Generalized Degrees of Freedom of the Symmetric Cache-Aided MISO Broadcast Channel With Partial CSIT

Enrico Piovano, Hamdi Joudeh and Bruno Clerckx

Abstract—We consider the cache-aided MISO broadcast channel (BC) in which a multi-antenna transmitter serves K single-antenna receivers, each equipped with a cache memory. The transmitter has access to partial knowledge of the channel state information. For a symmetric setting, in terms of channel strength levels, partial channel knowledge levels and cache sizes, we characterize the generalized degrees of freedom (GDoF) 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 GDoF outer bound for a parallel MISO BC under channel uncertainty are used for the converse. We further show that the characterized order-optimal GDoF is also attained in a decentralized setting, where no coordination is required for content placement in the caches.

Index Terms—Channel state information at the transmitter, coded caching, cache-aided interference management.

I. INTRODUCTION

Traffic over wireless networks is predominantly becoming content-oriented, a transformation mainly driven by the advent of multimedia applications, especially video-on-demand services [2]. For this type of traffic, there is often a large content library out of which users request specific files. The content library is typically generated well before transmission, creating the opportunity to pre-store (i.e. cache) parts of the content at different nodes across the network during off-peak times, when the network resources are under utilized. This cached information is then used during peak times, when users are actively requesting content and competing for wireless spectrum, to reduce the transmission load over the network [3]. Therefore, such cache-aided networks often operate in two phases: a *placement phase* which takes place during off-peak times, and a *delivery phase* which takes place during peak times [4].

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 communica-

tion links. This was taken up by Maddah-Ali and Niesen in [4], where caching was investigated in the context of a broadcast network in which one transmitter (server) communicates with multiple users, equipped with cache memories, over a shared noiseless link. In addition to the obvious local caching gains, Maddah-Ali and Niesen revealed a (hidden) *global caching gain* which scales with the aggregate size of cache memories distributed across the network, 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.

Global caching gains were initially demonstrated assuming a centralized setting, were centrally coordinated placement takes place [4]. While the placement phase takes place during off-peak hours before user demands are known to the transmitter, it was still assumed that it was carried out in a centrally coordinated manner in which the number and identity of active users during the delivery phase are known beforehand. This is often difficult to satisfy in practical networks, particularly in wireless settings where users enjoy a high degree of mobility. This called for developing a decentralized version of coded-caching, where placement is randomized and hence independent of the identity and number of active users during the delivery phase [5]. Surprisingly, it was shown in [5] that decentralization comes at a low price, achieving an order-optimal performance comparable to the centralized scheme.

The coded-caching framework above has been further extended in many directions. Such developments were recently surveyed in [6], in which challenges and open problems are also discussed. One of the main open problems identified in [6] is the capacity characterization of cache-aided wireless networks.

A. Cache-Aided Wireless Networks

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 Degrees of Freedom (DoF) metric and the Generalized Degrees of Freedom (GDoF) metric. 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 [7]–[18].

Amongst the main insights derived from the above studies 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 [7]–[10]. 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 [8], [19] and/or where channel state information at the transmitter (CSIT) is imperfect [15]–[18].

B. The Cache-Aided MISO Broadcast Channel

In this paper, we focus on the cache-aided multiple-input-single-output broadcast channel (MISO BC), in which a K-antenna transmitter serves K single-antenna users, where each user is equipped with a cache memory. Note that the K 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 [15]–[17].

Studying the classical MISO BC (with no caches) reveals that spatial multiplexing gains (i.e. DoF) of this channel suffer losses under imperfect CSIT. For example, the extreme case of finite precision CSIT causes a total collapse of the DoF to 1, where all (DoF) benefits of multiple transmitting antennas are lost [20]. The availability of partial instantaneous CSIT can help salvage some of the lost gains, achieving DoF between 1 and K depending on the CSIT quality. The complementary role of coded-caching in such scenarios was first observed in [15]. 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 [15] that coded-caching can offset the loss due to partial CSIT, up to a certain CSIT quality given the cache size.

The DoF metric, however, can be very pessimistic, as best exemplified by the DoF collapse in [20]. 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: 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 [21]–[23]. The cache-aided MISO BC was studied under the GDoF framework in [17],

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 [24]. In particular, [24] 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 [15] and then further investigated in [18], [19], [24], plays a central role in achieving and interpreting the order-optimal GDoF of the cache-aided MISO BC under partial CSIT as we show through our results. Next, we highlight the main contributions of this paper.

