88, 137], where it is not allowed to cache combinations of multiple packets as a single packet. Note that this is more restrictive than the delivery phase considered in Chapter 4, where no assumptions about the placement scheme were made. However, it turned out that a scheme with uncoded placement was sufficient to obtain order-optimal performance.

Besides considering *centralized placement*, in which coordination amongst nodes during the placement phase is allowed, we also consider as in Chapter 4 *decentralized placement* where no coordination amongst receivers is allowed during the placement phase. Centralized placement at the transmitters, however, is always assumed throughout this work, as transmitters are considered to be fixed nodes in the network.

5.3.2. Delivery Phase

In this phase, each receiver Rx_j reveals its request for an arbitrary file W_{d_j}, where $d_j \in [N]$. The tuple of all user demands is denoted by $\mathbf{d} = (d_1, \ldots, d_K)$. As each receiver Rx_j has the subset of requested packets, given by $\{\mathbf{w}_{d_j,f}\}_{f=1}^{F} \cap \mathcal{U}_j$, pre-stored in its cache memory, the transmitters are required to deliver the remaining packets given by $\{\mathbf{w}_{d_j,f}\}_{f=1}^{F} \setminus \mathcal{U}_j$, for all $j \in [K_{\mathrm{R}}]$. Given the demands $\mathbf{d}$ and the receiver caching realization $\{\mathcal{U}_j\}_{j=1}^{K_{\mathrm{R}}}$, the set of all packets to be delivered is given by

$$\mathcal{D}\big(\mathbf{d}, \{\mathcal{U}_j\}_{j=1}^{K_{\mathrm{R}}}\big) = \bigcup_{j=1}^{K_{\mathrm{R}}} \big\{\{\mathbf{w}_{d_j,f}\}_{f=1}^{F} \setminus \mathcal{U}_j\big\}.$$

Packet Splitting and Encoding: Unlike the placement phase, in which the breaking of packets is not allowed, we assume that each packet to be transmitted in the delivery phase is split into two subpackets, as communication is carried out over two parallel subchannels. In particular, each packet $\mathbf{w}_{n,f}$ is split as

$$\mathbf{w}_{n,f} = \big(\mathbf{w}_{n,f}^{(\mathrm{p})}, \mathbf{w}_{n,f}^{(\mathrm{n})}\big)$$

where $\mathbf{w}_{n,f}^{(\mathrm{p})}$ and $\mathbf{w}_{n,f}^{(\mathrm{n})}$ are referred to as the P-subpacket and the N-subpacket, respectively. Without loss of generality, we assume that $\mathbf{w}_{n,f}^{(\mathrm{p})}$ and $\mathbf{w}_{n,f}^{(\mathrm{n})}$ consist of the first qB bits and the last $\bar{q}B$ bits of $\mathbf{w}_{n,f}$, respectively, where the *splitting ratio* $q \in [0, 1]$ is a design parameter and $\bar{q} \triangleq 1 - q$. Moreover, while q may depend on β (i.e. long-term channel parameters), we assume that q is fixed at the beginning of the delivery phase and is not allowed to depend on the fading coefficients or the user demands. From the above, each transmitter cache $\mathcal{P}_i$ is split into $\mathcal{P}_i^{(\mathrm{p})}$ and $\mathcal{P}_i^{(\mathrm{c})}$, containing P-subpackets and N-subpackets respectively. Similarly, a set of packets to be delivered $\mathcal{D}$ is split into $\mathcal{D}^{(\mathrm{p})}$ and $\mathcal{D}^{(\mathrm{c})}$.

Each subpacket cached by the transmitters is encoded into a *coded subpacket* using an independent random Gaussian code. In particular, a coding scheme $\psi^{(\mathrm{p})} : \mathbb{F}_2^{qB} \to \mathbb{C}^{\tilde{B}^{(\mathrm{p})}}$ of rate

$\beta \log P + o(\log P)$ is used to encode P-subpackets, while a scheme $\psi^{(\mathrm{n})} : \mathbb{F}_2^{(1-q)B} \to \mathbb{C}^{\tilde{B}^{(\mathrm{n})}}$ of rate $\bar{\beta} \log P + o(\log P)$ is used to encode N-subpackets [3]. The coded versions of the P-subpacket $\mathbf{w}_{n,f}^{(\mathrm{p})}$ and the N-subpacket $\mathbf{w}_{n,f}^{(\mathrm{n})}$, defined as $\tilde{\mathbf{w}}_{n,f}^{(\mathrm{p})} \triangleq \psi^{(\mathrm{p})}(\mathbf{w}_{n,f}^{(\mathrm{p})})$ and $\tilde{\mathbf{w}}_{n,f}^{(\mathrm{n})} \triangleq \psi^{(\mathrm{n})}(\mathbf{w}_{n,f}^{(\mathrm{n})})$ respectively, are given in terms of channel symbols as

$$\tilde{\mathbf{w}}_{n,f}^{(\mathrm{p})} = \big(\tilde{W}_{n,f}^{(\mathrm{p})}(1), \dots, \tilde{W}_{n,f}^{(\mathrm{p})}(\tilde{B}^{(\mathrm{p})})\big) \tag{5.3}$$

$$\tilde{\mathbf{w}}_{n,f}^{(\mathrm{n})} = \big(\tilde{W}_{n,f}^{(\mathrm{n})}(1), \dots, \tilde{W}_{n,f}^{(\mathrm{n})}(\tilde{B}^{(\mathrm{n})})\big). \tag{5.4}$$

It is clear that a coded P-subpacket carries a DoF of β, while a coded N-subpacket carries a DoF of $\bar{\beta}$, which is in tune with the single link capacities of the corresponding subchannels.

Block Structure: Communication of coded subpackets is carried out independently over the P-subchannel and the N-subchannel. Communication in the P-subchannel takes place over $H^{(\mathrm{p})}$ blocks, each referred to as a P-block and spanning $\tilde{B}^{(\mathrm{p})}$ channel uses, while communication in the N-subchannel takes place over $H^{(\mathrm{n})}$ blocks, each referred to as a N-block and spanning $\tilde{B}^{(\mathrm{n})}$ channel uses.

The goal in each P-block $m^{(\mathrm{p})} \in [H^{(\mathrm{p})}]$ is to deliver a subset of P-subpackets $\mathcal{D}_{m^{(\mathrm{p})}}^{(\mathrm{p})} \subseteq \mathcal{D}^{(\mathrm{p})}$ to a subset of receivers, denoted by $\mathcal{R}_{m^{(\mathrm{p})}}^{(\mathrm{p})}$, such that one P-subpacket is intended exactly for one receiver. Similarly, in each N-block $m^{(\mathrm{n})} \in [H^{(\mathrm{n})}]$, the goal is to deliver the N-subpackets in $\mathcal{D}_{m^{(\mathrm{n})}}^{(\mathrm{n})} \subseteq \mathcal{D}^{(\mathrm{n})}$ to the subset of receivers $\mathcal{R}_{m^{(\mathrm{n})}}^{(\mathrm{n})}$. At the end of the communication, for each receiver Rx$_j$ to be able to retrieved its requested file, the sets of delivered subpackets and the content of the cache memory $\mathcal{U}_j$ should satisfy

$$W_{d_j}^{(\mathrm{p})} \triangleq \{\mathbf{w}_{d_j,f}^{(\mathrm{p})}\}_{f=1}^{F} \subset \left(\bigcup_{m^{(\mathrm{p})}=1}^{H^{(\mathrm{p})}} \mathcal{D}_{m^{(\mathrm{p})}}^{(\mathrm{p})} \right) \cup \mathcal{U}_j^{(\mathrm{p})} \tag{5.5}$$

$$W_{d_j}^{(\mathrm{n})} \triangleq \{\mathbf{w}_{d_j,f}^{(\mathrm{n})}\}_{f=1}^{F} \subset \left(\bigcup_{m^{(\mathrm{n})}=1}^{H^{(\mathrm{n})}} \mathcal{D}_{m^{(\mathrm{n})}}^{(\mathrm{n})} \right) \cup \mathcal{U}_j^{(\mathrm{n})} \tag{5.6}$$

where $\mathcal{U}_j^{(\mathrm{p})}$ and $\mathcal{U}_j^{(\mathrm{n})}$ are the portions of $\mathcal{U}_j$ that correspond to P-subpackets and N-subpackets respectively, i.e. the first qB bits and the last $\bar{q}B$ bits, respectively, of packets in $\mathcal{U}_j$. Similarly, $W_{d_j}^{(\mathrm{p})}$ and $W_{d_j}^{(\mathrm{n})}$ are the portions of W_{d_j} that correspond to P-subpackets and N-subpackets respectively. As in [31], we adopt one-shot linear delivery schemes in each subchannel, i.e. *each encoded channel symbol is beamformed in one channel use, where spreading over multiple channel uses is not allowed.*

Transmit Linear Beamforming: Transmission of coded subpackets in each P-block and N-block is carried out using linear beamforming. In particular, consider the $m^{(\mathrm{p})}$-th P-block, where $m^{(\mathrm{p})} \in [H^{(\mathrm{p})}]$. The transmitter Tx$_i$, $i \in [K_T]$, transmits a linear combination of the P-subpackets in $\mathcal{P}_i^{(\mathrm{p})}$

[3] Note that both the number of packets F and the number of bits per packet B may grown infinitely large.

and $\mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}$ given by

$$X_i^{(\mathrm{p})}(t) = \sum_{\substack{(n,f): \\ \mathbf{w}^{(\mathrm{p})}_{n,f} \in \mathcal{P}_i^{(\mathrm{p})} \cap \mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}}} v^{(\mathrm{p})}_{i,n,f}(t) \cdot \tilde{W}^{(\mathrm{p})}_{n,f}(t), \; t \in \big[(m^{(\mathrm{p})}-1)\tilde{B}^{(\mathrm{p})}+1 : m^{(\mathrm{p})}\tilde{B}^{(\mathrm{p})}\big] \tag{5.7}$$

where $[t_1 : t_2] \triangleq \{t_1, t_1+1, \ldots, t_2\}$. In (5.7), each $v^{(\mathrm{p})}_{i,n,f}(t)$ is a complex beamforming coefficient used at time t over the P-subchannel, which is allowed to depend on the channel coefficients of the P-subchannel due to perfect CSIT (e.g. as in [31]). Similarly, for the $m^{(\mathrm{n})}$-th N-block, where $m^{(\mathrm{n})} \in [H^{(\mathrm{n})}]$, Tx$_i$ transmits a linear combination of the P-subpackets in $\mathcal{P}_i^{(\mathrm{n})}$ and $\mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}}$ given by

$$X_i^{(\mathrm{n})}(t) = \sum_{\substack{(n,f): \\ \mathbf{w}^{(\mathrm{n})}_{n,f} \in \mathcal{P}_i^{(\mathrm{n})} \cap \mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}}}} v^{(\mathrm{n})}_{i,n,f}(t) \cdot \tilde{W}^{(\mathrm{n})}_{n,f}(t), \; t \in \big[(m^{(\mathrm{n})}-1)\tilde{B}^{(\mathrm{n})}+1 : m^{(\mathrm{n})}\tilde{B}^{(\mathrm{n})}\big] \tag{5.8}$$

where each $v^{(\mathrm{n})}_{i,n,f}(t)$ is a complex beamforming coefficient, which is not allowed to depend on the channel coefficients of the N-subchannel due to no CSIT. Note that in (5.7) and (5.8), we implicitly assume that $\tilde{W}^{(\mathrm{p})}_{n,f}(t) = \tilde{W}^{(\mathrm{p})}_{n,f}(t \bmod \tilde{B}^{(\mathrm{p})})$, $\tilde{W}^{(\mathrm{p})}_{n,f}(0) = \tilde{W}^{(\mathrm{p})}_{n,f}(\tilde{B}^{(\mathrm{p})})$, $\tilde{W}^{(\mathrm{n})}_{n,f}(t) = \tilde{W}^{(\mathrm{n})}_{n,f}(t \bmod \tilde{B}^{(\mathrm{n})})$ and $\tilde{W}^{(\mathrm{n})}_{n,f}(0) = \tilde{W}^{(\mathrm{n})}_{n,f}(\tilde{B}^{(\mathrm{n})})$, to maintain consistency with (5.3) and (5.4). Moreover, the coded subpackets and beamforming coefficients are designed such that the transmit power constraints are respected.

Receive Linear Combining: Transmit signals pass through the channel modeled in (5.1) and (5.2). The signals received by Rx$_j$, $j \in [K_\mathrm{R}]$, in the P-block $m^{(\mathrm{p})}$ and the N-block $m^{(\mathrm{n})}$ are given by

$$\mathbf{y}_j^{(\mathrm{p})}(m^{(\mathrm{p})}) = \Big(Y_j^{(\mathrm{p})}(t) : t \in \big[(m^{(\mathrm{p})}-1)\tilde{B}^{(\mathrm{p})}+1 : m^{(\mathrm{p})}\tilde{B}^{(\mathrm{p})}\big]\Big) \tag{5.9}$$

$$\mathbf{y}_j^{(\mathrm{n})}(m^{(\mathrm{n})}) = \Big(Y_j^{(\mathrm{n})}(t) : t \in \big[(m^{(\mathrm{n})}-1)\tilde{B}^{(\mathrm{n})}+1 : m^{(\mathrm{n})}\tilde{B}^{(\mathrm{n})}\big]\Big) \tag{5.10}$$

where $\big(Y(t) : t \in [t_1 : t_2]\big) \triangleq \big(Y(t_1), \ldots, Y(t_2)\big)$. Focusing on the P-subchannel first and following the linear scheme proposed in [31], each receiver Rx$_j$ in $\mathcal{R}^{(\mathrm{p})}_{m^{(\mathrm{p})}}$ uses the content of its cache to subtract the interference of the undersidered subpackets in $\mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}$, transmitted in the P-block $m^{(\mathrm{p})}$, $m^{(\mathrm{p})} \in [H^{(\mathrm{p})}]$. This is achieved through a linear combination $\mathcal{L}^{(\mathrm{p})}_{j,m^{(\mathrm{p})}}(\mathbf{y}_j^{(\mathrm{p})}(m^{(\mathrm{p})}), \tilde{\mathcal{U}}_j^{(\mathrm{p})})$ formed to recover $\mathbf{w}^{(\mathrm{p})}_{d_j,f} \in \mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}$, where $\tilde{\mathcal{U}}_j^{(\mathrm{p})}$ denotes the set of coded P-subpackets cached at Rx$_j$. The communication in the $m^{(\mathrm{p})}$-th P-block is successful if there exists linear combinations at the transmitters (i.e. beamformers) and linear combinations at the receivers such that for all Rx$_j$ in $\mathcal{R}^{(\mathrm{p})}_{m^{(\mathrm{p})}}$, we have

$$\mathcal{L}^{(\mathrm{p})}_{j,m^{(\mathrm{p})}}(\mathbf{y}_j^{(\mathrm{p})}(m^{(\mathrm{p})}), \tilde{U}_j^{(\mathrm{p})}) = \sqrt{P^\beta}\tilde{\mathbf{w}}^{(\mathrm{p})}_{d_j,f} + \mathbf{z}_j^{(\mathrm{p})}(m^{(\mathrm{p})}) \tag{5.11}$$

where $\mathbf{z}_j^{(\mathrm{p})}(m^{(\mathrm{p})})$ is a sequence of $\mathcal{N}_\mathbb{C}(0,1)$ noise samples. The point-to-point channel in (5.11) has a capacity of $\beta \log P + o(\log P)$, and therefore $\tilde{\mathbf{w}}^{(\mathrm{p})}_{d_j,f}$ is reliably communicated as qB grows large.

In a similar manner, considering the N-block $m^{(\mathrm{n})}$, $m^{(\mathrm{n})} \in [H^{(\mathrm{n})}]$, each receiver Rx_j in $\mathcal{R}^{(\mathrm{n})}_{m^{(\mathrm{n})}}$ forms a linear combination $\mathcal{L}^{(\mathrm{n})}_{j,m^{(\mathrm{n})}}(\mathbf{y}^{(\mathrm{n})}_j(m^{(\mathrm{n})}), \tilde{\mathcal{U}}^{(\mathrm{n})}_j)$ to recover $\mathbf{w}^{(\mathrm{n})}_{d_j,f} \in \mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}}$, where $\tilde{\mathcal{U}}^{(\mathrm{n})}_j$ denotes the set of coded N-subpackets cached at Rx_j. The communication in the $m^{(\mathrm{n})}$-th N-block is successful if there exists linear combinations at the transmitters and linear combinations at the receivers such that

$$\mathcal{L}^{(\mathrm{n})}_{j,m^{(\mathrm{n})}}(\mathbf{y}^{(\mathrm{n})}_j(m^{(\mathrm{n})}), \tilde{\mathcal{U}}^{(\mathrm{n})}_j) = \sqrt{P^{\bar{\beta}}}\tilde{\mathbf{w}}^{(\mathrm{n})}_{d_j,f} + \mathbf{z}^{(\mathrm{n})}_j(m^{(\mathrm{n})}) \tag{5.12}$$

where the point-to-point channel channel in (5.12) has a capacity $\bar{\beta}\log P + o(\log P)$, and therefore $\tilde{\mathbf{w}}^{(\mathrm{n})}_{d_j,f}$ is reliably communicated as $\bar{q}B$ grows large.

5.3.3. Delivery Time and DoF

As already mentioned in the previous sections, the performance metric utilized in this chapter is the DoF. However, we have to adapt the definition of DoF to take into account the we are considering one-shot linear DoF schemes. To achieve this, we start by defining a metric for the delivery time similar as the GNDT in Section 4.3.2, which in turn will help to define the DoF. We start by defining the unit of the delivery time, i.e. the packet-time-slot. One packet-time-slot is defined as the optimal time required to communicate a single packet to a single user, under no caching and no interference, as $P \to \infty$. This is achieved by setting $q = \beta$, and hence communicating βB bits over the P-subchannel at rate $\beta \log P + o(\log P)$ bits per channel use and $\bar{\beta}B$ bits over the N-subchannel at rate $\bar{\beta}\log P + o(\log P)$ bits per channel use. Therefore, a packet-time-slot is equivalent to $\frac{B}{\log P}$ uses of the channel (or time instances). It follows that an achievable sum-DoF can be interpreted as an achievable sum-rate, measured in packets per packet-time-slots as $P \to \infty$.

In general, for any feasible linear delivery scheme as described in Section 5.3.2, each P-subpacket consists of qB bits and is delivered in one P-block over the point-to-point channel in (5.11) at rate $\beta \log P + o(\log P)$. It follows that a P-block has a duration of $\frac{q}{\beta}$ packet-time-slots. Similarly, each N-subpacket consists of $\bar{q}B$ bits and is delivered over the point-to-point channel in (5.12) at rate $\bar{\beta}\log P + o(\log P)$, and hence an N-block has a duration of $\frac{\bar{q}}{\bar{\beta}}$ packet-time-slots. It follows that the delivery time for a feasible scheme is given by $H = \max\left\{\frac{q}{\beta}H^{(\mathrm{p})}, \frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})}\right\}$ packet-time-slots, and the achievable sum-DoF is given by $\frac{|\mathcal{D}|}{H}$. Therefore, for fixed caching realization $\left(\{\mathcal{P}_i\}_{i=1}^{K_\mathrm{T}}, \{\mathcal{U}_j\}_{j=1}^{K_\mathrm{R}}\right)$ and splitting ratio q, which are independent of user demands, the maximum achievable one-shot linear sum-DoF (DoF for short) for the worst case demands is given by

$$\mathsf{DoF}_{\mathrm{L}}^{(\{\mathcal{P}_i\}_{i=1}^{K_\mathrm{T}}, \{\mathcal{U}_j\}_{j=1}^{K_\mathrm{R}}, q)} = \inf_{\mathbf{d}} \sup_{\substack{H^{(\mathrm{p})},\, H^{(\mathrm{n})},\\ \{\mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}\}_{m^{(\mathrm{p})}=1}^{H^{(\mathrm{p})}}, \{\mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}}\}_{m^{(\mathrm{n})}=1}^{H^{(\mathrm{n})}}}} \frac{\left|\mathcal{D}\left(\mathbf{d}, \{\mathcal{U}_j\}_{j=1}^{K_\mathrm{R}}\right)\right|}{\max\left\{\frac{q}{\beta}H^{(\mathrm{p})}, \frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})}\right\}}. \tag{5.13}$$

Note that, differently from the definition of the GDoF in Section 4.3.2, we do not include here the contribution of the local caching gain. This in turn allows to simplify the calculations later on and to directly compare the results in this chapter with the ones in [31].

The formulation in Eq. (5.13) leads to the definition of the *one-shot linear DoF* of the network as the maximum achievable one-shot linear sum-DoF over all caching realizations and splitting ratios, i.e.

$$\begin{aligned} \mathsf{DoF}_{\mathrm{L}}^{*}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta) = \sup_{\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_j\}_{j=1}^{K_{\mathrm{R}}},q} \quad & \mathsf{DoF}_{\mathrm{L}}^{(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_j\}_{j=1}^{K_{\mathrm{R}}},q)} \\ \text{s.t.} \;\; & |\mathcal{P}_i| = \mu_{\mathrm{T}} NF,\ \forall i \in [K_{\mathrm{T}}] \\ & |\mathcal{U}_j| = \mu_{\mathrm{R}} NF,\ \forall j \in [K_{\mathrm{R}}] \\ & q \in [0,1]. \end{aligned} \tag{5.14}$$

5.4. Main Results

In this sections we present the main results of the chapter. The proofs are deferred to subsequent sections and appendices. We start with the centralized setting and then move on to the decentralized setting.

5.4.1. Centralized Setting

Theorem 5.1. *For the cache-aided wireless network described in Section 4.3, assuming centralized placement, an achievable one-shot linear sum-DoF is given by*

$$\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta) = \beta \cdot \min\{K_{\mathrm{T}}\mu_{\mathrm{T}} + K_{\mathrm{R}}\mu_{\mathrm{R}}, K_{\mathrm{R}}\} + \bar{\beta} \cdot \min\{1 + K_{\mathrm{R}}\mu_{\mathrm{R}}, K_{\mathrm{R}}\}. \tag{5.15}$$

Moreover, $\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta)$ *satisfies*

$$\frac{\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta)}{\mathsf{DoF}_{\mathrm{L}}^{*}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta)} \geq \frac{1}{2}, \tag{5.16}$$

where $\mathsf{DoF}_{\mathrm{L}}^{*}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta)$ *is the one-shot linear DoF of the network as defined in* (5.14).

The proof of Theorem 5.1 is presented in Section 5.6 and employs the result derived in Section 5.5. From Theorem 5.1, the result in [31, Th. 1] is recovered by setting $\beta = 1$ (P-subchannel only). In this case, we know from [31] that perfect CSIT and caches at the transmitters allow cooperation and $\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},1)$ scales with the aggregate memory of all transmitters and receivers. On the other hand, when $\beta = 0$ (N-subchannel only), all DoF benefits of transmitter-side cooperation are annihilated following the result in Section 4.4, and the achievable one-shot linear sum-DoF in Theorem 5.1 reduces to the sum-DoF achieved with one transmitter [24]. In this case, the original Maddah-Ali and Niesen scheme [24] is implemented, where the XoR takes place over the air through superposition of coded packets, and $\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},0)$ scales with the aggregate memory of the receivers only. For general β, $\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta)$ takes the form

$$\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},\beta) = \beta \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}},\mu_{\mathrm{R}},0), \tag{5.17}$$

which is achieved by choosing an adequate splitting ratio q (as a function of β) in order to best utilize the two subchannels. Once q is chosen, the P-subpackets and N-subpackets are then delivered over the P-subchannel and N-subchannel as for the cases with $\beta = 1$ and $\beta = 0$, respectively. The result in Eq. (5.17) can be easily linked to the result in Eq. (4.14). In particular, by considering full transmitter cooperation and the same number of transmitters and receivers, i.e. $\mu_{\mathrm{T}} = 1$ and $K_{\mathrm{T}} = K_{\mathrm{R}}$, it can be readily seen that (5.17) coincides with (4.14), while not considering the local caching gain (for more details look at Section 5.3.3).

5.4.2. Decentralized Setting

In this part we consider the decentralized setting where centrally coordinated placement is only allowed at the transmitters and not at the receivers side during the placement phase.

Theorem 5.2. *For the cache-aided wireless network described in Section 4.3, under decentralized placement in which centrally coordinated placement is only allowed at the transmitters and not at the receivers, an achievable one-shot linear sum-DoF is given by*

$$\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) = \beta \cdot \frac{1}{\sum_{l=0}^{K_{\mathrm{R}}-1} \frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l-1}}{\min\{K_{\mathrm{T}}\mu_{\mathrm{T}}+l, K_{\mathrm{R}}\}}} + \bar{\beta} \cdot \frac{K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}}}. \tag{5.18}$$

Moreover, $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ *satisfies*

$$\frac{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)}{\mathsf{DoF}_{\mathrm{L}}^{*}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)} \geq \frac{1}{3}. \tag{5.19}$$

The proof of Theorem 5.2 is presented in Section 5.7. Choosing $\beta = 1$ in Theorem 5.2 is equivalent to consider decentralized placement for the setting of [31]. Hence, given the order-optimality result, also in the decentralized version of [31] the sum-DoF scales with the aggregate cache memory of all transmitters and receivers. Proving order-optimality for the case $\beta = 1$ is the main technical challenge of the chapter and it corresponds to an important result on its own. Hence, we summarize it as a theorem, which directly follows from Theorem 5.2 by substituting $\beta = 1$.

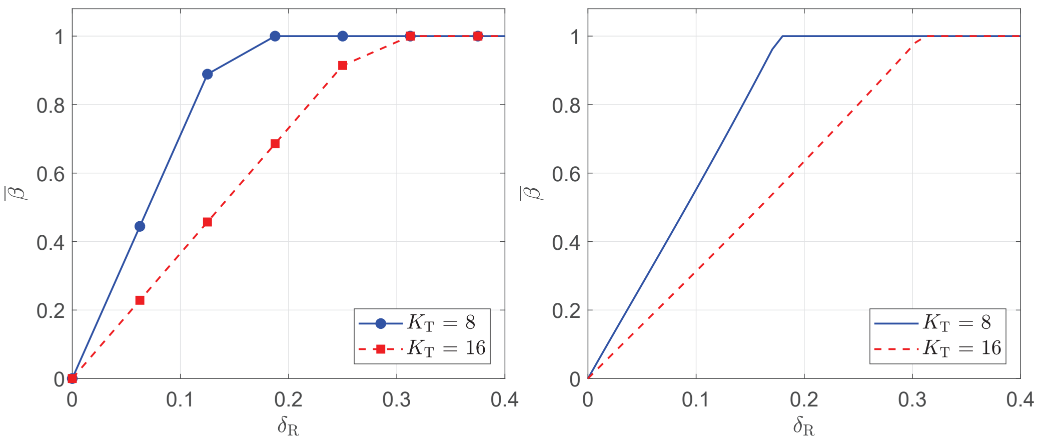
Theorem 5.3. *For the cache-aided wireless network with perfect CSIT described in [31], under decentralized placement in which centrally coordinated placement is only allowed at the transmitters and not at the receivers, an achievable one-shot linear sum-DoF is given by*

$$\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}) = \cdot \frac{1}{\sum_{l=0}^{K_{\mathrm{R}}-1} \frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l-1}}{\min\{K_{\mathrm{T}}\mu_{\mathrm{T}}+l, K_{\mathrm{R}}\}}}. \tag{5.20}$$

Moreover, $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}})$ *satisfies*

$$\frac{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}})}{\mathsf{DoF}_{\mathrm{L}}^{*}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}})} \geq \frac{1}{3}. \tag{5.21}$$

Note that β has been omitted in the notation as assumed equal to 1.



(a) Centralized Setting

(b) Decentralized Setting

Figure 5.1.: Tradeoff between δ_{R} and $\bar{\beta}$ for networks with $K_{\mathrm{R}} = 16$, $K_{\mathrm{T}} \in \{8, 16\}$, $\mu_{\mathrm{R}} = 1/16$ and $\mu_{\mathrm{T}} = 1/2$.

On the other hand, $\beta = 0$ reduces the setup to the decentralized setting in [79] in a DoF sense (the smaller multiplicative gap is due to uncoded placement and linear delivery). In general, similar to Theorem 5.1, $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$ takes the form

$$\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) = \beta \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0). \tag{5.22}$$

Moreover, one could easily conclude from Theorem 5.1 and Theorem 5.2 that centralized placement at the receivers can only lead to at most a factor of 3 improvement over decentralized placement. Furthermore, we observe through numerical simulations that this multiplicative factor does not exceed 1.5, which is in agreement with the result obtained in Chapter 4.1.

5.4.3. Tradeoff Between Receiver Cache Size and CSIT Budget

In this part, we investigate the implications of Theorem 5.1 and Theorem 5.2 by considering the tradeoff between the receiver cache memory size and the CSIT budget. For this purpose, we start by assuming that CSIT is perfectly available across all signalling dimensions, captured by $\beta = 1$ (equivalently $\bar{\beta} = 0$). For given μ_{T} and μ_{R}, an achievable delivery time under centralized placement, denoted by $H_{\mathrm{C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)$, is easily derived from the one-shot linear sum-DoF in Theorem 5.1. Now suppose that the CSIT budget is reduced, e.g. by providing feedback for a fraction of sub-carriers. This yields $H_{\mathrm{C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1 - \bar{\beta}) \geq H_{\mathrm{C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)$, where $\bar{\beta}$ is interpreted as the reduction in CSIT budget. We are interested in the corresponding increase in receiver cache size, i.e. $\delta_{\mathrm{R}} \in [0, 1 - \mu_{\mathrm{R}}]$, such that $H_{\mathrm{C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}} + \delta_{\mathrm{R}}, 1 - \bar{\beta}) = H_{\mathrm{C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)$. Note that a similar tradeoff is defined for the decentralized case through $H_{\mathrm{D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}} + \delta_{\mathrm{R}}, 1 - \bar{\beta}) = H_{\mathrm{D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)$.

The tradeoff between μ_{R} and $\bar{\beta}$ is evaluated numerically and illustrated in Fig. 5.1 for both centralized and decentralized cases. In particular, we consider a network of $K_{\mathrm{R}} = 16$ receivers with $\mu_{\mathrm{R}} = 1/16$ and $\mu_{\mathrm{T}} = 1/2$. The number of transmitters K_{T} is varied between 8 and 16. It can be seen that the tradeoff is sharper for $K_{\mathrm{T}} = 8$ compared to $K_{\mathrm{T}} = 16$ in the sense that a higher reduction in CSIT $\bar{\beta}$ can be achieved for a smaller increase in receiver cache size given by δ_{R}. This is due to the fact that at most 8 orthogonal beams can be created (through e.g. zero-forcing) in the setting with $K_{\mathrm{T}} = 8$, while $K_{\mathrm{T}} = 16$ allows up to 16 orthogonal beams. This makes the latter setting more dependent on CSIT in general, hence requiring a higher increase in cache size to compensate for the same reduction in CSIT budget.

5.4.4. Related Setups

It is worthwhile highlighting that the results in Theorem 5.1 and Theorem 5.2 can be easily applied to other related setups. In particular, the N-subchannel can be replaced by a $(K_{\mathrm{T}}+1)$-th transmitter, operating on a different frequency (e.g. a WiFi access point of femtocell), and connected to all transmitter caches through a multicast capacitated link (captured by $\bar{\beta}$) [124]. In practise, this scenario is realized when the receivers are connected to a wireless cellular networks over the P-subchannel and they are also in proximity of a WiFi access point of a femtocell base station, which represents the aforementioned $(K_{\mathrm{T}} + 1)$-th transmitter. Note that, in this case, the ergodic fading assumptions of our original setting can be relaxed, particularly if perfect CSI is also available at the $(K_{\mathrm{T}} + 1)$-th transmitter.

The results also extend to the multi-server setting of [100] with wired (noiseless) linear networks, in which the parallel subchannels correspond to scenarios where servers can reach receivers through two parallel networks: a fully connected linear interference network, which corresponds to the P-subchannel, and a multicast networks, which corresponds to the N-subchannel.

5.5. One-Shot Linear Sum-DoF Upper-Bound

In this section, we obtain an upper-bound of the one-shot linear sum-DoF of the network given as follows.

Theorem 5.4. *For the cache-aided wireless network described in Section 4.3, the one-shot linear sum-DoF of the network, defined in* (5.14)*, is bounded above as*

$$\mathsf{DoF}_{\mathrm{L}}^{*}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) \leq \beta \cdot \min\left\{\frac{K_{\mathrm{T}}\mu_{\mathrm{T}} + K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\} + \bar{\beta} \cdot \min\left\{\frac{1 + K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}. \tag{5.23}$$

It is easily seen that by denoting the right-hand side of (5.23) as $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta)$, we have

$$\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) = \beta \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0). \tag{5.24}$$

The expression in (5.24) proofs useful further on when the upper-bound in Theorem 5.4 is employed to prove the converse parts of Theorem 5.1 and Theorem 5.2.

The rest of this section is dedicated to proving Theorem 5.4. We start with the observation that under average distinct demands, as opposed to worst-case demands, there is a precise characterization for the number of packets to be delivered to the receivers [31]. Since the performance under average demands is no worse than that under worst-case demands, the one-shot linear sum-DoF in (5.14) is bounded above by

$$\mathsf{DoF}^{*}_{\mathrm{L}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) \leq \frac{K_{\mathrm{R}} F(1-\mu_{\mathrm{R}})}{\overline{H}}, \tag{5.25}$$

where $\overline{H}$ is a lower-bound on the delivery time under average demands rather than worst-case demands. Note that the above relaxation is commonly used to obtain upper-bounds in cache-aided setups, e.g. the proof in Section 4.5 and [24, 31, 32]. Next, we follow the same general footsteps of [31, Sec. V] to characterize and then find a lower-bound for $\overline{H}$. The steps borrowed from [31] are explained in less detail, while we elaborate more on the new challenges that arise due to packet splitting over the two subchannels.

5.5.1. Upper-Bound on the Number of Subpackets Reliably Delivered Per Block

First, let us fix the caching realization $\big(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}}, \{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}}\big)$, user demand vector $\mathbf{d}$ and splitting ratio q. As described in Section 5.3.2, in each P-block or N-block, a subset of P-subpackets or N-subpacket are delivered over the P-subchannel or the N-subchannel, respectively. Let $\big\{\mathbf{w}^{(\mathrm{p})}_{n_l,f_l}\big\}_{l=1}^{L^{(\mathrm{p})}}$ be a set of $L^{(\mathrm{p})}$ P-subpackets to be delivered to $L^{(\mathrm{p})}$ distinct receivers over one P-block, and $\big\{\mathbf{w}^{(\mathrm{n})}_{n_l,f_l}\big\}_{l=1}^{L^{(\mathrm{n})}}$ be a set of $L^{(\mathrm{n})}$ N-subpackets to be delivered to $L^{(\mathrm{n})}$ distinct receivers over one N-block. In order for the receivers to successfully decode the transmitted subpackets, $L^{(\mathrm{p})}$ and $L^{(\mathrm{n})}$ must satisfy

$$L^{(\mathrm{p})} \leq \min_{l\in[L^{(\mathrm{p})}]} \big\{|\mathcal{R}_l| + |\mathcal{T}_l|\big\} \tag{5.26}$$

$$L^{(\mathrm{n})} \leq \min_{l\in[L^{(\mathrm{n})}]} |\mathcal{R}_l| + 1 \tag{5.27}$$

where, for any $l \in [L^{(\mathrm{p})}]$ or $l \in [L^{(\mathrm{n})}]$, $\mathcal{T}_l$ and $\mathcal{R}_l$ are the sets of transmitters and receivers, respectively, which store the packet $\mathbf{w}_{n_l,f_l} = \big(\mathbf{w}^{(\mathrm{p})}_{n_l,f_l}, \mathbf{w}^{(\mathrm{n})}_{n_l,f_l}\big)$ in their caches.

The inequality in (5.26) follows directly from [31, Lem. 3]. On the other hand, the inequality in (5.27) can be shown to hold by following the same general steps used to prove [31, Lem. 3], while observing that the generic channel matrices and the lack of CSIT make the zero-forcing conditions in the proof of [31, Lem. 3] impossible to satisfy almost surely. This in turn eliminates the transmitter cooperation gain. A more detailed explanation is given in Appendix C.2.

5.5.2. Integer Program Formulation

For any P-block and N-block indexed by $m^{(\mathrm{p})}$ and $m^{(\mathrm{n})}$ respectively, the sets of subpackets $\mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}$ and $\mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}}$ to be delivered are deemed feasible *only if* their cardinalities satisfy (5.26) and (5.27). Hence by keeping the caching realization, demand vector and splitting ratio fixed, the following integer programming problems yields a lower-bound on the delivery time:

$$
\begin{aligned}
\min \quad & \max\Big\{\frac{q}{\beta}H^{(\mathrm{p})}, \frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})}\Big\} \\
\text{s.t.} \quad & \bigcup_{m^{(\mathrm{p})}=1}^{H^{(\mathrm{p})}} \mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}} = \bigcup_{r=1}^{K_{\mathrm{R}}} \Big(W^{(\mathrm{p})}_{d_r} \setminus \mathcal{U}^{(\mathrm{p})}_r\Big) \\
& \bigcup_{m^{(\mathrm{n})}=1}^{H^{(\mathrm{n})}} \mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}} = \bigcup_{r=1}^{K_{\mathrm{R}}} \Big(W^{(\mathrm{n})}_{d_r} \setminus \mathcal{U}^{(\mathrm{n})}_r\Big) \\
& \mathcal{D}^{(\mathrm{p})}_{m^{(\mathrm{p})}}, \mathcal{D}^{(\mathrm{n})}_{m^{(\mathrm{n})}} \text{ are feasible, } \forall m^{(\mathrm{p})} \in [H^{(\mathrm{p})}],\ \forall m^{(\mathrm{n})} \in [H^{(\mathrm{n})}].
\end{aligned}
\tag{5.28}
$$

The optimal value for the above problem is denoted by $H^{*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}$.

5.5.3. From Worst-Case to Average Demands and Optimizing Over Caching Realizations and Splitting Ratios

Given a caching realization $\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}}, \{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}}\right)$, each file W_n, with $n \in [N]$, is split into $(2^{K_{\mathrm{T}}} - 1)(2^{K_{\mathrm{R}}})$ subfiles $\{W_{n,\mathcal{T},\mathcal{R}}\}_{\mathcal{T}\subseteq_{\emptyset}[K_{\mathrm{T}}],\mathcal{R}\subseteq[K_{\mathrm{R}}]}$, where $W_{n,\mathcal{T},\mathcal{R}}$ denotes the subfile of file W_n cached by transmitters in $\mathcal{T}$ and receivers in $\mathcal{R}$, and $\mathcal{T} \subseteq_{\emptyset} [K_{\mathrm{T}}]$ denotes $\mathcal{T} \subseteq [K_{\mathrm{T}}], \mathcal{T} \neq \emptyset$. Denoting the number of packets in $W_{n,\mathcal{T},\mathcal{R}}$ as $a_{n,\mathcal{T},\mathcal{R}}$, we may write an optimization problem to minimize $H^{*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}$, for the worst-case demands, over all caching realizations and splitting ratios.

As in [31], we further lower-bound the delivery time by considering average demands instead of worst-case demands. In particular, by taking the average over the set of all possible $\pi(N, K_{\mathrm{R}}) = \frac{N!}{(N-K_{\mathrm{R}})!}$ permutations of distinct receiver demands, denote by $\mathcal{P}_{N,K_{\mathrm{R}}}$, we write the problem:

$$
\begin{aligned}
\min_{\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},q} \quad & \frac{1}{\pi(N,K_{\mathrm{R}})} \sum_{\mathbf{d}\in\mathcal{P}_{N,K_{\mathrm{R}}}} H^{*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)} \\
\text{s.t.} \quad & \sum_{\mathcal{T}\subseteq_{\emptyset}[K_{\mathrm{T}}]} \sum_{\mathcal{R}\subseteq[K_{\mathrm{R}}]} a_{n,\mathcal{T},\mathcal{R}} = F,\ \forall n \in [N] \\
& \sum_{n=1}^{N} \sum_{\substack{\mathcal{T}\subseteq[K_{\mathrm{T}}]:\\ i\in\mathcal{T}}} \sum_{\mathcal{R}\subseteq[K_{\mathrm{R}}]} a_{n,\mathcal{T},\mathcal{R}} \leq \mu_{\mathrm{T}} NF,\ \forall i \in [K_{\mathrm{T}}] \\
& \sum_{n=1}^{N} \sum_{\mathcal{T}\subseteq_{\emptyset}[K_{\mathrm{T}}]} \sum_{\substack{\mathcal{R}\subseteq[K_{\mathrm{R}}]:\\ j\in\mathcal{R}}} a_{n,\mathcal{T},\mathcal{R}} \leq \mu_{\mathrm{R}} NF,\ \forall j \in [K_{\mathrm{R}}] \\
& q \in [0,1], a_{n,\mathcal{T},\mathcal{R}} \geq 0, \forall n \in [N], \forall \mathcal{T} \subseteq_{\emptyset} [K_{\mathrm{T}}], \forall \mathcal{R} \subseteq [K_{\mathrm{R}}].
\end{aligned}
\tag{5.29}
$$

The optimum objective for the above problem is denoted by $\overline{H}$, which appears in the bound in (5.25). In what follows, we are interested in further lower bounding $\overline{H}$.

5.5.4. Decoupling the P and N Subchannel and Optimizing Over Caching Realizations

To obtain a lower-bound for $\bar{H}$, we consider optimizing over caching realizations for the P-subchannel and N-subchannel independently. To facilitate this, we start by observing that $H^{*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}$ in (5.29), the optimum objective of (5.28) is bounded below as

$$H^{*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)} \geq \max\left\{\frac{q}{\beta}H^{(\mathrm{p})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)},\frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}\right\} \tag{5.30}$$

where $H^{(\mathrm{s})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}$, $\mathrm{s} \in \{\mathrm{p}, \mathrm{n}\}$, is the optimum objective of the optimization problem

$$\begin{aligned} \min \quad & H^{(\mathrm{s})} \\ \text{s.t.} \quad & \bigcup_{m^{(\mathrm{s})}=1}^{H^{(\mathrm{s})}} \mathcal{D}_{m^{(\mathrm{s})}}^{(\mathrm{s})} = \bigcup_{r=1}^{K_{\mathrm{r}}} \left(W_{d_r}^{(\mathrm{s})} \setminus \mathcal{U}_r^{(\mathrm{s})}\right) \\ & \mathcal{D}_{m^{(\mathrm{s})}}^{(\mathrm{s})} \text{ is feasible, } \forall m^{(\mathrm{s})} \in [H^{(\mathrm{s})}]. \end{aligned} \tag{5.31}$$

The lower-bound in (5.30) is derived directly from problem (5.28), e.g. the P-subchannel term on the right-hand side of (5.30) is obtained by relaxing all N-subchannel components in the objective and constraints of problem (5.28). Denoting the average demand operator $\frac{1}{\pi(N,K_{\mathrm{R}})}\sum_{\mathbf{d}\in\mathcal{P}_{N,K_{\mathrm{R}}}}(\cdot)$ by $\mathbb{E}_{\mathbf{d}}(\cdot)$ for brevity, it follows that the objective function of problem (5.29) is lower bounded as

$$\begin{aligned} \mathbb{E}_{\mathbf{d}}\left(H^{*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}\right) &\geq \mathbb{E}_{\mathbf{d}}\left(\max\left\{\frac{q}{\beta}H^{(\mathrm{p})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)},\frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}\right\}\right) \\ &\geq \max\left\{\frac{q}{\beta}\,\mathbb{E}_{\mathbf{d}}\left(H^{(\mathrm{p})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}\right),\frac{\bar{q}}{\bar{\beta}}\,\mathbb{E}_{\mathbf{d}}\left(H^{(\mathrm{n})*\left(\{\mathcal{P}_i\}_{i=1}^{K_{\mathrm{T}}},\{\mathcal{U}_i\}_{i=1}^{K_{\mathrm{R}}},\mathbf{d},q\right)}\right)\right\} \end{aligned} \tag{5.32}$$

where the inequality in (5.32) follows from the convexity of the pointwise maximum function and Jensen's inequality. Next, we plug the lower-bound in (5.32) into (5.29) from which we obtain a lower-bound on $\overline{H}$. Moreover, for any given splitting ratio q, we optimize over caching realizations independently for the P-subchannel and N-subchannel through

$$\begin{aligned}
\min_{\{\mathcal{P}_i\}_{i=1}^{K_\mathrm{T}},\{\mathcal{U}_i\}_{i=1}^{K_\mathrm{R}}} \quad & \frac{1}{\pi(N,K_\mathrm{R})}\sum_{\mathbf{d}\in\mathcal{P}_{N,K_\mathrm{R}}} H^{(\mathrm{s})*\left(\{\mathcal{P}_i\}_{i=1}^{K_\mathrm{T}},\{\mathcal{U}_i\}_{i=1}^{K_\mathrm{R}},\mathbf{d},q\right)} \\
\text{s.t.} \quad & \sum_{\mathcal{T}\subseteq_\emptyset[K_\mathrm{T}]}\sum_{\mathcal{R}\subseteq[K_\mathrm{R}]} a_{n,\mathcal{T},\mathcal{R}} = F,\ \forall n\in[N] \\
& \sum_{n=1}^{N}\sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ i\in\mathcal{T}}}\sum_{\mathcal{R}\subseteq[K_\mathrm{R}]} a_{n,\mathcal{T},\mathcal{R}} \le \mu_\mathrm{T}NF,\ \forall i\in[K_\mathrm{T}] \\
& \sum_{n=1}^{N}\sum_{\mathcal{T}\subseteq_\emptyset[K_\mathrm{T}]}\sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ j\in\mathcal{R}}} a_{n,\mathcal{T},\mathcal{R}} \le \mu_\mathrm{R}NF,\ \forall j\in[K_\mathrm{R}] \\
& a_{n,\mathcal{T},\mathcal{R}}\ge 0, \forall n\in[N], \forall\mathcal{T}\subseteq_\emptyset[K_\mathrm{T}], \forall\mathcal{R}\subseteq[K_\mathrm{R}],
\end{aligned} \tag{5.33}$$

for which we denote the optimum objective function as $\overline{H^{(\mathrm{s})}}^{(q)}$, $\mathrm{s}\in\{\mathrm{p},\mathrm{n}\}$. This yields the lower-bound on $\overline{H}$ given by

$$\overline{H} \ge \min_{q\in[0,1]} \max\left\{\frac{q}{\beta}\overline{H^{(\mathrm{p})}}^{(q)}, \frac{\bar{q}}{\bar{\beta}}\overline{H^{(\mathrm{n})}}^{(q)}\right\}. \tag{5.34}$$

The two components $\overline{H^{(\mathrm{p})}}^{(q)}$ and $\overline{H^{(\mathrm{n})}}^{(q)}$ can be separately lower bounded as

$$\overline{H^{(\mathrm{p})}}^{(q)} \ge \frac{K_\mathrm{R}F(1-\mu_\mathrm{R})^2}{K_\mathrm{T}\mu_\mathrm{T}+K_\mathrm{R}\mu_\mathrm{R}} \tag{5.35}$$

$$\overline{H^{(\mathrm{n})}}^{(q)} \ge \frac{K_\mathrm{R}F(1-\mu_\mathrm{R})^2}{1+K_\mathrm{R}\mu_\mathrm{R}}. \tag{5.36}$$

The lower-bound in (5.35) follows directly from [31, Lem. 4]. On the other hand, the lower-bound in (5.36) is derived in Appendix C.1 by employing the same techniques in the proof of [31].