C. Main Contributions and Organization

We consider a 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 [21]–[23]. 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 [23]. For this setting, the main contributions are twofold, as stated below:

1) We characterize the optimal GDoF up to a constant multiplicative factor, which is independent of all system parameters. This order-optimal GDoF characterization is derived while allowing central coordination during the placement phase of the achievability scheme.
2) We show that the order-optimal 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* architecture [9]. 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 [4], [5]. 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, commonly employed for the classical MISO BC with partial CSIT [23], [25], [26], and essentially operates the physical channel at some point of its multiple multicast GDoF region.

One of the technical challenges in characterizing the optimal GDoF for the above setting is the converse, i.e. deriving an outer bound which is within a constant multiplicative factor

---

[1]The same can be said about [15], where the DoF under partial CSIT can be equivalently interpreted as the GDoF under no CSIT (see Section III-A2). No converse is given in [15], except for the trivial case where perfect CSIT is available.

from the achievable GDoF. Under partial CSIT, the conventional cut-set-based argument in [4] fails when employed on its own (see also [9]–[11] for variants of such argument). Alternatively, we derive an outer bound by marrying the approach in [4] with a robust GDoF outer 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 [20], [22], [23]. Specifically, in this novel adaptation of the approach in [20], [22], [23] to cache-aided networks, caches at receivers are replaced with equivalent parallel side links, and then an upper bound on the 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 GDoF achieved under decentralized placement. This intractability is circumvented by observing that the decentralized achievable GDoF is bounded below by a centralized-like achievable 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.

In addition to the contributions highlighted above, we derive several insights from the optimal GDoF characterization, which generalize former observations obtained in special cases of the considered setting [15], [17], [20], [23]. Such insights, and how they relate to previous observations, can be found in Section III-A. As for the remainder of the paper, the organization is as follows. Section II introduces the considered setting and problem. Section III presents the two main results and related insights. In Section IV, we derive an outer bound which is employed in the following two sections to show order optimality. In Section V and Section VI, we prove the two main results, the centralized setting result and the decentralized setting result respectively. Section VII concludes the paper.

II. PROBLEM SETTING

Consider a 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 U_i of size MF bits, where $M \in [0, N]$. We define the *normalized cache size* as

$$\mu \triangleq \frac{M}{N} \qquad (1)$$

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. 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
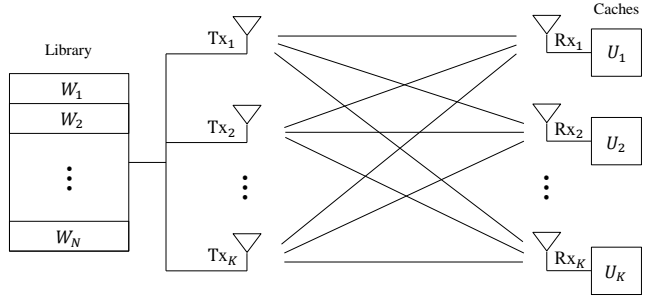

Fig. 1. A wireless network in which a transmitter of K antennas, $\text{Tx}_1, \ldots, \text{Tx}_K$, serves K single-antenna receivers, $\text{Rx}_1, \ldots, \text{Rx}_K$. The transmitter has access to a library of N files, while each receiver Rx_i is equipped with a cache memory U_i.

antenna (or radio head) as the *j-th transmitter* henceforth, while *transmitters* refers to the K transmit antennas jointly.

The network operates in two phases, *a placement phase* and a *delivery phase* [4]. 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 $U_i = \phi_i(\mathcal{W})$. 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 U_i. This is described in more detail below.

A. Physical Channel

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

$$Y_i(t) = \sum_{j=1}^{K} \sqrt{a_{ij}} G_{ij}(t) X_j(t) + Z_i(t) \qquad (2)$$

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 $\mathbb{E}(|X_j(t)|^2) \leq 1$ 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. 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 $t \in [T]$.

1) 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 [21], [22], channel gains are expressed in terms of the SNR as

$$a_{ii} = P \ \text{ and } \ a_{ij} = P^\alpha, \ \forall i, j \in [K], \ i \neq j \qquad (3)$$

where the parameter $\alpha \geq 0$ quantifies the strength of cross-links. The exponents of P in (3), i.e. 1 and α, are known as the channel strength parameters or levels. The channel model in (2) is rewritten as

$$Y_i(t) = \sqrt{P}G_{ii}(t)X_i(t) + \sum_{j=1, j \neq i}^{K} \sqrt{P^\alpha}G_{ij}(t)X_j(t) + Z_i(t) \qquad (4)$$

which is the model used throughout the paper. The results in this paper 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 [23], the regime $\alpha > 1$ poses new challenges and remains an open problem even for the classical MISO BC (with no caches) under partial CSIT.