Since in the problem in (5.33) the total number of subpackets per block delivered over either of the two subchannels is $K_\mathrm{R}F\,(1-\mu_\mathrm{R})$, and no more than K_R subpackets can be delivered simultaneously, we obtain $\overline{H^{(\mathrm{s})}}^{(q)} \ge \frac{K_\mathrm{R}F(1-\mu_\mathrm{R})}{K_\mathrm{R}}$. Combining this with the lower-bounds in (5.35) and (5.36), we obtain

$$\overline{H^{(\mathrm{p})}}^{(q)} \ge \frac{K_\mathrm{R}F(1-\mu_\mathrm{R})}{\min\left\{\frac{K_\mathrm{T}\mu_\mathrm{T}+K_\mathrm{R}\mu_\mathrm{R}}{1-\mu_\mathrm{R}}, K_\mathrm{R}\right\}} \tag{5.37}$$

$$\overline{H^{(\mathrm{n})}}^{(q)} \ge \frac{K_\mathrm{R}F(1-\mu_\mathrm{R})}{\min\left\{\frac{1+K_\mathrm{R}\mu_\mathrm{R}}{1-\mu_\mathrm{R}}, K_\mathrm{R}\right\}}. \tag{5.38}$$

It is evident that the above lower-bounds do not depend on the value of q, and by combining (5.37) and (5.38) with (5.34), it follows that

$$\overline{H} \ge \min_{q\in[0,1]} \max\left\{\frac{q}{\beta}\cdot\frac{K_\mathrm{R}F(1-\mu_\mathrm{R})}{\min\left\{\frac{K_\mathrm{T}\mu_\mathrm{T}+K_\mathrm{R}\mu_\mathrm{R}}{1-\mu_\mathrm{R}}, K_\mathrm{R}\right\}}, \frac{\bar{q}}{\bar{\beta}}\cdot\frac{K_\mathrm{R}F(1-\mu_\mathrm{R})}{\min\left\{\frac{1+K_\mathrm{R}\mu_\mathrm{R}}{1-\mu_\mathrm{R}}, K_\mathrm{R}\right\}}\right\}. \tag{5.39}$$

5.5.5. Optimizing Over Splitting Rations and Combing Bounds

The splitting ration q that minimizes the right-hand side of (5.39), which we denote by q^*, must satisfy

$$\frac{q^*}{\beta} \cdot \frac{K_{\mathrm{R}} F(1-\mu_{\mathrm{R}})}{\min\left\{\frac{K_{\mathrm{T}}\mu_{\mathrm{T}}+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}} = \frac{\bar{q^*}}{\bar{\beta}} \cdot \frac{K_{\mathrm{R}} F(1-\mu_{\mathrm{R}})}{\min\left\{\frac{1+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}},$$

as any other q leads to a larger value for the right-hand side of (5.39). By considering q^*, we obtain[4]

$$\bar{H} \geq \frac{K_{\mathrm{R}} F(1-\mu_{\mathrm{R}})}{\beta \cdot \min\left\{\frac{K_{\mathrm{T}}\mu_{\mathrm{T}}+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\} + \bar{\beta} \cdot \min\left\{\frac{1+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}}. \tag{5.40}$$

Combining the lower-bound in (5.40) with the upper-bound in (5.25), we obtain

$$\mathsf{DoF}^*_{\mathrm{L}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) \leq \beta \cdot \min\left\{\frac{K_{\mathrm{T}}\mu_{\mathrm{T}}+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\} + \bar{\beta} \cdot \min\left\{\frac{1+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}$$

which concludes the proof of Theorem 5.4.

5.6. Centralized Setting: Proof of Theorem 5.1

Equipped with the upper-bound in Theorem 5.4, we are now ready to prove the main results of the chapter. We start with Theorem 5.1 in this section and proceed to Theorem 5.2 in the following section.

5.6.1. Achievability of Theorem 5.1

Placement Phase

The placement phase is analogous to the one in [31]. Interestingly, as in Section 4.6, this implies that the placement phase is not required to depend on the value of β. As in [31], each file W_n, $n \in [N]$, is partitioned into $\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}\binom{K_{\mathrm{R}}}{K_{\mathrm{R}}\mu_{\mathrm{R}}}$ disjoint subfiles of equal size, denoted by

$$W_n = \{W_{n,\mathcal{T},\mathcal{R}}\}_{\substack{\mathcal{T}\subseteq[K_{\mathrm{T}}]:|\mathcal{T}|=K_{\mathrm{T}}\mu_{\mathrm{T}} \\ \mathcal{R}\subseteq[K_{\mathrm{R}}]:|\mathcal{R}|=K_{\mathrm{R}}\mu_{\mathrm{R}}}}.$$

Note that each subfile contains $\frac{F}{\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}\binom{K_{\mathrm{R}}}{K_{\mathrm{R}}\mu_{\mathrm{R}}}}$ packets. Each transmitter Tx_i stores subfiles given by $\mathcal{P}_i = \{W_{n,\mathcal{T},\mathcal{R}} : i \in \mathcal{T}\}$, while each receiver Rx_j stores subfiles given by $\mathcal{U}_j = \{W_{n,\mathcal{T},\mathcal{R}} : j \in \mathcal{R}\}$. It is easy to verify that such placement strategy satisfies the memory size constraints at both transmitters and receivers, and that each receiver caches $\mu_{\mathrm{R}} F$ packets from each file.

[4] For any real numbers x, y and q such that $\frac{q}{x} = \frac{1-q}{y}$, it is easy to verify that $\frac{q}{x} = \frac{1}{x+y}$.

Delivery Phase

During the delivery phase, each receiver Rx_j requests for a file W_{d_j}. As Rx_j has all the subfiles $W_{d_j,\mathcal{T},\mathcal{R}}$ with $j \in \mathcal{R}$ cached in its memory, it only requires the remaining subfiles given by $W_{d_j,\mathcal{T},\mathcal{R}}$ with $j \notin \mathcal{R}$. As shown in Section 5.3.2, each packet $\mathbf{w}_{d_j,f}$ to be delivered is split into two subpackets, i.e. $\mathbf{w}_{d_j,f} = \left(\mathbf{w}^{(\mathrm{p})}_{d_j,f}, \mathbf{w}^{(\mathrm{n})}_{d_j,n}\right)$. We refer to the set of P-subpackets of $W_{d_j,\mathcal{T},\mathcal{R}}$ as the P-subfile $W^{(\mathrm{p})}_{d_j,\mathcal{T},\mathcal{R}}$, and the set of N-packets of $W_{d_j,\mathcal{T},\mathcal{R}}$ as the N-subfile $W^{(\mathrm{n})}_{d_j,\mathcal{T},\mathcal{R}}$. The P-subfiles are delivered over the P-subchannel using the linear scheme in [31]. On the other hand, the N-subfiles are delivered over the N-subchannel using the original coded-multicasting scheme in [24], with the difference that superposition of coded N-subpackets over the air is used instead of XoR operations before encoding, as the latter is infeasible due to the distributed nature of transmitters. Decoding of subpackets at the receivers is carried out after taking the appropriate linear combinations, e.g. see (5.11) and (5.12). Each Rx_j retrieves all missing P-subfiles and N-subfile and hence the file W_{d_j} is recovered.

Achievable One-Shot Linear sum-DoF

Since each user has $\mu_{\mathrm{R}}F$ packets from each file stored in its cache memory, a total of $K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})$ packets are delivered during the delivery phase, split into $K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})$ P-subpackets and $K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})$ N-subpackets delivered over the P-subchannel and N-subchannel, respectively. In what follows, we denote $K_{\mathrm{R}}\mu_{\mathrm{R}}$ and $K_{\mathrm{T}}\mu_{\mathrm{T}}$ by $m_{\mathrm{C,R}}$ and $m_{\mathrm{C,T}}$ respectively. From [31], we know that $\min\{m_{\mathrm{C,T}} + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}$ P-subpackets are delivered in each P-block, and hence

$$H^{(\mathrm{p})}_{\mathrm{C}} = \frac{K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})}{\min\{m_{\mathrm{C,T}} + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}}.$$

On the other, we know from [24] that $\min\{1 + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}$ N-subpackets are delivered in each N-block. Therefore, we obtain

$$H^{(\mathrm{n})}_{\mathrm{C}} = \frac{K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})}{\min\{1 + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}}.$$

It follows that the delivery time in packet-time-slot is given by $H_{\mathrm{C}} = \max\left\{\frac{q}{\beta}H^{(\mathrm{p})}_{\mathrm{C}}, \frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})}_{\mathrm{C}}\right\}$. Next, we choose the splitting ratio q as follow:

$$q = \frac{\beta \cdot \min\{m_{\mathrm{C,T}} + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}}{\beta \cdot \min\{m_{\mathrm{C,T}} + m_{\mathrm{C,R}}, K_{\mathrm{R}}\} + \bar{\beta} \cdot \min\{1 + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}}.$$

It can be verified that the above splitting ratio satisfies $\frac{q}{\beta}H^{(\mathrm{p})}_{\mathrm{C}} = \frac{\bar{q}}{\bar{\beta}}H^{(\mathrm{n})}_{\mathrm{C}}$. This value of q minimizes the duration of the communication which in turn maximizes the achievable sum-DoF. Note that q increases with β, due to the fact that a larger β implies that the P-subchannel occupies a larger fraction of the bandwidth, hence carrying larger portions of each packet. As one may anticipate, we obtain $q = 0$ and $q = 1$ at the two extremes $\beta = 0$ and $\beta = 1$, respectively. With such value of

q we obtain

$$H_{\mathrm{C}} = \frac{K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})}{\beta \cdot \min\{m_{\mathrm{C,T}} + m_{\mathrm{C,R}}, K_{\mathrm{R}}\} + \bar{\beta} \cdot \min\{1 + m_{\mathrm{C,R}}, K_{\mathrm{R}}\}}. \tag{5.41}$$

From (5.41) and the fact that a total of $K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})$ packets are delivered during the delivery phase, the result in (5.15) directly follows. This concludes the proof of achievability.

5.6.2. Converse of Theorem 5.1

From [31], we know that for $\beta = 1$, we have $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)/\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) \leq 2$. We show that when $\beta = 0$, we also have $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)/\mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \leq 2$. Consider the two cases:

1. $\mu_{\mathrm{R}} \leq \frac{1}{2}$: In this case, from (5.23) in Theorem 5.4 we obtain

$$\begin{aligned} \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) &= \min\left\{\frac{1 + K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\} \\ &\leq \min\left\{\frac{1 + K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-1/2}, K_{\mathrm{R}}\right\} \\ &\leq 2 \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0). \end{aligned}$$

2. $\mu_{\mathrm{R}} > \frac{1}{2}$: In this case, the achievability part implies that

$$\begin{aligned} \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) &= \min\{1 + K_{\mathrm{R}}\mu_{\mathrm{R}}, K_{\mathrm{R}}\} \\ &> \min\{1 + K_{\mathrm{R}}/2, K_{\mathrm{R}}\} \\ &> \frac{K_{\mathrm{R}}}{2}. \end{aligned}$$

   Since $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \leq K_{\mathrm{R}}$, we obtain $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \leq 2 \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)$.

Now we extend the above to any $\beta \in [0, 1]$. From the two above constant factor inequalities for $\beta = 1$ and $\beta = 0$, and the decomposition of the lower-bound and the upper-bound in (5.17) and (5.24), we obtain

$$\begin{aligned} \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) &= \beta \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \\ &\leq 2\beta \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + 2\bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \\ &= 2 \cdot \mathsf{DoF}_{\mathrm{L,C}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta). \end{aligned}$$

This completes the proof of Theorem 5.1.

5.7. Decentralized Setting: Proof of Theorem 5.2

In this section, we present a proof of Theorem 5.2 starting with the achievability and then the converse.

5.7.1. Achievability of Theorem 5.2

Placement Phase

As in the centralized setting, the placement phase does not depend on β. Each file $W_n, n \in [N]$, is partitioned into $\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}$ disjoint subfiles of equal size, denoted by $W_n = \{W_{n,\mathcal{T}}\}_{\mathcal{T}\subseteq[K_{\mathrm{T}}]:|\mathcal{T}|=K_{\mathrm{T}}\mu_{\mathrm{T}}}$, where each subfile contains $\frac{F}{\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}}$ packets. Each transmitter Tx$_i$ then stores subfile given by $\mathcal{P}_i = \{W_{n,\mathcal{T}} : i \in \mathcal{T}\}$. On the other end, placement at the receivers is done in a decentralized manner similar to Section 4.7 and [79]. In particular, each receiver Rx$_i$ stores $\mu_{\mathrm{R}}F$ packets from each file, chosen uniformly at random. Therefore, each packet of each file is stored in some subset of users $\tilde{\mathcal{R}} \subseteq [K_{\mathrm{R}}]$, where $|\tilde{\mathcal{R}}| \in \{0, 1, \ldots, K_{\mathrm{R}}\}$. For any $n \in [N]$, we use $W_{n,\mathcal{T},\tilde{\mathcal{R}}}$ to denote the packets of file W_n which are stored by transmitters in $\mathcal{T}$ and receivers in $\tilde{\mathcal{R}}$, where $W_{n,\mathcal{T},\tilde{\mathcal{R}}}$ is referred to as a mini-subfile henceforth. It follows that W_n can be reconstructed from $\big\{W_{n,\mathcal{T},\tilde{\mathcal{R}}} : \mathcal{T} \subseteq [K_{\mathrm{T}}], |\mathcal{T}| = K_{\mathrm{T}}\mu_{\mathrm{T}}, \tilde{\mathcal{R}} \subseteq [K_{\mathrm{R}}]\big\}$.

Delivery Phase

Each receiver Rx$_j$ requests for a file W_{d_j}, hence the transmitters have to deliver all mini-subfiles $W_{d_j,\mathcal{T},\tilde{\mathcal{R}}}$ with $j \notin \tilde{\mathcal{R}}$. Each packet to be delivered is split as in the centralized case, and we use $W^{(\mathrm{p})}_{d_j,\mathcal{T}}$ (P-subfile) and $W^{(\mathrm{n})}_{d_j,\mathcal{T}}$ (N-subfile) to denote the sets of P-subpackets and N-subpackets of $W_{d_j,\mathcal{T}}$, respectively. Similarly, we use $W^{(\mathrm{p})}_{d_j,\mathcal{T},\tilde{\mathcal{R}}}$ (P-mini-subfile) and $W^{(\mathrm{n})}_{d_j,\mathcal{T},\tilde{\mathcal{R}}}$ (N-mini-subfile) to denote the sets of P-subpackets and N-subpackets of $W_{d_j,\mathcal{T},\tilde{\mathcal{R}}}$, respectively.

The P-mini-subfiles are delivered over the P-subchannel, where the delivery takes place over K_{R} sub-phases indexed by $l \in \{0, 1, \ldots, K_{\mathrm{R}} - 1\}$. In the l-th sub-phase, the transmitters delivers all $W^{(\mathrm{p})}_{d_j,\mathcal{T},\tilde{\mathcal{R}}}$ with $|\tilde{\mathcal{R}}| = l$. Note that l goes up to $K_{\mathrm{R}} - 1$ since for $|\tilde{\mathcal{R}}| = K_{\mathrm{R}}$, the corresponding P-mini-subfiles are cached by all receivers. For each sub-phase l, the delivery in the P-subchannel is reminiscent of the centralized P-subchannel delivery in Section 5.6.1, with the difference that $m_{\mathrm{C,R}}$ in the centralized setting is replaced with l here (i.e. smaller multicasting gain), as this sub-phase considers subfiles which are cached by exactly l users. It follows that $\min\{m_{\mathrm{C,T}} + l, K_{\mathrm{R}}\}$ P-subpackets are transmitted simultaneously.

On the other hand, the N-mini-subfiles are delivered over the N-subchannel using the original decentralized coded-multicasting scheme in [79], while using over the air superposition instead of XoR. Each receiver then obtains all missing mini-subfiles and recovers the demanded file.

Achievable One-Shot Linear sum-DoF

We start be focusing on the delivery time over the P-subchannel. Consider the l-th sub-phase and an arbitrary subset of users $\tilde{\mathcal{R}}$ with size l. For each P-subfile $W^{(\mathrm{p})}_{n,\mathcal{T}}$, $n \in [N]$, stored by some subset $\mathcal{T}$ of users), the probability that any of its P-subpackets is stored by any of the users in $\tilde{\mathcal{R}}$ is given by μ_{R}, as each such user caches $\mu_{\mathrm{R}}F$ random P-subpackets from each file. Hence, the probability that a P-subpacket is stored by exactly the l users of $\tilde{\mathcal{R}}$ is given by $\mu_{\mathrm{R}}^l(1 - \mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}$. It follows that the expected number of P-subpackets of $W^{(\mathrm{p})}_{n,\mathcal{T}}$ stored by each user in $\tilde{\mathcal{R}}$ is given by

$\frac{\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F}{\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}} + o(F)$ when $F \to \infty$. The term $o(F)$ is omitted henceforth. As there is a total of $\binom{K_{\mathrm{R}}}{l}$ subsets of l users, there is a total of $\frac{\binom{K_{\mathrm{R}}}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F}{\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}}$ P-subpackets of $W_{n,\mathcal{T}}^{(\mathrm{p})}$ which are cached by exactly l users. We now proceed to calculate number of P-subpackets of $W_{d_j}^{(\mathrm{p})}$ stored by exactly l users and have to be delivered to receiver Rx_j. For each $\mathcal{T}$, receiver Rx_j has all P-mini-subfiles $W_{d_j,\mathcal{T},\tilde{\mathcal{R}}}^{(\mathrm{p})}$, with $|\tilde{\mathcal{R}}| = l$ and $j \in \tilde{\mathcal{R}}$, cached in its memory. Hence, Rx_j already has $\frac{\binom{K_{\mathrm{R}}-1}{l-1}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F}{\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}}$ P-subpackets of $W_{d_j,\mathcal{T}}^{(\mathrm{p})}$ which are cached by exactly l users. It follows that the number of P-subpackets of $W_{d_j,\mathcal{T}}^{(\mathrm{p})}$ unavailable at Rx_j, given by all P-mini-subfiles $W_{d_j,\mathcal{T},\tilde{\mathcal{R}}}^{(\mathrm{p})}$ with $|\tilde{\mathcal{R}}| = l$ and $j \notin \tilde{\mathcal{R}}$, is equal to $\frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F}{\binom{K_{\mathrm{T}}}{K_{\mathrm{T}}\mu_{\mathrm{T}}}}$. Considering all possible P-subfiles $W_{d_j,\mathcal{T}}^{(\mathrm{p})}$ for all $\mathcal{T}$, and as there are K_{R} receivers in total, the total number of P-subpackets which are stored by exactly l users and have to be delivered to all receivers in the l-th delivery sub-phase is given by

$$K_{\mathrm{R}}\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F.$$

We recall that in the l-th delivery sub-phase, a total of $\min\{m_{\mathrm{C,T}}+l, K_{\mathrm{R}}\}$ P-subpackets are delivered simultaneously over the P-subchannel. By summing over all K_{R} sub-phases, we obtain

$$H_{\mathrm{D}}^{(\mathrm{p})} = K_{\mathrm{R}}\sum_{l=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F}{\min\{m_{\mathrm{C,T}}+l, K_{\mathrm{R}}\}}.$$

Moving on to the N-subchannel, as the delivery of the N-mini-subfiles follows the coded-multicasting scheme of [79], it follows that

$$H_{\mathrm{D}}^{(\mathrm{n})} = K_{\mathrm{R}}\sum_{l=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l}F}{1+l} = \frac{1-\mu_{\mathrm{R}}}{\mu_{\mathrm{R}}}\left(1-(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}}\right)F.$$

From the above, it follows that the delivery time is given by $H_{\mathrm{D}} = \max\left\{\frac{q}{\beta}H_{\mathrm{D}}^{(\mathrm{p})}, \frac{\bar{q}}{\bar{\beta}}H_{\mathrm{D}}^{(\mathrm{n})}\right\}$ packet-time-slots. As for the centralized case, we choose q such that $\frac{q}{\beta}H_{\mathrm{D}}^{(\mathrm{p})} = \frac{\bar{q}}{\bar{\beta}}H_{\mathrm{D}}^{(\mathrm{n})}$, which in turn minimizes the duration of the communication and hence maximizes the achievable sum-DoF. Hence, we choose

$$q = \frac{\beta\cdot\frac{1}{\sum_{l=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l-1}}{\min\{K_{\mathrm{R}},K_{\mathrm{T}}\mu_{\mathrm{T}}+l\}}}}{\beta\cdot\frac{1}{\sum_{l=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l-1}}{\min\{K_{\mathrm{R}},K_{\mathrm{T}}\mu_{\mathrm{T}}+l\}}} + \bar{\beta}\cdot\frac{K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}}}}.$$

From the above choice of q and the values of $H_{\mathrm{D}}^{(\mathrm{p})}$ and $H_{\mathrm{C}}^{(\mathrm{p})}$, it follows that

$$H_{\mathrm{D}} = \frac{K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})}{\beta\cdot\frac{1}{\sum_{l=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{l}\mu_{\mathrm{R}}^{l}(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-l-1}}{\min\{m_{\mathrm{C,T}}+l,K_{\mathrm{R}}\}}} + \bar{\beta}\frac{K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}}}}. \tag{5.42}$$

As a total of $K_{\mathrm{R}}F(1-\mu_{\mathrm{R}})$ packets are delivered during the delivery phase, the result in (5.18) directly follows from (5.42), which concludes the proof of achievability.

5.7.2. Converse of Theorem 5.2

In this part, we prove (5.19) through the following steps:

- The first step of the proof is to show that when $\beta = 0$, we have the constant factor

$$\frac{\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)}{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)} \leq 3. \tag{5.43}$$

- The following step is to show that the one-shot linear DoF ratio in (5.43), with $\beta = 0$, is an upper-bound for the ratio with $\beta = 1$, i.e.

$$\frac{\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)}{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1)} \leq \frac{\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)}{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)}. \tag{5.44}$$

- Equipped with (5.43) and (5.44), we proceed ad follows:

$$\begin{aligned}\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta) &= \beta \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + \bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)\\ &\leq 3\beta \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 1) + 3\bar{\beta} \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)\\ &= 3 \cdot \mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \beta).\end{aligned}$$

It can be seen that the last of the three above steps concludes the proof of Theorem 5.2. Therefore, the remainder of this part is dedicated to proving the inequalities in (5.43) and (5.44).

Proof of (5.43)

First, we recall that $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) = \frac{K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}}}$. Combining this with $(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}} \geq 0$ and the Bernoulli inequality $(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}} \geq 1 - K_{\mathrm{R}}\mu_{\mathrm{R}}$, we obtain

$$\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \geq \max\Big\{K_{\mathrm{R}}\mu_{\mathrm{R}}, 1\Big\}. \tag{5.45}$$

For the trivial case of $K_{\mathrm{R}} = 1$, it is easy to see that $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) = \mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) = 1$. For the case of $K_{\mathrm{R}} = 2$, we have $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \geq 1$ from (5.45) and $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \leq 2$ from (5.23) in Theorem 5.4. Hence for this case, (5.43) holds. Similarly, for the case $K_{\mathrm{R}} = 3$, we have $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \geq 1$ and $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \leq 3$ from which (5.43) also holds. Therefore, without loss of generality, we assume that $K_{\mathrm{R}} \geq 4$ henceforth. We proceed by considering the following cases:

1. $\mu_{\mathrm{R}} \leq 1/K_{\mathrm{R}}$: For this case we have

$$\begin{aligned}\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) &= \min\Big\{\frac{K_{\mathrm{R}}\mu_{\mathrm{R}}+1}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\Big\} \\ &\leq \min\Big\{\frac{1+1}{1-1/K_{\mathrm{R}}}, K_{\mathrm{R}}\Big\} \\ &\leq \min\Big\{\frac{8}{3}, K_{\mathrm{R}}\Big\} \leq 3.\end{aligned}$$

Combining the above with $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \geq 1$, we conclude that (5.43) holds.

2. $\mu_{\mathrm{R}} \in (1/K_{\mathrm{R}}, 2/K_{\mathrm{R}}]$: For this case, we start by defining the function

$$f(\mu_{\mathrm{R}}) = 3\mu_{\mathrm{R}} + \frac{1}{K_{\mathrm{R}}\mu_{\mathrm{R}}}.$$

The function $f(\mu_{\mathrm{R}})$ is convex in $[0, \infty)$, and hence $f(\mu_{\mathrm{R}}) \leq \max\big(f(\frac{1}{K_{\mathrm{R}}}), f(\frac{2}{K_{\mathrm{R}}})\big)$ over the interval of interest $\mu_{\mathrm{R}} \in (1/K_{\mathrm{R}}, 2/K_{\mathrm{R}}]$. Moreover, it is easy to verify that $f(\frac{1}{K_{\mathrm{R}}}) = \frac{3}{K_{\mathrm{R}}} + 1 \leq \frac{7}{4}$ and $f(\frac{2}{K_{\mathrm{R}}}) = \frac{6}{K_{\mathrm{R}}} + \frac{1}{2} \leq 2$. Therefore, $f(\mu_{\mathrm{R}}) = 3\mu_{\mathrm{R}} + \frac{1}{K_{\mathrm{R}}\mu_{\mathrm{R}}} \leq 2$ for all K_{R} and μ_{R} of interest. Combining this with (5.45) and (5.23), we obtain

$$\begin{aligned}\frac{\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)}{\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0)} &\leq \min\Big\{\frac{K_{\mathrm{R}}\mu_{\mathrm{R}}+1}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\Big\} \cdot \frac{1}{\max\{K_{\mathrm{R}}\mu_{\mathrm{R}}, 1\}} \\ &\leq \Big(1 + \frac{1}{K_{\mathrm{R}}\mu_{\mathrm{R}}}\Big) \cdot \frac{1}{1-\mu_{\mathrm{R}}} \leq 3\end{aligned}$$

where the last inequality is equivalent to $3\mu_{\mathrm{R}} + \frac{1}{K_{\mathrm{R}}\mu_{\mathrm{R}}} \leq 2$. Therefore, (5.43) holds in this case.

3. $\mu_{\mathrm{R}} \in (2/K_{\mathrm{R}}, 1/2]$: For this case we have

$$\begin{aligned}\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) &= \min\Big\{\frac{K_{\mathrm{R}}\mu_{\mathrm{R}}+1}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\Big\} \\ &\leq \min\Big\{\frac{K_{\mathrm{R}}\mu_{\mathrm{R}}+1}{1-1/2}, K_{\mathrm{R}}\Big\} \\ &= \min\Big\{2K_{\mathrm{R}}\mu_{\mathrm{R}}+2, K_{\mathrm{R}}\Big\} \\ &\leq \min\Big\{3K_{\mathrm{R}}\mu_{\mathrm{R}}, K_{\mathrm{R}}\Big\}.\end{aligned}$$

Combining the above with $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \geq K_{\mathrm{R}}\mu_{\mathrm{R}}$, it follows that (5.43) holds.

4. $\mu_{\mathrm{R}} > 1/2$: For this last case we have $\mathsf{DoF}_{\mathrm{L,D}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \geq \max\{K_{\mathrm{R}}\mu_{\mathrm{R}}, 1\} > K_{\mathrm{R}}/2$. Combining this with $\mathsf{DoF}_{\mathrm{L,ub}}(\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, 0) \leq K_{\mathrm{R}}$, it follows that (5.43) holds, hence concluding the proof.

Proof of (5.44)

From (5.18) and (5.23), the inequality in (5.44) can be expressed as

$$\frac{\min\left\{\frac{1+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}}{\left(\sum_{m=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{m}\mu_{\mathrm{R}}^m(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-1-m}}{1+m}\right)^{-1}} \geq \frac{\min\left\{\frac{K_{\mathrm{T}}\mu_{\mathrm{T}}+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}}{\left(\sum_{m=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{m}\mu_{\mathrm{R}}^m(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-1-m}}{\min\{K_{\mathrm{T}}\mu_{\mathrm{T}}+m, K_{\mathrm{R}}\}}\right)^{-1}}. \tag{5.46}$$

Defining the function $J(r)$ as

$$J(r) = \frac{\min\left\{\frac{r+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}, K_{\mathrm{R}}\right\}}{\left(\sum_{m=0}^{K_{\mathrm{R}}-1}\frac{\binom{K_{\mathrm{R}}-1}{m}\mu_{\mathrm{R}}^m(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-1-m}}{\min\{r+m, K_{\mathrm{R}}\}}\right)^{-1}} \tag{5.47}$$

it can be seen that (5.46) is equivalent to $J(1) \geq J(K_{\mathrm{T}}\mu_{\mathrm{T}})$. In the following, we show that that $J(1) \geq J(r)$ for all $r \geq 1$. As a consequence, $J(1) \geq J(r)$ will also hold for integer values of r, hence for any $K_{\mathrm{T}}\mu_{\mathrm{T}}$ which is assumed to be integer for the decentralized setting and hence in Theorem 5.2 and in (5.46).

It is readily seen that for $r \geq K_{\mathrm{R}}(1-2\mu_{\mathrm{R}})$, the numerator in (5.47) becomes K_{R}, and the function $J(r)$ decrease with r. Therefore, without loss of generality, we only consider the interval $r \in [1, K_{\mathrm{R}}(1-2\mu_{\mathrm{R}})]$ in what follows. Equivalently, for any K_{R} and r, we consider values of μ_{R} that satisfy $\mu_{\mathrm{R}} \leq \frac{1}{2}\left(1-\frac{r}{K_{\mathrm{R}}}\right)$.

Next, the inequality in (5.46) is equivalently rewritten as

$$\frac{1+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}\sum_{m=0}^{K_{\mathrm{R}}-1}\binom{K_{\mathrm{R}}-1}{m}\frac{\mu_{\mathrm{R}}^m(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-1-m}}{1+m} \geq \frac{r+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}\sum_{m=0}^{K_{\mathrm{R}}-1}\binom{K_{\mathrm{R}}-1}{m}\frac{\mu_{\mathrm{R}}^m(1-\mu_{\mathrm{R}})^{K_{\mathrm{R}}-1-m}}{\min\{r+m, K_{\mathrm{R}}\}}.$$

After rearranging the terms and removing redundant factors, the above is expressed as

$$\sum_{m=0}^{K_{\mathrm{R}}-1}\frac{1+K_{\mathrm{R}}\mu_{\mathrm{R}}}{1+m}\binom{K_{\mathrm{R}}-1}{m}\left(\frac{\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}\right)^m \geq \sum_{m=0}^{K_{\mathrm{R}}-1}\frac{r+K_{\mathrm{R}}\mu_{\mathrm{R}}}{\min\{r+m, K_{\mathrm{R}}\}}\binom{K_{\mathrm{R}}-1}{m}\left(\frac{\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}\right)^m,$$

which is further rewritten as

$$\sum_{m=0}^{K_{\mathrm{R}}-1}\frac{\zeta(K_{\mathrm{R}}+1)+1}{1+m}\binom{K_{\mathrm{R}}-1}{m}\zeta^m \geq \sum_{m=0}^{K_{\mathrm{R}}-1}\frac{\zeta(K_{\mathrm{R}}+r)+r}{\min\{r+m, K_{\mathrm{R}}\}}\binom{K_{\mathrm{R}}-1}{m}\zeta^m, \tag{5.48}$$

where $\zeta = \frac{\mu_{\mathrm{R}}}{1-\mu_{\mathrm{R}}}$, which is constrained as $\zeta \in \left[0, \frac{K_{\mathrm{R}}-r}{K_{\mathrm{R}}+r}\right]$ for given K_{R} and r. After further rearrangement of terms, the inequality in (5.48) is rewritten as

$$p(\zeta) \triangleq \sum_{m=0}^{K_{\mathrm{R}}} c_m \cdot \zeta^m \geq 0, \tag{5.49}$$

where $p(\zeta)$ is a polynomial in the variable ζ with coefficients given by

$$c_m = \begin{cases} 0, \quad m=0 \\ \frac{1-r}{K_\mathrm{R}}, \quad m=K_\mathrm{R} \\ \binom{K_\mathrm{R}-1}{m-1}\cdot\left(\frac{K_\mathrm{R}+1}{m}-\frac{K_\mathrm{R}+r}{\min\{r+m-1,K_\mathrm{R}\}}\right)+\binom{K_\mathrm{R}-1}{m}\cdot\left(\frac{1}{m+1}-\frac{r}{\min\{r+m,K_\mathrm{R}\}}\right), \quad m\in[1,K_\mathrm{R}-1]_\mathbb{Z}. \end{cases}$$

Note that in the above, we use $[a,b]_\mathbb{Z}$ to denote the set of all integers that are in the interval $[a,b]$, i.e. $[a,b]_\mathbb{Z}\triangleq[a,b]\cap\mathbb{Z}$. At this point, it is clear that the problem reduces to showing that $p(\zeta)\geq 0$ for $\zeta\in\left[0,\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right]$. To this end, we derive the following property of $p(\zeta)$.

Lemma 5.1. *The polynomial $p(\zeta)$ is quasiconcave and hence satisfies the following inequality:*

$$p(\zeta)\geq\min\left(p(0),p\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)\right),\ \forall\zeta\in\left[0,\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right]. \tag{5.50}$$

The proof of (5.50) is rather involved and hence is deferred to Appendix C.3. From Lemma 5.1, it follows that to prove that the inequality in (5.49) holds, it is sufficient to show that $p(0)\geq 0$ and $p\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)\geq 0$. Note that the case with $\zeta=0$ is trivial as $p(0)=0$. Hence, it remains to show that $p\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)\geq 0$ holds true. For this, we require the following inequality.

Lemma 5.2. *[138]. For any positive integer $K\in\mathbb{Z}_+$ and real number $r\in[1,K]$, we have*

$$\sum_{m=1}^{K}\frac{m}{\min\{r+m-1,K\}}\binom{K}{m}\left(\frac{K-r}{K+r}\right)^m\leq\frac{K-r+2}{K+r}\left[\left(\frac{2K}{K+r}\right)^K-1\right]. \tag{5.51}$$

The final step of the proof is to show that the inequality $p\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)\geq 0$ is an instance of Lemma 5.2, and hence holds true. Equivalently, we consider (5.48). By plugging $\zeta=\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}$ into (5.48) and multiplying both sides by $\frac{K_\mathrm{R}+r}{K_\mathrm{R}}$, the inequality $p\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)\geq 0$ is equivalently expressed as

$$\sum_{m=0}^{K_\mathrm{R}-1}\frac{K_\mathrm{R}-r+2}{1+m}\binom{K_\mathrm{R}-1}{m}\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)^m\geq\sum_{m=0}^{K_\mathrm{R}-1}\frac{K_\mathrm{R}+r}{\min\{r+m,K_\mathrm{R}\}}\binom{K_\mathrm{R}-1}{m}\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)^m.$$

By rearranging the above inequality and using the fact that $\binom{K_\mathrm{R}}{m+1}=\binom{K_\mathrm{R}-1}{m}\frac{K_\mathrm{R}}{m+1}$, we obtain

$$\frac{K_\mathrm{R}-r+2}{K_\mathrm{R}+r}\sum_{m=1}^{K_\mathrm{R}}\binom{K_\mathrm{R}}{m}\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)^m\geq\sum_{m=0}^{K_\mathrm{R}-1}\frac{K_\mathrm{R}}{\min\{r+m,K_\mathrm{R}\}}\binom{K_\mathrm{R}-1}{m}\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)^{m+1}. \tag{5.52}$$

By employing $\binom{K_\mathrm{R}}{m+1}=\binom{K_\mathrm{R}-1}{m}\frac{K_\mathrm{R}}{m+1}$ one more time, we finally arrive at

$$\frac{K_\mathrm{R}-r+2}{K_\mathrm{R}+r}\left[\left(\frac{2K_\mathrm{R}}{K_\mathrm{R}+r}\right)^{K_\mathrm{R}}-1\right]\geq\sum_{m=1}^{K_\mathrm{R}}\frac{m}{\min\{r+m-1,K_\mathrm{R}\}}\binom{K_\mathrm{R}}{m}\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)^m. \tag{5.53}$$

where in going from (5.52) to (5.53), we used the binomial identity to obtain $\sum_{m=1}^{K_\mathrm{R}}\binom{K_\mathrm{R}}{m}\left(\frac{K_\mathrm{R}-r}{K_\mathrm{R}+r}\right)^m=$

$\left(\frac{2K_{\mathrm{R}}}{K_{\mathrm{R}}+r}\right)^{K_{\mathrm{R}}} - 1$. At this point, it is evident that the inequality in (5.53) holds true due to (5.51) in Lemma 5.2. Therefore, (5.49) holds and the proof of (5.44) is complete.

5.8. Summary of the Chapter

In this chapter, we considered the problem of cache-aided interference management in a wireless network where each node is equipped with a cache memory and transmission occurs over two parallel channels, one for which perfect CSIT is available and another for which no CSIT is available. Focusing on strategies with uncoded placement and separable one-shot linear delivery schemes, we characterized the optimal one-shot linear sum-DoF to within a multiplicative factor of 2. We further considered a decentralized setting in which content caching at the receivers is randomized. For this decentralized setting, we characterized the optimal one-shot linear sum-DoF to within a multiplicative factor of 3. Our results generalize and expand upon previous one-shot linear sum-DoF results in literature, namely [100] and [31], by including the parallel no CSIT (or multicast) channel and by considering decentralization at the receivers. The order optimality proof for the decentralized setting posed a number of technical challenges, which were circumvented by involved mathematical manipulations and employing the notion of quasiconcavity. Moreover, the results in this chapter are the first important steps towards the characterization of the information-theoretic limits of cache-aided interference networks under partial CSIT and partial cooperation.

6. Conclusion

In the last two decades, a great deal of research has made significant progress towards the understanding of the capacity limits of multiantenna wireless networks under perfect CSIT. While this has been a milestone in the information-theoretic literature, the acquisition of perfect channel state information is hardly achieved in practical networks. Initial studies and deployments strived to apply multiantenna techniques to scenarios with partial CSIT. However, recent breakthroughs in the study of capacity approximation frameworks such as the DoF or the GDoF unveiled that such approach is fundamentally flawed as it fails to achieve the information theoretic limits of the channels. While the design of capacity-achieving robust interference management strategies is rather complicated on the basis of the current information-theoretic techniques, the use of DoF or GDoF metrics has allowed to shed some light on this fundamental problem.

This thesis made progress towards the design of optimal interference management strategies for multiantenna wireless networks with partial instantaneous CSIT. In particular, we characterized the optimal sum-DoF (or sum-GDoF) and DoF regions for different kind of network settings. A two-fold approach was taken in each of these settings: 1) an outer-bound (or upper-bound) was first derived 2) an achievable scheme which attains such outer-bound was constructed.

We first considered classical content-oblivious networks, where no content can be predicted and prestored in advance. In Chapter 2 we derived the optimal DoF region of the K-user MISO BC with arbitrary CSIT levels. On the basis of previous results in the literature which had established the optimal sum-DoF, we first derived an outer-bound of the optimal DoF region. We then proved the achievability of such outer-bound by considering a rate-splitting strategy with flexible power allocation for the private codewords and flexible allocation of the DoF of the common codeword. To show the achievability, we introduced a novel and unconventional approach where, instead of characterizing and showing the achievability of the corner points, we characterized and showed the achievability of each facet of the polyhedral outer-bound.

In Chapter 3 we extended the work in Chapter 2 by studying the DoF behavior of the overloaded MISO BC, where the number of users is larger than the number of transmitting antennas. To simplify the analysis, we considered a setup where a number of users equal to the number of transmitting antennas have partial CSIT and the remaining users have no CSIT. We first proposed a scheme based on power partitioning where all users are simultaneously served in a non-orthogonal manner and we showed that it achieves a strict DoF gain compared to an orthogonal scheme where the two subsets of users with and without partial CSIT are independently served. We then showed that a generalized version of such power partitioning scheme could achieve the entire DoF region for the considered setting.

We next moved from content-oblivious networks to content-aware networks, where the edge-nodes can predict and prestore part of the most popular content in their cache memories. In Chapter 4 we considered the symmetric K-user cache-aided MISO BC with partial CSIT, where each user is equipped with a cache memory where it can prestore part of the content library. We characterized the optimal sum-GDoF of the network up to a constant multiplicative factor of 12 for all system parameters. Furthermore, we showed that such sum-GDoF characterization is robust to decentralization, where no coordination is allowed during the placement phase. The construction of the GDoF upper-bound extended a family of robust outer bounds based on the aligned image sets approach, initially developed in the context of classical networks with no caches, to cache-aided networks. On the other hand, the achievability schemes relied on the interplay between coded-caching, to enable coded multicasting opportunities, and rate-splitting, to enable spatial multiplexing gains.

In Chapter 5 we extended the work in Chapter 4 by considering a cache-aided interference network with an arbitrary number of transmitters and receivers, where each transmitter or receiver can store a fraction of the content library. We assumed that the transmitters and receivers could communicate through two parallel channels, one for which perfect CSIT is available and another for which no CSIT is available. The partial CSIT can be then seen as the fraction of the bandwidth given by channel with perfect CSIT. By assuming separable one-shot linear delivery schemes and uncoded placement, we derived the optimal sum-DoF of the network up to a constant multiplicative factor of 2 for all system parameters. While this result was obtained by assuming centralized placement, we showed that order-optimality is still attained in a decentralized setting, where a centrally coordinated placement is not allowed at the receivers side. Our results generalized and expanded upon previous one-shot linear DoF results in literature, namely [100] and [31], by including the parallel no CSIT (or multicast) channel and by considering decentralization at the receivers. The order optimality proof for the decentralized setting posed a number of technical challenges, which were circumvented by involved mathematical manipulations and employing the notion of quasiconcavity.

From a more philosophical and abstract perspective, this thesis makes a step forward towards a deeper understanding of the uncertainties arising in communication systems. Looking at the history of Information Theory it is possible to appreciate the effort made by the researchers to decouple these uncertainties and simplify the study of the fundamental limits of communications by analyzing them individually. The most understood uncertainty is the noise. Starting from the groundbreaking work of Shannon in 1948, a significant progress has been made in the study of communication systems affected by noise only. This deep understanding has allowed a big leap forward in the design of high-speed wireless communication systems. However, while dealing with noise allows to serve at high data rate a single user in isolation, the situation becomes more complicated in a multi-user scenario due the presence of interference. The ability to manage interference is intimately related to another intrinsic uncertainty of communication systems, which is the quality of the channel state information at the transmitter. The results in this thesis make progress towards the understading of the effect of imperfect channel state information in the performance of multi-user setups in wireless communication systems. Hopefully this will provide insights in the design of novel interference management techniques which will dramatically improve the performance of the future generation of communication systems.

Recently, a new job has brought me in the field of natural language processing. My task is to make sense of sentences, i.e. extract their content. Again, this task is intimately related to the understanding of uncertainties. In fact, the main uncertainty here is the form of the sentence, i.e. the different ways to convey the same content. Even if in a completely different field, the experience I have acquired by studying uncertainties in communication systems has allowed me to better understand how to deal with the uncertainties arising in natural language processing.

6.1. Future Work

This PhD thesis has been the result of a 4 years journey, which started in October 2015, made of exciting as well as frustrating periods. The main lesson learnt during this long journey has been to never give up. PhD is made of an uncountable number of down moments, where I often felt completely lost. During these moments, I could not find any new interesting direction or idea to work on, or I had to deal with problems which seemed insurmountable. In fact, during this journey, many problems were investigate and, while some of them could be solved, others still remain open today. For instance, in Chapter 3 we studied the DoF of the overloaded MISO BC by restricting the setup to the case where a number of users equal to the number of transmitting antennas have partial CSIT and the remaining users have no CSIT. This in turn allows, with some minor modifications, to utilize the sum-DoF upper-bound in [22] in order to derive an outer-bound of the optimal DoF region. While this is an interesting problem on its own, our original direction was to consider a symmetric setting where all users have partial CSIT. However, the main challenge which we were not able to solve is the derivation of a robust sum-DoF upper-bound. It is unclear if this can be obtained by extending the aligned image set approach in [22] or it requires the development of novel techniques. This problem remains unsolved.

Regarding Chapter 4, for the original shared-link setting, recent efforts managed to reduce the constant multiplicative factor to 2 [86, 89]. Our observations through numerical simulations, which shows that the gap is much smaller than 12, provide hope that such tightening may also be possible for the order-optimal characterizations presented in our work.

Regarding Chapter 5, our initial intention was to study the cache-aided interference channel with the same partial instantaneous CSIT as defined in the previous chapters. However, this was posing very difficult challenges in the design of the achievability scheme as the multicasting messages cannot be jointly encoded when they belong to different transmitters. Hence, instead of partial CSIT, we considered a PN-parallel channel model, inspired by the works in [73, 125] which have made the link between wireless networks where CSI is only reported for a fraction of the bandwidth and wireless networks where CSI is reported for the entire bandwidth but with a certain quality (for instance, receivers report the CSI over a certain number of bits), as explained in Chapter 5. Hence, we leave as future work the study of cache-aided interference management where the transmitters have a partial instantaneous knowledge of the CSIT of the users. To summarize, our work leaves open a number of problems, and we summarize some of them below:

- **Optimal Sum-DoF and optimal DoF region of the Overloaded MISO BC:** In Chapter 3 we studied the DoF of the overloaded MISO BC by restricting the setup to the case where a number of users equal to the number of transmitting antennas have partial CSIT and the remaining users have no CSIT. As aforementioned, as important extension of this work would be to consider an overloaded setting where all users have partial CSIT (restricting for simplicity to a symmetric setting). However, the main challenge here would the derivationn of the sum-DoF upper-bound and DoF region outer-bound.

- **Characterization of order-optimal sum-GDoF of the K-user cache-aided MISO BC with arbitrary CSIT levels:** In Chapter 4 we studied the GDoF of the cache-aided MISO BC for a symmetric setting of the CSIT levels. An interesting, even though very intricate generalization, would be to characterize order-optimal sum-GDoF for arbitrary CSIT levels of the users. The major difficulty here is the potential explosion in the number of channel parameters. Therefore it is not surprising that such asymmetric GDoF characterizations are still open even in classical networks [42, 43].

- **Reduction of the constant multiplicative factor of 12 for the order-optimal sum-GDoF of the K-user cache-aided MISO BC:** For the original shared-link setting, recent efforts managed to reduce the constant multiplicative factor to 2 [86, 89]. As aforementioned, our observations through numerical simulations show that the gap is much smaller than 12, and this provides hope that such tightening may also be possible for the order-optimal characterizations presented in our work.

- **Fundamental limits of cache-aided interference management in heterogeneous parallel channels**: In Chapter 5 we studied the cache-aided interference channel by assuming uncoded placement and one-shot linear delivery schemes. An intriguing direction would be to explore the fundamental limits of the considered setup while relaxing such restrictions. While we expect uncoded placement to still be order-optimal, the delivery scheme will likely rely on interference alignment and symbol spreading. This direction builds upon and benefit from recent results reported in [34, 106].

- **Fundamental limits of cache-aided interference management with partial CSIT**: As already afroremensioned, another important direction would be to extend the model in Chapter 5 by considering partial instantaneous CSIT instead of parallel subchannels.

- **Extension of the cache-aided interference setup to F-RAN networks:** Another interesting direction would be to extend the setup and results in Chapter 5 to F-RAN architectures, where decentralized placement can also be afforded at the transmitters due to the supporting cloud [118–120]. Such direction will also be relevant to D2D networks underlaying a cellular infrastructure, that performs the role of the cloud, which can benefit from the lower complexity one-shot linear schemes.

Before concluding, I would like to mention an important direction of research which it has only been touched in this thesis but it could be of potential interest as PhD topic for a new

PhD student. In this thesis we wanted to characterize the fundamental limits of robust interference management and we had to consider the DoF/GDoF frameworks to make this characterization analytically tractable. While this is interesting from an information-theoretic perspective, still leaves open the question whether this is implementable in practical wireless communication networks. Hence, I would suggest to a new PhD student to consider the problem of robust interference management in the finite SNR regime. By considering classical scenarios without caches, previous works have tackled this problem by, for instance, formulating it as an optimization problem for the precoders of the codewords [45,46]. However, no works have considered robust interference management with caches. As mentioned above, an important direction of research, which would be crucial to evaluate the impact of caching in practical wireless networks, would be then the design of robust interference management techniques in the finite SNR regime for cache-aided wireless networks.

A. Proofs for Chapter A

Proof of the optimal DoF region $\mathcal{D}^*$

The DoF region $\mathcal{D}^*$ described by the inequalities in (3.9) and (3.10) is a K_{R}-dimensional polyhedral region. As in Chapter 2, we prove the optimality of $\mathcal{D}^*$ by showing that it is simultaneously achievable and an outer-bound of the optimal DoF region.

Achievability: In this section we prove the achievability of $\mathcal{D}^*$. Before we delve into the general case, we first characterize the achievable DoF region obtained by switching off all users in $\mathcal{K}_0$ (forcing their DoF to zero). This is equivalent to projecting $\mathcal{D}^*$ onto the K_{T} dimensional subspace characterized by $d_{K_{\mathrm{T}}+1}, \dots, d_{K_{\mathrm{R}}} = 0$. It is readily seen that this setup corresponds to the K-user MISO BC studied in Chapter 2 for the case where all users have the same CSIT quality. The corresponding DoF region can be then obtained from Theorem 2.2 and it is given by the lemma below. This region is then utilized as a building block to prove the achievability of $\mathcal{D}^*$.

Lemma A.1. *For a MISO BC with $K_{\mathrm{R}} = K_{\mathrm{T}}$ users and CSIT quality $\beta \in [0,1]$ for all users, the optimal DoF region $\mathcal{D}_{K_{\mathrm{R}}=K_{\mathrm{T}}}$ is given by*

$$d_k \geq 0, \quad \forall k \in \mathcal{K}_{\mathrm{R}} \tag{A.1}$$

$$\sum_{k \in \mathcal{S}} d_k \leq 1 + (|\mathcal{S}| - 1)\beta, \quad \forall \mathcal{S} \subseteq \mathcal{K}_{\mathrm{R}}, |\mathcal{S}| \geq 1 \tag{A.2}$$

where $\mathcal{K}_{\mathrm{R}}$ denotes the set of users $\{1, \dots, K_{\mathrm{R}}\}$.

We can now proceed to show the achievability of the region $\mathcal{D}^*$. First, defining $d_\Sigma = \sum_{i \in \mathcal{K}_0} d_i$, the problem is equivalent to showing that all the non-negative tuples $(d_1, \dots, d_M, d_\Sigma)$ that satisfy

$$d_i \geq 0, d_\Sigma \geq 0 \quad \forall i \in \mathcal{K}_\beta \tag{A.3}$$

$$\sum_{i \in \mathcal{S}} d_i + d_\Sigma \leq 1 + (|\mathcal{S}| - 1)\beta, \quad \forall \mathcal{S} \subseteq \mathcal{K}_\beta, |\mathcal{S}| \geq 1 \tag{A.4}$$

are achievable. All tuples $(d_1, \dots, d_M, d_{M+1}, \dots, d_K)$ are then obtained by splitting, in all possible variants, the values of d_Σ among users in $\mathcal{K}_0$. The proof follows the same steps as the proof of Theorem 2.2 in Section 2.8. In this case, the induction is done over the number of users in $\mathcal{K}_\beta$, denoted as K_β and equal to K_{T}. The case $K_\beta = 1$ is trivial. We assume that the hypothesis holds for $K_\beta = 1, \dots, k-1$. As before, we show that each facet of the polyhedron is achievable. Starting with the hyperplanes in (A.4), for each subset $\mathcal{S} \subseteq \mathcal{K}_\beta, |\mathcal{S}| \geq 1$, we need to show that all the non-negative tuples $(d_1, \dots, d_k, d_\Sigma)$ that satisfy

$$\begin{cases} \sum_{i\in\mathcal{S}} d_i + d_\Sigma = 1 + (|\mathcal{S}|-1)\beta \\ \sum_{i\in\bar{\mathcal{S}}} d_i + d_\Sigma \leq 1 + (|\bar{\mathcal{S}}|-1)\beta, \forall\bar{\mathcal{S}} \subseteq \mathcal{K}_\beta, \bar{\mathcal{S}} \neq \mathcal{S}, |\bar{\mathcal{S}}| \geq 1 \end{cases}$$

are achievable. Following the same steps as in Section 2.8, it can be verified that the above conditions are equivalent to

$$\begin{cases} \sum_{i\in\mathcal{S}} d_i + d_\Sigma = 1 + (|\mathcal{S}|-1)\beta & \\ d_i \geq \beta, & \forall i \in \mathcal{S} \\ d_i \leq \beta, & \forall i \in \mathcal{K}_\beta \setminus \mathcal{S}. \end{cases}$$

Each DoF tuple is achieved through power partitioning by allocating powers scaling as $O(P^\beta)$ to private symbols of users $i \in \mathcal{S}$, and powers scaling as $O(P^{d_i})$ to private symbols of users $i \in \mathcal{K}_\beta \backslash \mathcal{S}$. On top, we consider all possible power partitions $\Lambda \in [\beta, 1]$ and for each partition, the common symbol's DoF is split, in all possible variants, among users $k \in \mathcal{S}$ only, while $d_\Sigma = 1 - \Lambda$.

Considering the facets contained in the hyperplanes in (A.4), we have two cases. The first is given by $d_\Sigma = 0$ and it reduces to k users with CSIT β and k antennas as in Lemma A.1. The second case considers any $j \in \mathcal{K}_\beta$ and we have

$$\begin{cases} d_j = 0 \\ \sum_{i\in\mathcal{S}} d_i + d_\Sigma \leq 1 + (|\mathcal{S}|-1)\beta, \forall\mathcal{S} \subseteq \mathcal{K}_\beta \setminus \{j\}, |\mathcal{S}| \geq 1. \end{cases}$$

This corresponds to the region in (A.3) and (A.4) considering the $k-1$ users in $\mathcal{K}_\beta$. Using the same argument as before, this region is achievable by induction. Moreover, all facets of the polyhedron are achievable, all the remaining points can be achieved by time-sharing

Converse: The converse is based on the sum-DoF upper-bound obtained in [22] and it is similar as the one in Section 2.7 in Chapter 2. For an arbitrary subset of users $\mathcal{U} \subseteq \mathcal{K}$, the sum-DoF is upperbounded by

$$\sum_{k\in\mathcal{U}} d_k \leq 1 + \beta(|\mathcal{S}|-1)^+ \tag{A.5}$$

where $\mathcal{S} = \mathcal{U} \cap \mathcal{K}_\beta$. We increase the number of transmitter antennas to K_R and then enhance the quality of one of the users in $\mathcal{S}$ to 1 (if $\mathcal{S}$ is empty we pick any other user). Since the previous steps provide an outer-bound and cannot harm the DoF, (A.5) directly follows from [22, Theorem 1]. By removing all redundant inequalities, the outer-bound coincides with the region $\mathcal{D}^*$.

B. Proofs for Chapter 4

B.1. Proof of Lemma 4.1

The proof is based on the approach in [22, 41, 43], where outer bounds under finite precision and partial CSIT are derived. We follow the same overall steps in these works, while specializing to the specific setup considered here. For simplicity and notational briefness, we focus on real channels. The extension to complex channels follows along the lines of [22, 41]. We consider $s = K$ users. For general $s \leq K$, the exact same steps follow while considering only the corresponding s rate bounds.

B.1.1. Deterministic Channel Model

The first step is to convert the channel into a deterministic equivalent with inputs and outputs all being integers. This is given by

$$\bar{Y}_i(t) = \lfloor G_{ii}(t)\bar{X}_i(t) \rfloor + \sum_{j=1, j\neq i}^{K} \lfloor \bar{P}^{\alpha-1} G_{ij}(t)\bar{X}_j(t) \rfloor \tag{B.1}$$

$$\bar{B}_i(t) = \bar{A}_i(t) \tag{B.2}$$

where $\bar{P} = \sqrt{P}$, $\bar{X}_i(t) \in \{0, 1, \ldots, \lfloor \bar{P} \rfloor\}$ and $\bar{A}_i(t) \in \{0, 1, \ldots, \lfloor \bar{P}^{\gamma} \rfloor\}$, $\forall i \in [K]$. It can be shown that a sum-GDoF upper-bound for the deterministic channel is also a sum-GDoF upper-bound for the original channel using the same steps in [22]. Therefore we focus on the deterministic channel henceforth.

B.1.2. Fanos Inequality and Differences of Entropies

For notational brevity, we define $M_i \triangleq \left(W_{d_i^1}, \ldots, W_{d_i^{\lfloor N/s \rfloor}}\right)$ to denote the set of messages to be delivered to user i. Moreover, we define $M_{[i:K]} \triangleq M_i, \ldots, M_K$. Using Fano's inequality, for user k we have

$$\begin{aligned} nR_k &\leq I\left(M_k; \bar{Y}_k^n, \bar{B}_k^n \mid M_{[k+1:K]}, \mathcal{G}\right) + o(n) & \text{(B.3)} \\ &\leq H\left(\bar{Y}_k^n, \bar{B}_k^n \mid M_{[k+1:K]}, \mathcal{G}\right) - H\left(\bar{Y}_k^n, \bar{B}_k^n \mid M_{[k:K]}, \mathcal{G}\right) + o(n). & \text{(B.4)} \end{aligned}$$

After omitting $o(n)$ and $o\left(\log(P)\right)$ terms, we obtain

$$n\sum_{k=1}^{K} R_k \leq n(1+\gamma)\log(\bar{P}) + \sum_{k=2}^{K} \underbrace{H\left(\bar{Y}_{k-1}^{n}, \bar{B}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}\right) - H\left(\bar{Y}_{k}^{n}, \bar{B}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}\right)}_{H_k^{\Delta}}. \tag{B.5}$$

Hence, the focus becomes to bound the differences of entropies $H_2^{\Delta}, \ldots, H_K^{\Delta}$.

B.1.3. Bounding the Differences of Entropies

Focusing on the term H_k^{Δ}, $k \in [2:K]$, we proceed as follows:

$$\begin{aligned} H_k^{\Delta} &= H\left(\bar{Y}_{k-1}^{n}, \bar{B}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}\right) - H\left(\bar{Y}_{k}^{n}, \bar{B}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}\right) && \text{(B.6)} \\ &= H\left(\bar{Y}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}\right) - H\left(\bar{Y}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}\right) \\ &\quad + H\left(\bar{B}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_{k-1}^{[n]}\right) - H\left(\bar{B}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_{k}^{n}\right) && \text{(B.7)} \\ &\leq H\left(\bar{Y}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}\right) - H\left(\bar{Y}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}\right) + n\log\left(\bar{P}^{\gamma}+1\right). && \text{(B.8)} \end{aligned}$$

In the above, (B.7) is obtained from the chain rule, while (B.8) follows from $H\big(\bar{B}_k^n \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_k^n\big) \geq 0$ and $H\big(\bar{B}_{k-1}^n \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_{k-1}^n\big) \leq H\big(\bar{B}_{k-1}^n\big) \leq \sum_{t=1}^{n} H\big(\bar{B}_{k-1}(t)\big) \leq n\log\left(\bar{P}^{\gamma}+1\right)$. Now it remains to bound the difference of entropies $H\left(\bar{Y}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}\right) - H\left(\bar{Y}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}\right)$ under partial CSIT and the bounded density assumptions as described in Section 4.3.1. This difference is bounded above as

$$H\left(\bar{Y}_{k-1}^{n} \mid M_{[k:K]}, \mathcal{G}\right) - H\left(\bar{Y}_{k}^{n} \mid M_{[k:K]}, \mathcal{G}\right) \leq n\big(1-(\alpha-\beta)\big)\log(\bar{P}) + o\big(\log(\bar{P})\big). \tag{B.9}$$

The inequality in (B.9) follows directly from [43] (see the proofs of [43, Th. 1] and [43, Th. 2]), and is obtained using the aligned image sets approach [22]. Intuitively, under perfect CSIT (i.e. $\beta = \alpha$), the transmitter uses zero-forcing to create a maximal difference of entropies, in a GDoF sense, between $\bar{Y}_{k-1}^{n}$ and $\bar{Y}_{k}^{n}$. On the other hand, when all paths have equal strengths and the CSIT is limited to finite precision (i.e. $\alpha = 1$ and $\beta = 0$), a positive difference of entropies in a GDoF sense cannot be created. Between the two extremes, the transmitter benefits from path-loss and partial CSIT, through power control and zero-forcing, to create a positive difference of entropies which is bounded above by 1, in a GDoF sense.

By combining the bounds in (B.9) and (B.8), we obtain

$$H_k^{\Delta} \leq n\big(\gamma + 1 - (\alpha-\beta)\big)\log(\bar{P}) + o\big(\log(\bar{P})\big). \tag{B.10}$$

The bound in (B.10) holds for all $k \in [2:K]$. By plugging (B.10) into (B.5), the result in (4.24) directly follows.

B.2. Proof of Lemma 4.2

First, let us rewrite the signal model in (4.4) in vector form as (for brevity, the time index is omitted)

$$Y_i = \sqrt{P}\,[\hat{G}_{i1} \cdots \hat{G}_{iK}]\,\mathbf{Q}_i\,\mathbf{X} + \sqrt{P^{1-\beta}}\,[\tilde{G}_{i1} \cdots \tilde{G}_{iK}]\,\mathbf{Q}_i\,\mathbf{X} + Z_i \tag{B.11}$$

where $\mathbf{X} \triangleq [X_1 \cdots X_K]^{\mathsf{T}}$ is the signal transmitted from the K transmitters and $\mathbf{Q}_i$ is a $K \times K$ diagonal matrix with 1 as the (i, i)-th entry and $\sqrt{P^{\alpha-1}}$ as the remaining diagonal entries. Note that we ignore the time index for brevity. The messages $W^{(\mathrm{c})}$ and $W_1^{(\mathrm{p})}, \ldots, W_K^{(\mathrm{p})}$ are encoded into unit power independent Gaussian codewords $X^{(\mathrm{c})}$ and $X_1^{(\mathrm{p})}, \ldots, X_K^{(\mathrm{p})}$, respectively. The transmitted signal is then constructed as

$$\mathbf{X} = \mathbf{D}\left(\sqrt{1 - P^{\beta-\alpha}}\mathbf{V}^{(\mathrm{c})}X^{(\mathrm{c})} + \sqrt{P^{\beta-\alpha}}\sum_{k=1}^{K}\mathbf{V}_k^{(\mathrm{p})}X_k^{(\mathrm{p})}\right). \tag{B.12}$$

In the above, $\mathbf{D}$ is a $K \times K$ diagonal matrix where the (j, j)-th entry is $O(1)$ in P, and is chosen such that the power constraint $\mathbb{E}\left(|X_j|^2\right) \leq 1$ is not violated. $\mathbf{V}^{(\mathrm{c})}$ is a generic (random) unit vector and $\mathbf{V}_k^{(\mathrm{p})} \triangleq \left[V_{k1}^{(\mathrm{p})} \;\cdots V_{kK}^{(\mathrm{p})}\right]^{\mathsf{T}}$ is a zero-forcing unit vector designed using the channel estimates such that

$$\sqrt{P^{\alpha}}\left(\hat{G}_{i1}V_{k1}^{(\mathrm{p})} + \cdots + \sqrt{P^{1-\alpha}}\hat{G}_{ii}V_{ki}^{(\mathrm{p})} + \cdots + \hat{G}_{iK}V_{kK}^{(\mathrm{p})}\right) = 0, \; \forall i \neq k. \tag{B.13}$$

It is simple to verify from the zero-forcing condition that $V_{ki}^{(\mathrm{p})}$ cannot scale faster than $O(\sqrt{P^{\alpha-1}})$ for all $k \neq i$. Hence, the received signal of user i is rewritten as

$$Y_i = \sqrt{P}a_i^{(\mathrm{c})}X^{(\mathrm{c})} + \sqrt{P^{1+\beta-\alpha}}a_{ii}^{(\mathrm{p})}X_i^{(\mathrm{p})} + \sum_{k=1,k\neq i}^{K} a_{ik}^{(\mathrm{p})}X_k^{(\mathrm{p})} + Z_i \tag{B.14}$$

where $a_i^{(\mathrm{c})}$ and $a_{ik}^{(\mathrm{p})}$, for all $i, k \in [K]$, are all $O(1)$.

Each user i decodes $X^{(\mathrm{c})}$ by treating interference as noise and recovers $W^{(\mathrm{c})}$. As $X^{(\mathrm{c})}$ is received with power $O(P)$, while interference plus noise has power $O(P^{1+\beta-\alpha})$, it follows that $X^{(\mathrm{c})}$ supports a rate of $(\alpha - \beta)\log(P) + o\big(\log(P)\big)$. Then, each user i proceeds to remove the contribution of $X^{(\mathrm{c})}$ from the received signal and decodes its own $X_i^{(\mathrm{p})}$ while treating the remaining interference as noise, from which $W_i^{(\mathrm{p})}$ is recovered. As $X_i^{(\mathrm{p})}$ is received with power $O(P^{1+\beta-\alpha})$, while the remaining interference plus noise has power $O(1)$, it follows that $X_i^{(\mathrm{p})}$ supports a rate of $(1 + \beta - \alpha)\log(P) + o\big(\log(P)\big)$.

Remark B.1. *It is worthwhile highlighting that the achievable GDoF in Lemma 4.2 (shown in this appendix) can be inferred from [43]. One key difference, however, is that the MISO BC considered in [43] has private messages only, and rate-splitting is used to multicast part of the private messages as a common codeword decoded by all users. This relationship between the MISO BC with private messages and its counterpart with a common message under partial CSIT was first* *observed in [13].*

B.3. Proofs of Order Optimality

Here we provide proofs for the order-optimality parts of Theorem 4.1 and Theorem 4.2. We start with an instrumental lemma used throughout the proofs in the following subsections.

Lemma B.1. *For parameters K, μ and s defined previously, if $\frac{K}{s(1+K\mu)} \geq 1$, then the function given by*

$$f(\delta; K, \mu, s) = \frac{1 + (s-1)(1-\delta)}{K(1-\delta) + (1+K\mu)\delta} \tag{B.15}$$

is non-decreasing in $\delta \in [0, 1]$.

Proof. The derivative of $f(\delta; K, \mu, s)$ with respect to δ is given by $\frac{df}{d\delta} = -\frac{s(1+K\mu)-K}{(K(1-\delta)+(1+K\mu)\delta)^2}$, which is non-negative for $K \geq s(1 + K\mu)$. □

B.3.1. Order Optimality of $\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)$

We show here that for any μ, there exists a particular $s \in [K]$ such that $\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)/\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta) \leq 12$ for all $\delta \in [0, 1]$. We handle the two cases $K \leq 12$ and $K \geq 13$ separately. Starting with $K \leq 12$, consider a generic $\delta \in [0, 1]$. By setting $s = 1$ in (4.19), we get that $\mathsf{GNDT}_1^{\mathrm{lb}}(\mu, \delta) = 1 - \mu$. On the other hand, $\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta) \leq \mathsf{GNDT}_{\mathrm{C}}(\mu, 1) \leq K(1 - \mu)$. Hence, $\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)/\mathsf{GNDT}_1^{\mathrm{lb}}(\mu, \delta) \leq 12$.

Next, we consider $K \geq 13$. As in [24], we split the problem in three sub-cases: the sub-case $0 \leq \mu \leq \frac{1.1}{K}$, the sub-case $\frac{1.1}{K} < \mu \leq 0.092$ and the sub-case $0.092 < \mu \leq 1$. We start with $0 \leq \mu \leq \frac{1.1}{K}$. For $\delta = 1$, we have $\mathsf{GNDT}_{\mathrm{C}}(\mu, 1) \leq \mathsf{GNDT}_{\mathrm{C}}(0, 1) = K$. By setting $s = \lfloor 0.275K \rfloor$, we know from [24] that $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, 1) \geq K/12$. On the other hand, for a generic $\delta \in [0, 1]$, the following upper-bound holds

$$\frac{\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{C}}(0, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} = \underbrace{\frac{1 + (s-1)(1-\delta)}{K(1-\delta) + \delta}}_{f(\delta; K, 0, s)} \cdot \frac{K}{s\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)}. \tag{B.16}$$

Since $\frac{K}{s} \geq \frac{1}{0.275} > 1$, from Lemma B.1 it follows that $f(\delta; K, 0, s)$ is non-decreasing in $\delta \in [0, 1]$. Hence,

$$\frac{\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{C}}(0, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{C}}(0, 1)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, 1)} \leq 12. \tag{B.17}$$

We proceed to the sub-case $\frac{1.1}{K} < \mu \leq 0.092$. Let $\tilde{\mu}$ be the largest number in $[0, \mu]$ such that $K\tilde{\mu}$ is an integer. We know from [24] that $\mathsf{GNDT}_{\mathrm{C}}(\mu, 1) \leq \mathsf{GNDT}_{\mathrm{C}}(\tilde{\mu}, 1) \leq \frac{1}{\mu}$. By setting $s = \left\lfloor \frac{0.3}{\mu} \right\rfloor$, we also know from [24] that $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, 1) \geq \frac{1}{12\mu}$. Considering a generic $\delta \in [0, 1]$, we write

$$\frac{\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{C}}(\tilde{\mu}, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} = \underbrace{\frac{1 + (s-1)(1-\delta)}{K(1-\delta) + (1+K\tilde{\mu})\delta}}_{f(\delta; K, \tilde{\mu}, s)} \cdot \frac{K\,(1-\tilde{\mu})}{s\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)}. \tag{B.18}$$

As $\frac{K}{s(1+K\tilde{\mu})} \geq \frac{1}{0.3}\frac{K\mu}{1+K\mu} > 1$, Lemma B.1 implies that $f(\delta; K, \tilde{\mu}, s)$ is non-decreasing in $\delta \in [0, 1]$. Hence,

$$\frac{\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{C}}(\tilde{\mu},\delta)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{C}}(\tilde{\mu},1)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,1)} \leq 12. \tag{B.19}$$

Finally, we look at the sub-case $0.092 < \mu \leq 1$ and we consider a generic $\delta \in [0, 1]$. By setting $s = 1$, we get $\mathsf{GNDT}_{1}^{\mathrm{lb}}(\mu, \delta) = 1 - \mu$. Moreover, from [24], we know that $\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta) \leq \mathsf{GNDT}_{\mathrm{C}}(\mu, 1) \leq 12(1 - \mu)$. Hence $\mathsf{GNDT}_{\mathrm{C}}(\mu, \delta)/\mathsf{GNDT}_{1}^{\mathrm{lb}}(\mu, \delta) \leq 12$. This concludes the proof.

B.3.2. Order Optimality of $\mathsf{GNDT}_{\mathrm{D}}(\mu, \delta)$

As for the centralized setting, we show that for any μ, there exists a particular $s \in [K]$ such that $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu, \delta)/\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu, \delta) \leq 12$ for all $\delta \in [0, 1]$. We start with the following lemma.

Lemma B.2. *The value u, defined in* (4.47)*, satisfies $u \leq K\mu$ for all $\mu \in [0, 1)$.*

Proof. We focus on $\mu > 0$ as $u = 0$ for $\mu = 0$. By definition of u in (4.47), we have

$$\frac{K(1-\mu)}{K\mu}\left(1-(1-\mu)^{K}\right) = \frac{K(1-\mu)}{1+u} \tag{B.20}$$

which follows from $\mathsf{GNDT}_{\mathrm{D}}(\mu, 1) = \frac{K(1-\mu)}{K\mu}\left(1 - (1 - \mu)^{K}\right)$, as shown in [79]. Hence, showing that $u \leq K\mu$ it is equivalent to showing that

$$\frac{K(1-\mu)}{K\mu}\left(1-(1-\mu)^{K}\right) \geq \frac{K(1-\mu)}{1+K\mu} \tag{B.21}$$

$$\Rightarrow (K\mu+1)\left(1-(1-\mu)^{K}\right) \geq K\mu \tag{B.22}$$

$$\Rightarrow 1 \geq (K\mu+1)(1-\mu)^{K}. \tag{B.23}$$

The inequality in (B.23) is shown to hold by observing that $\mu > 0$ and $K\mu + 1 \leq (1 + \mu)^{K}$, from which we obtain $(K\mu + 1)(1 - \mu)^{K} \leq (1 + \mu)^{K}(1 - \mu)^{K} = (1 - \mu^{2})^{K} \leq 1$. Hence, $u \leq K\mu$ holds. □

Equipped with Lemma B.2, the remainder of the proof follows the same procedures in Appendix B.3.1. In particular, we consider the two cases $K \leq 12$ and $K \geq 13$. For the case $K \leq 12$, by setting $s = 1$ in (4.19), we get that $\mathsf{GNDT}_{1}^{\mathrm{lb}}(\mu, \delta) = 1 - \mu$. On the other hand, we have $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu, \delta) \leq \mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu, 1) \leq K(1-\mu)$. It follows that $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu, \delta)/\mathsf{GNDT}_{1}^{\mathrm{lb}}(\mu, \delta) \leq 12$.

Next, we focus on $K \geq 13$. As in [79], we consider three separate sub-cases: the sub-case $0 \leq \mu \leq 1/K$, the sub-case $1/K < \mu \leq 1/12$ and the sub-case $1/12 < \mu \leq 1$. We look at the sub-case $0 \leq \mu \leq 1/K$ first. For $\delta = 1$, we have $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu, 1) \leq K$, and by setting $s = \lfloor K/4 \rfloor$, we obtain $\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu, 1) \geq \frac{1}{12}K$ from [79]. On the other hand, for a generic $\delta \in [0, 1]$, we have

$$\frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\delta)} = \underbrace{\frac{1+(s-1)(1-\delta)}{K(1-\delta)+(1+u)\delta}}_{f(\delta;K,u/K,s)} \cdot \frac{K\,(1-\mu)}{s\left(1-\frac{M}{\lfloor\frac{N}{s}\rfloor}\right)}. \tag{B.24}$$

By applying Lemma B.2 to lower bound the value of u, we can write $\frac{K}{s(1+u)} \geq \frac{K}{\frac{K}{4}\cdot(1+K\mu)} > 1$. Hence, from Lemma B.1, the function $f(\delta; K, u/K, s)$ is non-decreasing in $\delta \in [0,1]$. It follows that

$$\frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,1)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,1)} \leq 12. \tag{B.25}$$

Next, we consider the sub-case $\frac{1}{K} < \mu \leq \frac{1}{12}$. From [79], we have $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,1) \leq \frac{1}{\mu}$, and by setting $s = \left\lfloor \frac{1}{4\mu} \right\rfloor$, we have $\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,1) \geq \frac{1}{12\mu}$. For a generic $\delta \in [0,1]$, we have

$$\frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)}{\mathsf{GNDT}_{s}^{\mathrm{lb}}(\mu,\delta)} = \underbrace{\frac{1+(s-1)(1-\delta)}{K(1-\delta)+(1+u)\delta}}_{f(\delta;K,u/K,s)} \cdot \frac{K\,(1-\mu)}{s\left(1-\frac{M}{\lfloor\frac{N}{s}\rfloor}\right)}. \tag{B.26}$$

By applying Lemma B.2, it follows that $\frac{K}{s(1+u)} \geq 4 \cdot \frac{K\mu}{1+K\mu} > 1$. Hence, from Lemma B.1, $f(\delta; K, u/K, s)$ is non-decreasing in $\delta \in [0,1]$. Therefore, the statement in (B.25) holds here as well.

Finally, we consider the remaining sub-case $1/12 < \mu \leq 1$ for a generic $\delta \in [0,1]$. By setting $s = 1$, we get $\mathsf{GNDT}_{1}^{\mathrm{lb}}(\mu,\delta) = 1-\mu$. Moreover, from [79], we know that $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta) \leq \mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,1) \leq \frac{1}{\mu} - 1$. Hence, $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)/\mathsf{GNDT}_{1}^{\mathrm{lb}}(\mu,\delta) \leq 12$. This concludes the proof.

B.4. Proof of Lemma 4.3

It readily seen from the definition of u in (4.47) that $\mathsf{GNDT}_{\mathrm{D}}(\mu,1) = \mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,1)$. It is also easy to verify that $\mathsf{GNDT}_{\mathrm{D}}(\mu,0) = \mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,0)$ and $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(1,\delta) = \mathsf{GNDT}_{\mathrm{D}}(1,\delta) = 0$. Therefore, we focus on $\delta \in (0,1)$ and $\mu \in [0,1)$. We define b_m, $m \in \{0,1,\ldots,K-1\}$, such that

$$b_m = \frac{K\binom{K-1}{m}\mu^m(1-\mu)^{K-m}}{K(1-\mu)}. \tag{B.27}$$

It can be shown that $\sum_{m=0}^{K-1} b_m = 1$ as follows

$$\sum_{m=0}^{K-1} b_m = \frac{1}{K(1-\mu)} \sum_{m=0}^{K-1} K\binom{K-1}{m}\mu^m(1-\mu)^{K-m} \tag{B.28}$$

$$= \sum_{m=0}^{K-1} \binom{K-1}{m}\mu^m(1-\mu)^{K-1-m} = 1 \tag{B.29}$$

where (B.29) follows from the binomial identity[1]. Hence, the inequality in (4.46) is equivalently written as

$$\sum_{m=0}^{K-1} \frac{b_m}{K(1-\delta)+(1+m)\delta} \leq \frac{1}{K(1-\delta)+(1+u)\delta} \tag{B.30}$$

$$\Rightarrow \sum_{m=0}^{K-1} \frac{b_m}{c_m+v} \leq \frac{1}{\tilde{c}+v}. \tag{B.31}$$

where $v \triangleq K(1-\delta)$, $c_m \triangleq (1+m)\delta$ and $\tilde{c} \triangleq (1+u)\delta$. By rearrangement of (B.31), we obtain

$$(\tilde{c}'+1)\sum_{m=0}^{K-1} \frac{b_m}{c'_m+1} \leq 1. \tag{B.32}$$

where $\tilde{c}' = \tilde{c}/v$ and $c'_m = c_m/v$ By the definition of u in (4.47), for any $\delta \in (0,1)$, we have

$$\sum_{m=0}^{K-1} \frac{b_m}{(1+m)\delta} = \frac{1}{(1+u)\delta} \tag{B.33}$$

$$\Rightarrow \sum_{m=0}^{K-1} \frac{b_m}{c_m} = \frac{1}{\tilde{c}} \tag{B.34}$$

$$\Rightarrow \sum_{m=0}^{K-1} \frac{b_m}{c'_m} = \frac{1}{\tilde{c}'}. \tag{B.35}$$

By plugging $\tilde{c}'$ from (B.35) into (B.32), we obtain

$$\left(\frac{1}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m}} + 1\right)\sum_{m=0}^{K-1} \frac{b_m}{c'_m+1} \leq 1. \tag{B.36}$$

Hence, showing that (B.36) holds implies that (4.46) holds for $\delta \in (0,1)$. This is shown next.

Let us define the function $f(v) = \frac{v}{1+v}$, which is concave in $\mathbb{R}_+ \setminus \{0\}$. Moreover, consider the points $\left\{\frac{1}{c'_0}, \ldots, \frac{1}{c'_{K-1}}\right\}$ in $\mathbb{R}_+ \setminus \{0\}$. From $\sum_{m=0}^{K-1} b_m = 1$, which is obtained from (B.29), and by

[1] Recall that the binomial identity is given by $(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$.

applying Jensen's inequality, we have

$$\sum_{m=0}^{K-1} b_m f\left(\frac{1}{c'_m}\right) \leq f\left(\sum_{m=0}^{K-1} \frac{b_m}{c'_m}\right) \tag{B.37}$$

$$\Rightarrow \sum_{m=0}^{K-1} b_m \frac{1}{c'_m + 1} \leq \frac{\sum_{m=0}^{K-1} \frac{b_m}{c'_m}}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m} + 1} \tag{B.38}$$

$$\Rightarrow \left(\frac{\sum_{m=0}^{K-1} \frac{b_m}{c'_m} + 1}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m}}\right)\left(\sum_{m=0}^{K-1} \frac{b_m}{c'_m + 1}\right) \leq 1 \tag{B.39}$$

$$\Rightarrow \left(\frac{1}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m}} + 1\right)\sum_{m=0}^{K-1} \frac{b_m}{c'_m + 1} \leq 1 \tag{B.40}$$

which is the inequality in (B.36). This concludes the proof.

B.5. Proof of Corollary 4.1

First, for $\mu = 0$ we have that $\mathsf{GDoF}_{\mathrm{C}}(0,\delta) = \mathsf{GDoF}_{\mathrm{D}}(0,\delta) = K(1-\delta)+\delta$, while for $\mu = 1$ we have that $\mathsf{GNDT}_{\mathrm{C}}(1,\delta) = \mathsf{GNDT}_{\mathrm{D}}(1,\delta) = 0$. Therefore, we focus on $\mu \in (0,1)$ in what follows. The multiplicative factor of 1.5 in (4.49) can be shown by considering the three following cases:

1. $K \geq 3$: From Theorem 4.1, it follows that $\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)$ is bounded above by

$$\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta) \leq (1-\delta)\frac{K}{1-\mu} + \delta\frac{1+K\mu}{1-\mu} \tag{B.41}$$

where (B.41) holds with equality for $\mu \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}\}$, as expressed in (4.13). For the remaining points in $\mu \in [0,1]$, the achievable sum-GDoF upper-bound in (B.41) follows from

$$\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta) \geq \frac{K(1-\mu)}{K(1-\delta)+(1+K\mu)\delta} \tag{B.42}$$

which in turn holds as $\frac{K(1-\mu)}{K(1-\delta)+(1+K\mu)\delta}$ is convex in μ and $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta)$ is the lower convex envelope (see (4.10)). From Lemma 4.3, a lower-bound for $\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)$ is given by

$$\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta) \geq (1-\delta)\frac{K}{1-\mu} + \delta\frac{1+u}{1-\mu} \tag{B.43}$$

where $1+u = \frac{K(1-\mu)}{\mathsf{GNDT}_{\mathrm{D}}(\mu,1)}$ from (4.47). From [79], we know that $\mathsf{GNDT}_{\mathrm{D}}(\mu,1)$ can be written as

$$\mathsf{GNDT}_{\mathrm{D}}(\mu,1) = \frac{1-\mu}{\mu}\left(1-(1-\mu)^K\right). \tag{B.44}$$

It follows that $1+u$ is given by

$$1+u = \frac{K\mu}{1-(1-\mu)^K}. \tag{B.45}$$

From (B.41) and (B.43), the ratio between $\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)$ and $\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)$ is bounded above as

$$\frac{\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)}{\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)} \leq \frac{(1-\delta)K+\delta(1+K\mu)}{(1-\delta)K+\delta(1+u)} \leq \frac{1+K\mu}{1+u}. \tag{B.46}$$

where the rightmost inequality in (B.46) follows from $u \leq K\mu$, which in turn is obtained from Lemma B.2 in Appendix B.3.2. By plugging (B.45) into (B.46), we obtain

$$\frac{1+K\mu}{1+u} = \frac{1+K\mu}{K\mu}\left(1-(1-\mu)^K\right) \leq 1.5 \tag{B.47}$$

where the bound by 1.5 follows directly from [76, Lem. 1].

2. $K = 2$: For this case, we consider the two following subcases:
   - $\mu \in (0, 1/2]$: For this interval, we employ the same bounding techniques used for the case $K \geq 3$. Hence, from (B.46) and (B.47) we obtain

     $$\frac{\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)}{\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)} \leq \frac{1+2\mu}{2\mu}\left(1-(1-\mu)^2\right). \tag{B.48}$$

     It is readily seen that the right-hand-side of (B.48), which we denote as $g(\mu)$, is a concave parabola with a maximum at $\mu = 3/4$. Given the symmetry of the parabola, it follows that $g(\mu) \leq g(1/2) = 1.5$ for $\mu \in (0, 1/2]$.
   - $\mu \in [1/2, 1)$: For this interval, the bounding techniques used for the case $K \geq 3$ are loose. Alternatively, it can be easily shown from Theorem 4.1 that $\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta) = \frac{2}{1-\mu}$. Combining this with the upper-bound for $\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)$ in (B.43), we obtain

     $$\frac{\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)}{\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)} \leq \frac{2}{2(1-\delta)+(1+u)\delta} \leq \frac{2}{1+u} \tag{B.49}$$

     where the rightmost inequality in (B.49) follows from the fact that $1+u \leq 2$, which can be easily shown. By plugging (B.45) into (B.49), we obtain

     $$\frac{2}{1+u} = \frac{1}{\mu}\left(1-(1-\mu)^2\right) = 2-\mu. \tag{B.50}$$

     It is readily seen that $2-\mu \leq 1.5$ for $\mu \in [1/2, 1)$.

3. Case $K = 1$: In this case we have $\mathsf{GNDT}_{\mathrm{C}}(\mu,\delta) = \mathsf{GNDT}_{\mathrm{D}}(\mu,\delta) = 1-\mu$, hence (4.49) holds.

From the above three cases, the proof is complete. It is worthwhile highlighting that for the case $K = 2$, $\delta = 1$ and $\mu = 1/2$, we have $\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)/\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta) = 1.5$. Therefore, 1.5 is in fact the tightest possible upper-bound for $\mathsf{GDoF}_{\mathrm{C}}(\mu,\delta)/\mathsf{GDoF}_{\mathrm{D}}(\mu,\delta)$.

C. Proofs for Chapter 5

C.1. Proof of (5.36)

Note that $\overline{H^{(\mathrm{n})}}^{(q)}$ corresponds to the optimum objective value for the optimization problem in (5.33) when $\mathrm{s} = \mathrm{n}$. To bound this, we follow here the footsteps in the proof of [31, Lem. 4]. Starting from $H^{(\mathrm{n})*\left(\{\mathcal{P}_i\}_{i=1}^{K_\mathrm{T}},\{\mathcal{U}_i\}_{i=1}^{K_\mathrm{R}},\mathbf{d},q\right)}$ and by invoking (5.27), we obtain

$$H^{(\mathrm{n})*\left(\{\mathcal{P}_i\}_{i=1}^{K_\mathrm{T}},\{\mathcal{U}_i\}_{i=1}^{K_\mathrm{R}},q,\mathbf{d}\right)} \geq \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}} \sum_{r=1}^{K_\mathrm{R}} \sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ |\mathcal{T}|=i}} \sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ |\mathcal{R}|=j\\ r\notin\mathcal{R}}} \frac{a_{d_r,\mathcal{T},\mathcal{R}}}{j+1}. \tag{C.1}$$

By averaging over all possible demands, we obtain

$$\begin{aligned}\overline{H^{(\mathrm{n})}}^{\left(\{P\}_{i=1}^{\mathrm{K_T}},\{\mathcal{U}_i\}_{i=1}^{\mathrm{K}_\mathcal{T}},q\right)} &\geq \frac{1}{\pi(N,K_\mathrm{R})} \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}} \sum_{r=1}^{K_\mathrm{R}} \sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ |\mathcal{T}|=i}} \sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ |\mathcal{R}|=j\\ r\notin\mathcal{R}}} \pi(N-1,K_\mathrm{R}-1) \sum_{n=1}^{N} \frac{a_{n,\mathcal{T},\mathcal{R}}}{j+1}\\ &= \frac{1}{N} \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}-1} \frac{w_{i,j}}{j+1}.\end{aligned} \tag{C.2}$$

where, for any $i \in [K_\mathrm{T}]$ and $j \in [K_\mathrm{R}-1] \cup 0$, we define

$$w_{i,j} = \sum_{r=1}^{K_\mathrm{R}} \sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ |\mathcal{T}|=i}} \sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ |\mathcal{R}|=j\\ r\notin\mathcal{R}}} \sum_{n=1}^{N} a_{n,\mathcal{T},\mathcal{R}} = (K_\mathrm{R}-j) \sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ |\mathcal{T}|=i}} \sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ |\mathcal{R}|=j}} \sum_{n=1}^{N} a_{n,\mathcal{T},\mathcal{R}}. \tag{C.3}$$

It is readily seen that

$$K_\mathrm{R}\mu_\mathrm{R}NF \geq \sum_{r=1}^{K_\mathrm{R}} \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}} \sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ |\mathcal{T}|=i}} \sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ |\mathcal{R}|=j\\ r\in\mathcal{R}}} \sum_{n=1}^{N} a_{n,\mathcal{T},\mathcal{R}} \geq \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}-1} \frac{j}{K_\mathrm{R}-j} w_{i,j} \tag{C.4}$$

and

$$NF = \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}} \sum_{\substack{\mathcal{T}\subseteq[K_\mathrm{T}]:\\ |\mathcal{T}|=i}} \sum_{\substack{\mathcal{R}\subseteq[K_\mathrm{R}]:\\ |\mathcal{R}|=j}} \sum_{n=1}^{N} a_{n,\mathcal{T},\mathcal{R}} \geq \sum_{i=1}^{K_\mathrm{T}} \sum_{j=0}^{K_\mathrm{R}-1} \frac{1}{K_\mathrm{R}-j} w_{i,j}. \tag{C.5}$$

After applying the Cauchy-Schwarz inequality, we obtain

$$\sum_{j=0}^{K_{\mathrm{R}}-1} w_{i,j} \leq \sqrt{\sum_{j=0}^{K_{\mathrm{R}}-1} \frac{j+1}{K_{\mathrm{R}}-j} w_{i,j}} \cdot \sqrt{\sum_{j=0}^{K_{\mathrm{R}}-1} \frac{K_{\mathrm{R}}-j}{j+1} w_{i,j}}. \tag{C.6}$$

Applying the Cauchy-Schwarz inequality again, we obtain

$$\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} w_{i,j} \leq \sqrt{\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} \frac{j+1}{K_{\mathrm{R}}-j} w_{i,j}} \cdot \sqrt{\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} \frac{K_{\mathrm{R}}-j}{j+1} w_{i,j}}. \tag{C.7}$$

Moreover, from (C.4) and (C.5) we know that

$$\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} \frac{j+1}{K_{\mathrm{R}}-j} w_{i,j} \leq K_{\mathrm{R}} \mu_{\mathrm{R}} NF + NF. \tag{C.8}$$

It follows that

$$\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} w_{i,j} \leq \sqrt{K_{\mathrm{R}} \mu_{\mathrm{R}} NF + NF} \cdot \sqrt{\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} \frac{K_{\mathrm{R}}-j}{j+1} w_{i,j}}. \tag{C.9}$$

Furthermore, from [31] we know that

$$\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} w_{i,j} \geq K_{\mathrm{R}} N (1-\mu_{\mathrm{R}}) F. \tag{C.10}$$

Hence, it follows that

$$\begin{aligned} \overline{H^{(\mathrm{n})}}^{\left(\{\mathcal{P}_i\}_{i=1}^{\mathrm{K_T}}, \{\mathcal{U}_i\}_{i=1}^{\mathrm{K}_{\mathcal{T}}}, q\right)} &\geq \frac{1}{N} \sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} \frac{w_{i,j}}{j+1} \geq \frac{1}{K_{\mathrm{R}} N} \sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} \frac{K_{\mathrm{R}}-j}{j+1} w_{i,j} \\ &\geq \frac{1}{K_{\mathrm{R}} N} \cdot \frac{1}{K_{\mathrm{R}} \mu_{\mathrm{R}} NF + NF} \left(\sum_{i=1}^{K_{\mathrm{T}}} \sum_{j=0}^{K_{\mathrm{R}}-1} w_{i,j} \right)^2 \\ &\geq \frac{K_{\mathrm{R}} NF \left(1-\mu_{\mathrm{R}}\right)^2}{K_{\mathrm{R}} \mu_{\mathrm{R}} N + N} = \frac{K_{\mathrm{R}} F \left(1-\mu_{\mathrm{R}}\right)^2}{1 + K_{\mathrm{R}} \mu_{\mathrm{R}}} \end{aligned} \tag{C.11}$$

for any caching realization $\left(\{\mathcal{P}_i\}_{i=1}^{\mathrm{K_T}}, \{\mathcal{U}_i\}_{i=1}^{\mathrm{K}_{\mathcal{T}}}\right)$, which concludes the proof.

C.2. Proof of (5.27)

First, we will briefly revisit parts of the derivation in [31, Lem. 3] which in turn allows to proof Eq. (5.26). Next, we will describe the extra steps needed in order to prove (5.27). Suppose that in a certain N-block $L^{(\mathrm{n})}$ N-subpackets $\{\mathbf{w}_{1,1}^{(\mathrm{n})}, \mathbf{w}_{2,1}^{(\mathrm{n})}, \cdots, \mathbf{w}_{L,1}^{(\mathrm{n})}\}$ are scheduled to be communicated to L receivers $\{\mathrm{Rx}_1, \mathrm{Rx}_2, \ldots, \mathrm{Rx}_L\}$, respectively. As we are considering a specific N-block, we index the channel uses starting from 1 as $t = 1, 2, \cdots, \tilde{B}^{(\mathrm{n})}$. Each transmitter transmits a linear combination of the scheduled N-subpackets which are cached in its memory. From Section 5.3 each transmitter Tx_i, with $i \in [K_{\mathrm{T}}]$, will transmit the symbols given by

$$X_i^{(\mathrm{n})}(t) = \sum_{l\colon i \in \mathcal{T}_l} v_{i,l,1}^{(\mathrm{n})}(t) \tilde{W}_{l,1}^{(\mathrm{n})}(t).$$

where $\tilde{\mathbf{w}}_{l,1}^{(\mathrm{n})} = \big(\tilde{W}_{l,1}^{(\mathrm{n})}(1), \ldots, \tilde{W}_{l,1}^{(\mathrm{n})}(\tilde{B}^{(\mathrm{n})})\big)$ as in Eq. (5.4). Note that the beamforming coefficients of the N-subchannel cannot depend on the fading channel.

We can then write the received signal at Rx_j as

$$Y_j^{(\mathrm{n})}(t) = \sqrt{P^{\bar{\beta}}} \sum_{l=1}^{L} \sum_{i \in \mathcal{T}_l} G_{ji}^{(\mathrm{n})}(t) v_{i,l,1}^{(\mathrm{n})}(t) \tilde{W}_{l,1}^{(\mathrm{n})}(t) + Z^{(\mathrm{n})}(t). \tag{C.12}$$

Therefore, by applying the approach in [31, Lem. 3], we can convert the network as a new MISO interference channel (MISO IC) with L virtual transmitters $\{\widehat{\mathrm{Tx}}_l\}_{l=1}^{L}$, where each virtual transmitter has $|\mathcal{T}_l|$ antennas, and L single-antenna receivers $\{\mathrm{Rx}_j\}_{j=1}^{L}$, where the latter correspond to the receivers scheduled in such N-block. In this network $\widehat{\mathrm{Tx}}_l$ wants to deliver the N-subpacket $\mathbf{w}_{l,1}^{(\mathrm{n})}$ to Rx_l. Note that each antenna in the new network corresponds to a transmitter in the original network. In particular, each antenna of $\widehat{\mathrm{Tx}}_l$ corresponds to a transmitter which caches the N-subpacket $\mathbf{w}_{l,1}^{(\mathrm{n})}$ in its memory. This in turn implies that the channel vectors of the transmitters in the new MISO IC are correlated.

To bound the sum-DoF of this network we take the same approach of [31] which in turn is derived from [135]. Each virtual transmitter $\widehat{\mathrm{Tx}}_l$ in the constructed MISO BC will choose a beamforming vector $\mathbf{v}_l^{(\mathrm{n})}(t)$ to precode $\tilde{W}_{l,1}^{(\mathrm{n})}(t)$. Note that $\mathbf{v}_l^{(\mathrm{n})}(t) \in \mathbb{C}^{|\mathcal{T}_l| \times 1}$ consists of the set of complex beamforming coefficients $v_{i,l,1}^{(\mathrm{n})}(t)$, with $i \in \mathcal{T}_l$, chosen by the original transmitters corresponding to its antennas. We also denote the channel between Rx_j and $\widehat{\mathrm{Tx}}_l$ as the fading channel vector $\mathbf{g}_{jl}^{(\mathrm{n})}(t) \in \mathbb{C}^{|\mathcal{T}_l| \times 1}(t)$, which consists of the set of fading channel coefficients from the original transmitters corresponding to the antennas of $\widehat{\mathrm{Tx}}_l$ to Rx_j. The decodability conditions become then

$$\mathbf{g}_{jl}^{(\mathrm{n}),T}(t) \mathbf{v}_l^{(\mathrm{n})}(t) = 0, \quad \forall l \neq j \quad \text{s.t.} \quad j \notin \mathcal{R}_l \tag{C.13}$$

$$\mathbf{g}_{jj}^{(\mathrm{n}),T}(t) \mathbf{v}_j^{(\mathrm{n})}(t) \neq 0, \quad \forall j \in [L]. \tag{C.14}$$

Given that we are considering the N-subchannel, the beamforming vectors $\mathbf{v}_l^{(\mathrm{n})}(t)$ are independent from the channel vector coefficients $\mathbf{g}_{jl}^{(\mathrm{n})}(t)$. However, in order to remove interference in (C.13), the vector $\mathbf{v}_l^{(\mathrm{n})}(t)$ has still to belong to the null-space of $\mathbf{g}_{jl}^{(\mathrm{n})}(t)$, with $j \notin \mathcal{R}_l$. As the vector $\mathbf{g}_{jl}^{(\mathrm{n})}(t)$ has dimension $|\mathcal{T}_l|$, the null-space of $\mathbf{g}_{jl}^{(\mathrm{n})}(t)$ is a subspace of dimension $|\mathcal{T}_l| - 1$. As any beamforming vector $\mathbf{v}_l^{(\mathrm{n})}(t)$ is independent of $\mathbf{g}_{jl}^{(\mathrm{n})}(t)$, the probability of $\mathbf{v}_l^{(\mathrm{n})}(t)$ to belong to the null-space of $\mathbf{g}_{jl}^{(\mathrm{n})}(t)$ is almost surely 0. It follows that the condition in (C.13) cannot be satisfied almost surely. As a consequence, each N-subpacket scheduled to be delivered to a specific user has to be available in the caches of all the other users in order for them to remove it from the received signal. It follows that the condition $L \leq |\mathcal{R}_l| + 1$ has to be satisfied for all the scheduled N-subpackets $\mathbf{w}_{l,1}^{(\mathrm{n})}$, from which the result in Eq. (5.27) follows.

C.3. Proof of Lemma 5.1

Here we present a proof of the inequality in (5.50). We start with the following instrumental lemma.

Lemma C.1. *Consider a polynomial $\phi(\zeta) = \sum_{m=0}^{d} a_m \zeta^m$ for which there exists and integer N in $[-1, d]_{\mathbb{Z}}$ such that the coefficients of $\phi(\zeta)$ satisfy the following condition*

$$\begin{aligned} a_m &\geq 0, \quad m < N \\ a_m &> 0, \quad m = N \\ a_m &\leq 0, \quad m > N \end{aligned} \tag{C.15}$$

where the case $N = -1$ implies $a_0, \ldots, a_d \leq 0$. The polynomial $\phi(\zeta)$ is quasiconcave on $\zeta \in [0, \infty)$.

Proof. First, we note that for the cases: $N = -1$ (i.e. when $a_m \leq 0$ for all m), $N = 0$ and $N = 1$, the second derivative of $\phi(\zeta)$ is a polynomial with all coefficients not greater than zero. Therefore, $\phi(\zeta)$ is concave, and hence quasiconcave, on $\zeta \in [0, \infty)$. We proceed by induction. In particular, assume that the quasiconcavity hypothesis holds for all polynomials the satisfy the condition in (C.15) for integer $N = n$, where $n \geq 1$. Now consider a polynomial $\phi(\zeta)$ that satisfies the condition in (C.15) for $N = n + 1$. It is readily seen that the first derivative of $\phi(\zeta)$, denoted by $\phi'(\zeta)$, is a polynomial which satisfies the condition in (C.15) for $N = n$. Hence, $\phi'(\zeta)$ is quasiconcave by the induction hypothesis. Moreover, as $n \geq 1$, it follows from (C.15) that $\phi'(0) \geq 0$. It can be verified that $\phi'(0) \geq 0$ combined with the quasiconcavity of $\phi'(\zeta)$ guarantee that: either $\phi'(\zeta)$ is non-negative over $[0, \infty)$, or there exists $\zeta' \in [0, \infty)$ such that $\phi'(\zeta) \geq 0$ over the interval $[0, \zeta']$ and $\phi'(\zeta) \leq 0$ over the interval $[\zeta', \infty)$. It follows that $\phi(\zeta)$ is eithrer non-decreasing over $[0, \infty)$, or non-decreasing over $[0, \zeta']$ and non-increasing over $[\zeta', \infty)$. In both cases, $\phi(\zeta)$ is quasiconcave. This concludes the proof of Lemma C.1. □

Next, we show that the coefficients of the polynomial $p(\zeta)$ of interest satisfy the conditions in Lemma C.1. As this shows that $p(\zeta)$ is quasiconcave, the inequality in (5.50) directly follows by definition. The remainder of this appendix is dedicated to showing that $p(\zeta)$ is an instance of Lemma C.1.

The key step of this proof is to show that the sequence $\left\{ \frac{c_m}{\binom{K_\mathrm{R}-1}{m-1}} \right\}_{m=1}^{K_\mathrm{R}-1}$ is non-increasing. Supposing that this holds true, then this sequence would satisfy the condition of Lemma C.1, applied only to the indices $m \in [1, K_\mathrm{R} - 1]_{\mathbb{Z}}$. Since the sign of $\frac{c_m}{\binom{K_\mathrm{R}-1}{m-1}}$ is preserved by c_m, then $\{c_m\}_{m=1}^{K_\mathrm{R}-1}$ also satisfies the condition of Lemma C.1 over $m \in [1, K_\mathrm{R} - 1]_{\mathbb{Z}}$. Combining this with $c_0 = 0$ and $c_{K_\mathrm{R}} \leq 0$, it follows that $\{c_m\}_{m=0}^{K_\mathrm{R}}$ satisfies the condition of Lemma C.1, which in turn concludes the proof. Therefore, our problem reduces to showing that $\frac{c_m}{\binom{K_\mathrm{R}-1}{m-1}}$ in a non-increasing over $m \in [1, K_\mathrm{R} - 1]_{\mathbb{Z}}$.

First, it is readily seen that c_m can be written as

$$c_m = \binom{K_{\mathrm{R}}-1}{m-1} \left[\left(\frac{K_{\mathrm{R}}+1}{m} - \frac{K_{\mathrm{R}}+r}{\min\{r+m-1, K_{\mathrm{R}}\}} \right) + \frac{K_{\mathrm{R}}-m}{m} \left(\frac{1}{m+1} - \frac{r}{\min\{r+m, K_{\mathrm{R}}\}} \right) \right].$$

For briefness, we denote the coefficient $\frac{c_m}{\binom{K_{\mathrm{R}}-1}{m-1}}$ as c'_m. Hence, c'_m is given by

$$c'_m = \left(\frac{K_{\mathrm{R}}+1}{m} - \frac{K_{\mathrm{R}}+r}{\min\{r+m-1, K_{\mathrm{R}}\}} \right) + \frac{K_{\mathrm{R}}-m}{m} \left(\frac{1}{m+1} - \frac{r}{\min\{r+m, K_{\mathrm{R}}\}} \right).$$

Next, let us define the integer $\tilde{r} \in [1, K_{\mathrm{R}}-1]_{\mathbb{Z}}$ as $\tilde{r} \triangleq \lfloor r \rfloor = r - \epsilon$, where $\epsilon \in [0, 1)$. Using this definition, it can be shown that c'_m, $m \in [1, K_{\mathrm{R}}-1]_{\mathbb{Z}}$, may be expressed as:

$$c'_m = \begin{cases} d_m \triangleq \left(\frac{K_{\mathrm{R}}+1}{m} - \frac{K_{\mathrm{R}}+r}{r+m-1} \right) + \frac{K_{\mathrm{R}}-m}{m} \left(\frac{1}{m+1} - \frac{r}{r+m} \right), & m \in [1, K_{\mathrm{R}}-\tilde{r}-1]_{\mathbb{Z}} \\ \left(\frac{K_{\mathrm{R}}+1}{K_{\mathrm{R}}-\tilde{r}} - \frac{K_{\mathrm{R}}+\tilde{r}+\epsilon}{K_{\mathrm{R}}+\epsilon-1} \right) + \frac{\tilde{r}}{K_{\mathrm{R}}-\tilde{r}} \left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}} \right), & m = K_{\mathrm{R}}-\tilde{r} \\ e_m \triangleq \left(\frac{K_{\mathrm{R}}+1}{m} - \frac{K_{\mathrm{R}}+r}{K_{\mathrm{R}}} \right) + \frac{K_{\mathrm{R}}-m}{m} \left(\frac{1}{m+1} - \frac{r}{K_{\mathrm{R}}} \right), & m \in [K_{\mathrm{R}}-\tilde{r}+1, K_{\mathrm{R}}-1]_{\mathbb{Z}}. \end{cases}$$

Showing that c'_m is non-increasing in m is carried out through the two following steps:

1. We show that d_m and e_m are both non-increasing sequences in m . This guarantees that c'_m is non-increasing over both the intervals $[1, K_{\mathrm{R}}-\tilde{r}-1]_{\mathbb{Z}}$ and $[K_{\mathrm{R}}-\tilde{r}+1, K_{\mathrm{R}}-1]_{\mathbb{Z}}$.

2. We show that $c'_{K_{\mathrm{R}}-\tilde{r}} \leq d_{K_{\mathrm{R}}-\tilde{r}-1}$ and $c'_{K_{\mathrm{R}}-\tilde{r}} \geq e_{K_{\mathrm{R}}-\tilde{r}+1}$. This guarantees that c'_m is non-increasing over the entire interval $[1, K_{\mathrm{R}}-1]_{\mathbb{Z}}$.

Proof of Point 1): First, let us consider d_m. This can be rewritten as:

$$d_m = \frac{(K_{\mathrm{R}}-m+1)(r-1)}{m(m+r-1)} + \frac{(K_{\mathrm{R}}-m)(1-r)}{(m+1)(m+r)}. \tag{C.16}$$

For $r = 1$, we have $d_m = 0$ for all $m \in [1, K_{\mathrm{R}}-1]_{\mathbb{Z}}$. Hence, we consider $r \geq 1$. From (C.16), and after some rearrangements, the inequality $d_m \geq d_{m+1}$ which we wish to prove is equivalently written as

$$\frac{K_{\mathrm{R}}-m+1}{m(m+r-1)} - \frac{K_{\mathrm{R}}-m}{(m+1)(m+r)} \geq \frac{K_{\mathrm{R}}-m}{(m+1)(m+r)} - \frac{K_{\mathrm{R}}-m-1}{(m+2)(m+r+1)}. \tag{C.17}$$

Using the following notation $A = K_{\mathrm{R}} - m$, $B = m+1$ and $C = m+r$, (C.17) is rewritten as

$$\frac{A+1}{(B-1)(C-1)} - \frac{A}{BC} \geq \frac{A}{BC} - \frac{A-1}{(B+1)(C+1)}. \tag{C.18}$$

After further rearranging and simplifying, (C.18) becomes

$$ABC + B^2C + BC^2 \geq A - AB^2 - AC^2. \tag{C.19}$$

Since $A \geq 1, B \geq 2$ and $C \geq 2$, (C.19) always holds and hence d_m is non-increasing in m.

Next, we consider e_m. This can be rewritten as:

$$e_m = \frac{K_{\mathrm{R}}+1}{m} + \frac{K_{\mathrm{R}}}{m(m+1)} - \frac{1}{m+1} - \frac{r}{m} - 1 \tag{C.20}$$

From (C.20), it follows that $e_m \geq e_{m+1}$ is implied by

$$\frac{K_{\mathrm{R}}+1}{m} + \frac{K_{\mathrm{R}}}{m(m+1)} - \frac{1}{m+1} - \frac{r}{m} \geq \frac{K_{\mathrm{R}}+1}{m+1} + \frac{K_{\mathrm{R}}}{(m+1)(m+2)} - \frac{1}{m+2} - \frac{r}{m+1}. \tag{C.21}$$

After some rearrangements, the inequality in (C.21) becomes

$$(K_{\mathrm{R}}+1-r)(m+2) + 2K_{\mathrm{R}} - m \geq 0 \tag{C.22}$$

which holds as $m \geq 1$ and $K_{\mathrm{R}} \geq r$. Hence, e_m is a non-increasing in m and this part is complete.

Proof of Point 2): In order to show that $c'_{K_{\mathrm{R}}-\tilde{r}} \leq d_{K_{\mathrm{R}}-\tilde{r}-1}$, we only need to observe the following:

$$\begin{aligned} c'_{K_{\mathrm{R}}-\tilde{r}} &= \left(\frac{K_{\mathrm{R}}+1}{K_{\mathrm{R}}-\tilde{r}} - \frac{K_{\mathrm{R}}+\tilde{r}+\epsilon}{K_{\mathrm{R}}+\epsilon-1}\right) + \frac{\tilde{r}}{K_{\mathrm{R}}-\tilde{r}}\left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right) \\ &\leq \left(\frac{K_{\mathrm{R}}+1}{K_{\mathrm{R}}-\tilde{r}} - \frac{K_{\mathrm{R}}+\tilde{r}+\epsilon}{K_{\mathrm{R}}+\epsilon-1}\right) + \frac{\tilde{r}}{K_{\mathrm{R}}-\tilde{r}}\left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}+\epsilon}\right) \\ &= d_{K_{\mathrm{R}}-\tilde{r}} \\ &\leq d_{K_{\mathrm{R}}-\tilde{r}-1}. \end{aligned}$$

Next, we focus on showing that $c'_{K_{\mathrm{R}}-\tilde{r}} \geq e_{K_{\mathrm{R}}-\tilde{r}+1}$. We observe that $c'_{K_{\mathrm{R}}-\tilde{r}}$ can be expressed as:

$$\begin{aligned} c'_{K_{\mathrm{R}}-\tilde{r}} &= \left(\frac{K_{\mathrm{R}}+1}{K_{\mathrm{R}}-\tilde{r}} - \frac{K_{\mathrm{R}}+\tilde{r}+\epsilon}{K_{\mathrm{R}}+\epsilon-1}\right) + \frac{\tilde{r}}{K_{\mathrm{R}}-\tilde{r}}\left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right) \\ &= \left(\frac{\tilde{r}+1}{K_{\mathrm{R}}-\tilde{r}} - \frac{\tilde{r}+1}{K_{\mathrm{R}}+\epsilon-1}\right) + \frac{\tilde{r}}{K_{\mathrm{R}}-\tilde{r}}\left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right). \end{aligned} \tag{C.23}$$

On the other hand, $e_{K_{\mathrm{R}}-\tilde{r}+1}$ is given by:

$$\begin{aligned} e_{K_{\mathrm{R}}-\tilde{r}+1} &= \left(\frac{K_{\mathrm{R}}+1}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{K_{\mathrm{R}}+\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right) + \frac{\tilde{r}-1}{K_{\mathrm{R}}-\tilde{r}+1}\left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+2} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right) \\ &= \left(\frac{\tilde{r}}{K_{\mathrm{R}}-\tilde{r}+1} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right) + \frac{\tilde{r}-1}{K_{\mathrm{R}}-\tilde{r}+1}\left(\frac{1}{K_{\mathrm{R}}-\tilde{r}+2} - \frac{\tilde{r}+\epsilon}{K_{\mathrm{R}}}\right). \end{aligned} \tag{C.24}$$

By taking the difference of (C.23) and (C.24), we obtain

$$\begin{aligned} c'_{K_{\mathrm{R}}-\tilde{r}} - e_{K_{\mathrm{R}}-\tilde{r}+1} = \frac{K_{\mathrm{R}}+1-\tilde{r}-\epsilon}{(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+1)} &+ \frac{(\epsilon-1)(K_{\mathrm{R}}+\tilde{r}+\epsilon)}{K_{\mathrm{R}}(K_{\mathrm{R}}+\epsilon-1)} \\ &+ \frac{K_{\mathrm{R}}+\tilde{r}}{(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+1)(K_{\mathrm{R}}-\tilde{r}+2)}. \end{aligned} \tag{C.25}$$

After rearranging the terms in (C.25), it follows that $c'_{K_{\mathrm{R}}-\tilde{r}} - e_{K_{\mathrm{R}}-\tilde{r}+1} \geq 0$ is implied by the inequality

$$\underbrace{K_{\mathrm{R}}(K_{\mathrm{R}}+\epsilon-1)(K_{\mathrm{R}}+1-\tilde{r}-\epsilon)(K_{\mathrm{R}}-\tilde{r}+2)}_{l_1(\epsilon)}+\underbrace{K_{\mathrm{R}}(K_{\mathrm{R}}+\epsilon-1)(K_{\mathrm{R}}+\tilde{r})}_{l_2(\epsilon)}$$
$$+\underbrace{(\epsilon-1)(K_{\mathrm{R}}+\epsilon+\tilde{r})(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+1)(K_{\mathrm{R}}-\tilde{r}+2)}_{l_3(\epsilon)}\geq 0. \quad \text{(C.26)}$$

We denote the left-hand side of (C.26) by $l(\epsilon)=l_1(\epsilon)+l_2(\epsilon)+l_3(\epsilon)$. It is readily seen that $l_1(\epsilon)$ and $l_3(\epsilon)$ are second degree polynomials in the variable ϵ (i.e. parabolas). We consider the the three functions separately in order to derive a lower-bound on $l(\epsilon)$.

- $l_1(\epsilon)$: It can be easily verified that $l_1(\epsilon)$ is concave with a maximum value at $\epsilon^*=\frac{2-\tilde{r}}{2}$. Hence, $\epsilon^*\leq 0$ for $\tilde{r}\geq 2$ and $\epsilon^*=1/2$ for $\tilde{r}=1$. As a concave parabola is decreasing for $\epsilon\geq\epsilon^*$ and symmetric with respect to the maximum, it follows that for $\epsilon\in[0,1)$, we have

$$l_1(\epsilon)\geq l_1(1)=K_{\mathrm{R}}^2(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+2). \quad \text{(C.27)}$$

- $l_2(\epsilon)$: It is readily seen that for $\epsilon\in[0,1)$, the following holds

$$l_2(\epsilon)\geq l_2(0)=K_{\mathrm{R}}(K_{\mathrm{R}}-1)(K_{\mathrm{R}}+\tilde{r}). \quad \text{(C.28)}$$

- $l_3(\epsilon)$: This is a convex with a minimum value at $\epsilon^*=\frac{-K_{\mathrm{R}}-\tilde{r}+1}{2}<0$. Hence, for $\epsilon\in[0,1)$, we have

$$l_3(\epsilon)\geq l_3(0)=-(K_{\mathrm{R}}+\tilde{r})(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+1)(K_{\mathrm{R}}-\tilde{r}+2). \quad \text{(C.29)}$$

By summing over the lower-bounds in (C.27), (C.28) and (C.29), it follows that for $\epsilon\in[0,1)$, we have:

$$l(\epsilon)\geq K_{\mathrm{R}}(K_{\mathrm{R}}-1)(K_{\mathrm{R}}+\tilde{r})+(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+2)(\tilde{r}^2-\tilde{r}-K_{\mathrm{R}}). \quad \text{(C.30)}$$

Next, we express the right-hand side of the (C.30) as a function of K_{R}:

$$g(K_{\mathrm{R}})=K_{\mathrm{R}}(K_{\mathrm{R}}-1)(K_{\mathrm{R}}+\tilde{r})+(K_{\mathrm{R}}-\tilde{r})(K_{\mathrm{R}}-\tilde{r}+2)(\tilde{r}^2-\tilde{r}-K_{\mathrm{R}})=aK_{\mathrm{R}}^2+bK_{\mathrm{R}}+c,$$

where $a=\tilde{r}^2+2\tilde{r}-3$ and $b=-\tilde{r}(2\tilde{r}^2-3\tilde{r}+1)$. Finally, to show that $l(\epsilon)\geq 0$, it is sufficient to show $g(K_{\mathrm{R}})\geq 0$ for all $K_{\mathrm{R}}\geq\tilde{r}$. To this end, we observe that $g(K_{\mathrm{R}})=0$ for $\tilde{r}=1$, while $g(K_{\mathrm{R}})$ is a convex parabola with a minimum value at $\frac{\tilde{r}(\tilde{r}-1/2)}{\tilde{r}+3}\leq\tilde{r}$ for $\tilde{r}>1$. In latter case, $g(K_{\mathrm{R}})$ is increasing for $K_{\mathrm{R}}\geq\tilde{r}$. As $g(\tilde{r})\geq 0$, it follows that $g(K_{\mathrm{R}})\geq 0$ for all $K_{\mathrm{R}}\geq\tilde{r}$. This concludes the proof.

Bibliography

[1] NGMN Alliance, “5G white paper,” Feb. 2015.

[2] A. Gupta and R. K. Jha, “A Survey of 5G Network: Architecture and Emerging Technologies,” *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[3] M. Agiwal, A. Roy, and N. Saxena, “Next Generation 5G Wireless Networks: A Comprehensive Survey,” *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

[4] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5G,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[5] B. Clerckx and C. Oestges, *MIMO Wireless Networks: Channels, Techniques and Standards for Multi-antenna, Multi-user and Multi-cell Systems*. Academic Press, 2013.

[6] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge University Press, 2012.

[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[8] E. Telatar, “Capacity of Multi-antenna Gaussian Channels,” *Transactions on Emerging Telecommunications Technologies*, vol. 10, no. 6, pp. 585–595, 1999.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[10] G. Caire and S. Shamai, “On the achievable throughput of a multiantenna Gaussian broadcast channel,” *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[11] L. Zheng and D. N. C. Tse, “Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[12] H. Weingarten, Y. Steinberg, and S. S. Shamai, “The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel,” *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept 2006.

[13] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, “Degrees of Freedom of Time Correlated MISO Broadcast Channel With Delayed CSIT,” *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 315–328, Jan 2013.

[14] C. Hao, B. Rassouli, and B. Clerckx, “Achievable DoF Regions of MIMO Networks With Imperfect CSIT,” *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6587–6606, Oct 2017.

[15] C. Hao and B. Clerckx, “MISO Networks With Imperfect CSIT: A Topological Rate-Splitting Approach,” *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2164–2179, May 2017.

[16] E. Piovano and B. Clerckx, “Optimal DoF Region of the K-User MISO BC With Partial CSIT,” *IEEE Communications Letters*, vol. 21, no. 11, pp. 2368–2371, Nov 2017.

[17] E. Piovano, H. Joudeh, and B. Clerckx, “Overloaded multiuser MISO transmission with imperfect CSIT,” in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 34–38.

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

[19] N. Jindal, “MIMO Broadcast Channels With Finite-Rate Feedback,” *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, Nov 2006.

[20] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, “An overview of limited feedback in wireless communication systems,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341–1365, October 2008.

[21] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, “Multiuser MIMO Achievable Rates With Downlink Training and Channel State Feedback,” *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2845–2866, Jun 2010.

[22] A. G. Davoodi and S. A. Jafar, “Aligned Image Sets Under Channel Uncertainty: Settling Conjectures on the Collapse of Degrees of Freedom Under Finite Precision CSIT,” *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5603–5618, Oct. 2016.

[23] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, “Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution,” *IEEE Communications Magazine*, vol. 54, no. 5, pp. 98–105, May 2016.

[24] M. A. Maddah-Ali and U. Niesen, “Fundamental Limits of Caching,” *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[25] ——, “Cache-aided interference channels,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2015, pp. 809–813.

[26] ——, “Coding for caching: fundamental limits and practical challenges,” *IEEE Communications Magazine*, vol. 54, no. 8, pp. 23–29, Aug. 2016.

[27] M. Ji, G. Caire, and A. F. Molisch, “Fundamental Limits of Caching in Wireless D2D Networks,” *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, Feb 2016.

[28] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, “Femtocaching: Wireless video content delivery through distributed caching helpers,” in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 1107–1115.

[29] M. Gregori, J. Gmez-Vilardeb, J. Matamoros, and D. Gündüz, “Wireless Content Caching for Small Cell and D2D Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[30] M. A. Maddah-Ali and U. Niesen, “Cache-Aided Interference Channels,” *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1714–1724, March 2019.

[31] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Fundamental Limits of Cache-Aided Interference Management,” *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[32] J. Zhang and P. Elia, “Fundamental Limits of Cache-Aided Wireless BC: Interplay of Coded-Caching and CSIT Feedback,” *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.

[33] S. Yang, K. H. Ngo, and M. Kobayashi, “Content delivery with coded caching and massive MIMO in 5G,” *9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, pp. 370–374, Sept 2016.

[34] J. Hachem, U. Niesen, and S. N. Diggavi, “Degrees of Freedom of Cache-Aided Wireless Interference Networks,” *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5359–5380, Jul. 2018.

[35] X. Yi and G. Caire, “Topological coded caching,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2039–2043.

[36] E. Lampiris, J. Zhang, and P. Elia, “Cache-aided cooperation with no CSIT,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 2960–2964.

[37] E. Piovano, H. Joudeh, and B. Clerckx, “On coded caching in the overloaded MISO broadcast channel,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 2795–2799.

[38] ——, “Centralized and Decentralized Cache-Aided Interference Management in Heterogeneous Parallel Channels,” *arXiv:1812.01469*, 2018.

[39] E. Piovano, H. Joudeh, and B. Clerckx, "Robust Cache-Aided Interference Management Under Full Transmitter Cooperation," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 1540–1544.

[40] M. A. Maddah-Ali and D. Tse, "Completely Stale Transmitter Channel State Information is Still Very Useful," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4418–4431, July 2012.

[41] A. G. Davoodi and S. A. Jafar, "Transmitter Cooperation Under Finite Precision CSIT: A GDoF Perspective," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 6020–6030, Sep. 2017.

[42] ——, "Generalized Degrees of Freedom of the Symmetric K User Interference Channel Under Finite Precision CSIT," *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6561–6572, Oct. 2017.

[43] A. G. Davoodi, B. Yuan, and S. A. Jafar, "GDoF Region of the MISO BC: Bridging the Gap Between Finite Precision and Perfect CSIT," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 7208–7217, Nov. 2018.

[44] C. Hao, Y. Wu, and B. Clerckx, "Rate Analysis of Two-Receiver MISO Broadcast Channel With Finite Rate Feedback: A Rate-Splitting Approach," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3232–3246, Sep 2015.

[45] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.

[46] ——, "Robust Transmission in Downlink Multiuser MISO Systems: A Rate-Splitting Approach," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6227–6242, Dec. 2016.

[47] B. Rassouli, C. Hao, and B. Clerckx, "DoF Analysis of the MIMO Broadcast Channel With Alternating/Hybrid CSIT," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1312–1325, March 2016.

[48] T. Gou and S. A. Jafar, "Optimal Use of Current and Outdated Channel State Information: Degrees of Freedom of the MISO BC with Mixed CSIT," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1084–1087, July 2012.

[49] A. Lapidoth, S. Shamai, and M. A. Wigger, "On the capacity of fading MIMO broadcast channels with imperfect transmitter side-information," in *43rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2005.

[50] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, 1972.

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

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

[53] Wei Yu and J. M. Cioffi, “Sum capacity of Gaussian vector broadcast channels,” *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.

[54] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, “Multiple-antenna channel hardening and its implications for rate feedback and scheduling,” *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1893–1909, Sep. 2004.

[55] B. Hassibi and M. Sharif, “Fundamental Limits in Mimo Broadcast Channels,” *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1333–1344, Sep. 2007.

[56] N. Jindal, S. Vishwanath, and A. Goldsmith, “On the duality of gaussian multiple-access and broadcast channels,” *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

[57] T. M. Cover, “Comments on broadcast channels,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2524–2530, Oct 1998.

[58] P. Bergmans, “Random coding theorem for broadcast channels with degraded components,” *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 197–207, 1973.

[59] M. Costa, “Writing on dirty paper,” *IEEE transactions on information theory*, vol. 29, no. 3, pp. 439–441, 1983.

[60] A. Carleial, “Interference channels,” *IEEE Transactions on Information Theory*, vol. 24, no. 1, pp. 60–70, 1978.

[61] T. Han and K. Kobayashi, “A new achievable rate region for the interference channel,” *IEEE transactions on information theory*, vol. 27, no. 1, pp. 49–60, 1981.

[62] C. E. Shannon *et al.*, “Two-way communication channels,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[63] R. Ahlswede *et al.*, “The capacity region of a channel with two senders and two receivers,” *The annals of probability*, vol. 2, no. 5, pp. 805–814, 1974.

[64] A. Gamal and M. Costa, “The capacity region of a class of deterministic interference channels,” *IEEE Transactions on information Theory*, vol. 28, no. 2, pp. 343–346, 1982.

[65] H. Chong, M. Motani, H. K. Garg, and H. El Gamal, "On The HanKobayashi Region for the Interference Channel," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3188–3195, July 2008.

[66] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian Interference Channel Capacity to Within One Bit," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5534–5562, Dec. 2008.

[67] S. Shamai, "A broadcast strategy for the Gaussian slowly fading channel," in *Proceedings of IEEE International Symposium on Information Theory*, June 1997, pp. 150–.

[68] S. Shamai and A. Steiner, "A broadcast approach for a single-user slowly fading MIMO channel," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2617–2635, Oct 2003.

[69] A. Steiner and S. Shamai, "Achievable Rates with Imperfect Transmitter Side Information Using a Broadcast Transmission Strategy," *IEEE Transactions on Wireless Communications*, vol. 7, no. 3, pp. 1043–1051, March 2008.

[70] ——, "Multi-Layer Broadcasting over a Block Fading MIMO Channel," *IEEE Transactions on Wireless Communications*, vol. 6, no. 11, pp. 3937–3945, November 2007.

[71] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP journal on wireless communications and networking*, vol. 2018, no. 1, p. 133, 2018.

[72] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized Degrees of Freedom of the Symmetric Cache-Aided MISO Broadcast Channel With Partial CSIT," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5799–5815, Sep. 2019.

[73] H. Joudeh and B. Clerckx, "On the DoF of Parallel MISO BCs with Partial CSIT: Total Order and Separability," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[74] W. Santipach and M. L. Honig, "Asymptotic capacity of beamforming with limited feedback," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, June 2004, pp. 290–.

[75] ——, "Signature optimization for CDMA with limited feedback," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3475–3492, Oct 2005.

[76] Q. Yan, X. Tang, and Q. Chen, "On the gap between decentralized and centralized coded caching schemes," *arXiv:1605.04626*, 2016.

[77] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless Content Delivery Through Distributed Caching Helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[78] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 16–22, August 2016.

[79] M. A. Maddah-Ali and U. Niesen, "Decentralized Coded Caching Attains Order-Optimal Memory-Rate Tradeoff," *IEEE/ACM Transaction on Networking*, vol. 23, no. 4, pp. 1029–1040, Aug 2015.

[80] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for users with small buffers," *IET Communications*, vol. 10, no. 17, pp. 2315–2318, 2016.

[81] M. M. Amiri, Q. Yang, and D. Gündüz, "Coded caching for a large number of users," in *2016 IEEE Information Theory Workshop (ITW)*, Sep. 2016, pp. 171–175.

[82] S. Sahraei and M. Gastpar, "K users caching two files: An improved achievable rate," in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 620–624.

[83] M. Mohammadi Amiri and D. Gündüz, "Fundamental Limits of Coded Caching: Improved Delivery Rate-Cache Capacity Tradeoff," *IEEE Transactions on Communications*, vol. 65, no. 2, pp. 806–815, Feb 2017.

[84] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 135–139.

[85] C. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," in *2016 Information Theory and Applications Workshop (ITA)*, Jan 2016, pp. 1–6.

[86] H. Ghasemi and A. Ramamoorthy, "Improved Lower Bounds for Coded Caching," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4388–4413, Jul. 2017.

[87] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1691–1695.

[88] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The Exact Rate-Memory Tradeoff for Caching With Uncoded Prefetching," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1281–1296, Feb. 2018.

[89] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the Rate-Memory Tradeoff in Cache Networks Within a Factor of 2," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 647–663, Jan 2019.

[90] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2014, pp. 221–226.

[91] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "On the average performance of caching and coded multicasting with random demands," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 922–926.

[92] J. Llorca and A. M. Tulino, "Minimum cost caching-aided multicast under arbitrary demand," in *2013 Asilomar Conference on Signals, Systems and Computers*, Nov 2013, pp. 236–237.

[93] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 349–366, 2017.

[94] M. M. Amiri, Q. Yang, and D. Gündüz, "Decentralized coded caching with distinct cache capacities," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 734–738.

[95] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Caching and coded multicasting: Multiple groupcast index coding," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 881–885.

[96] M. Mohammadi Amiri, Q. Yang, and D. Gündüz, "Decentralized Caching and Coded Delivery With Distinct Cache Capacities," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4657–4669, Nov 2017.

[97] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical Coded Caching," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3212–3229, June 2016.

[98] J. Hachem, N. Karamchandani, and S. N. Diggavi, "Coded Caching for Multi-level Popularity and Access," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3108–3141, May 2017.

[99] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 56–60.

[100] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-Server Coded Caching," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[101] A. Ghorbel, M. Kobayashi, and S. Yang, "Content Delivery in Erasure Broadcast Channels With Cache and Feedback," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6407–6422, Nov. 2016.

[102] M. M. Amiri and D. Gündüz, "Cache-Aided Content Delivery Over Erasure Broadcast Channels," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 370–381, Jan. 2018.

[103] S. S. Bidokhti, M. Wigger, A. Yener, and A. E. Gamal, "State-adaptive coded caching for symmetric broadcast channels," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Oct 2017, pp. 646–650.

[104] S. S. Bidokhti, M. Wigger, and R. Timo, “Noisy Broadcast Networks With Receiver Caching,” *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6996–7016, Nov. 2018.

[105] M. M. Amiri and D. D. Gündüz, “On the Capacity Region of a Cache-Aided Gaussian Broadcast Channel with Multi-Layer Messages,” in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 1909–1913.

[106] F. Xu, M. Tao, and K. Liu, “Fundamental Tradeoff Between Storage and Latency in Cache-Aided Wireless Interference Networks,” *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[107] J. S. P. Roig, S. A. Motahari, F. Tosato, and D. Gündüz, “Fundamental limits of latency in a cache-aided 4x4 interference channel,” in *2017 IEEE Information Theory Workshop (ITW)*, Nov 2017, pp. 16–20.

[108] J. S. P. Roig, D. Gündüz, and F. Tosato, “Interference networks with caches at both ends,” in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[109] J. S. P. Roig, F. Tosato, and D. Gündüz, “Storage-Latency Trade-Off in Cache-Aided Fog Radio Access Networks,” in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[110] J. Zhang, F. Engelmann, and P. Elia, “Coded caching for reducing CSIT-feedback in wireless communications,” *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1099–1105, Sept 2015.

[111] Y. Cao, M. Tao, F. Xu, and K. Liu, “Fundamental Storage-Latency Tradeoff in Cache-Aided MIMO Interference Networks,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.

[112] K. Ngo, S. Yang, and M. Kobayashi, “Scalable Content Delivery With Coded Caching in Multi-Antenna Fading Channels,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 548–562, Jan 2018.

[113] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, “Physical-Layer Schemes for Wireless Coded Caching,” *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[114] Y. Cao and M. Tao, “Treating Content Delivery in Multi-Antenna Coded Caching as General Message Sets Transmission: A DoF Region Perspective,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3129–3141, June 2019.

[115] A. Sengupta, R. Tandon, and O. Simeone, “Fog-Aided Wireless Networks for Content Delivery: Fundamental Latency Tradeoffs,” *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[116] J. Kakar, S. Gherekhloo, and A. Sezgin, "Fundamental Limits on Delivery Time in Cloud- and Cache-Aided Heterogeneous Networks," *arXiv:1706.07627*, 2017.

[117] J. Zhang and O. Simeone, "Cloud-Edge Non-Orthogonal Transmission for Fog Networks with Delayed CSI at the Cloud," in *2018 IEEE Information Theory Workshop (ITW)*, Nov 2018, pp. 1–5.

[118] ——, "Fundamental Limits of Cloud and Cache-Aided Interference Management with Multi-Antenna Edge Nodes," *IEEE Transactions on Information Theory*, pp. 1–1, 2019.

[119] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, "Decentralized coded caching in wireless networks: Trade-off between storage and latency," in *2016 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 2443–2447.

[120] F. Xu and M. Tao, "Fundamental Limits of Decentralized Caching in Fog-RANs with Wireless Fronthaul," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 1430–1434.

[121] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog Radio Access Networks," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2029–2033.

[122] M. Tao, D. Gündüz, F. Xu, and J. S. P. Roig, "Content caching and delivery in wireless radio access networks," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 4724–4749, July 2019.

[123] J. Zhang and P. Elia, "Feedback-aided coded caching for the MISO BC with small caches," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[124] J. Chen, S. Yang, A. Özgür, and A. Goldsmith, "Achieving Full DoF in Heterogeneous Parallel Broadcast Channels With Outdated CSIT," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4154–4171, Jul. 2016.

[125] H. Joudeh and B. Clerckx, "On the Separability of Parallel MISO Broadcast Channels Under Partial CSIT: A Degrees of Freedom Region Perspective," *arXiv preprint arXiv:1905.01283*, 2019.

[126] D. N. C. Tse and S. V. Hanly, "Multiaccess fading channels. I. Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2815, Nov 1998.

[127] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channels. II. Delay-limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2831, Nov 1998.

[128] H. Joudeh and B. Clerckx, "DoF Region of the MISO BC with Partial CSIT: Proof by Inductive Fourier-Motzkin Elimination," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2019, pp. 1–5.

[129] A. Sengupta, R. Tandon, and O. Simeone, “Cloud and cache-aided wireless networks: Fundamental latency trade-offs,” *arXiv:1605.01690*, 2016.

[130] A. S. Avestimehr, S. N. Diggavi, C. Tian, and D. N. C. Tse, “An Approximation Approach to Network Information Theory,” *Found. Trends Commun. Inf. Theory*, vol. 12, no. 1-2, pp. 1–183, 2015.

[131] B. Yuan and S. A. Jafar, “Elevated multiplexing and signal space partitioning in the 2 User MIMO IC with partial CSIT,” in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–6.

[132] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Cache-Aided Interference Management in Wireless Cellular Networks,” *IEEE Transactions on Communications*, pp. 1–1, 2019.

[133] R. Tandon, S. A. Jafar, S. Shamai (Shitz), and H. V. Poor, “On the Synergistic Benefits of Alternating CSIT for the MISO Broadcast Channel,” *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4106–4128, July 2013.

[134] S. Lashgari, R. Tandon, and S. Avestimehr, “MISO Broadcast Channel With Hybrid CSIT: Beyond Two Users,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7056–7077, Dec 2016.

[135] M. Razaviyayn, G. Lyubeznik, and Z. Luo, “On the Degrees of Freedom Achievable Through Interference Alignment in a MIMO Interference Channel,” *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 812–821, Feb. 2012.

[136] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Cache-aided interference management in wireless cellular networks,” in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–7.

[137] Kai Wan, D. Tuninetti, and P. Piantanida, “On the optimality of uncoded cache placement,” in *2016 IEEE Information Theory Workshop (ITW)*, Sep. 2016, pp. 161–165.

[138] I. Pinelis, “An inequality involving a sum of power terms,” MathOverflow, uRL:https://mathoverflow.net/q/297696 (version: 2018-04-12). [Online]. Available: https://mathoverflow.net/q/297696