Remark 1. *As pointed out in [22], 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.*

2) Partial CSIT: 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

$$G_{ij}(t) = \hat{G}_{ij}(t) + \sqrt{P^{-\beta}}\tilde{G}_{ij}(t) \qquad (5)$$

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 [22], [23]. 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 $\sim \sqrt{P^\beta}$. Moreover, we assume throughout this work 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 [23].

Before we proceed, it is worth highlighting that channel state information at the receivers (CSIR) is assumed to be perfect. Moreover, in a slight abuse of notation, we henceforth use $\hat{\mathcal{G}}$ to denote the entire channel knowledge available to the transmitters.

B. 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 $X_j^T = \psi_j^{(T)}(\mathcal{W}, \mathbf{d}, U_1, \ldots, U_K, \hat{\mathcal{G}})$, 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 $\hat{W}_i = \eta_i^{(T)}(Y_i^T, U_i, \mathbf{d}, \mathcal{G})$, 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 $\mathcal{C}^{(T)} \triangleq \{\phi_i, \psi_i^{(T)}, \eta_i^{(T)} : i \in [K]\}$. 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

$$P_e^{(T)} \triangleq \max_{\mathcal{G} | \hat{\mathcal{G}}} \max_{\mathbf{d} \in [N]^K} \max_{i \in [K]} \mathbb{P}\left(\hat{W}_i \neq W_{d_i}\right) \qquad (6)$$

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

$$R \triangleq \frac{KF}{T}. \qquad (7)$$

For given system parameters, we say that the rate R is achievable if there exists a coding scheme, consisting of a sequence of codes $\{\mathcal{C}^{(T)} : T \in \mathbb{N}\}$ 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.

1) 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

$$\text{GDoF} \triangleq \lim_{P \to \infty} \frac{C(P)}{\log(P)}. \qquad (8)$$

Being an asymptotic (high-SNR) measure, it is well understood that the GDoF does not depend on P. On the other hand, while fixing the number of users K, we often write $\text{GDoF}(\mu, \alpha, \beta)$ 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 [8], [9], [16], and GDoF studies in classical networks under finite precision and partial CSIT on the other hand [22], [23].

2) Generalized Normalized Delivery Time: Instead of working directly with the GDoF, it is easier to derive the results in terms of a function of the reciprocal[2] $\frac{1}{\mathsf{GDoF}}$. Hence, we introduce the *generalized normalized delivery time* (GNDT), where the optimal GNDT is defined as

$$\mathsf{GNDT}(\mu, \alpha, \beta) \triangleq \frac{K}{\mathsf{GDoF}(\mu, \alpha, \beta)}. \tag{9}$$

The GNDT (or the *delivery time* as we refer to it throughout the paper) 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 $\log(P) + o(\log(P))$, 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, we say that the delivery time $\mathsf{GNDT}'(\mu, \alpha, \beta)$ is achievable if $\mathsf{GNDT}'(\mu, \alpha, \beta) \geq \mathsf{GNDT}(\mu, \alpha, \beta)$.

The GNDT generalizes the *normalized delivery time* (NDT) metric in [11] 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 (9) that the GDoF can be interpreted as the capacity in files per time-slot. Before we proceed, we remark that in this paper, as in [4], [5], [7]–[11], [16], [17], we 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.

C. 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. 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. This lack of coordination is known as decentralized placement [5]. 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.

III. MAIN RESULTS AND INSIGHTS

The main results of this paper are: 1) the GDoF characterization of the symmetric cache-aided MISO BC under partial

[2]This has been observed when dealing with the DoF in many works including [7], [9], [10], [16].

CSIT, described in Section II, to within a constant multiplicative gap, and 2) showing that such 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.

A. Centralized placement

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

$$\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta) \triangleq \frac{K(1 - \mu)}{K(1 - (\alpha - \beta)) + (1 + K\mu)(\alpha - \beta)} \tag{10}$$

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 1. *For the symmetric cache-aided MISO BC under partial CSIT described in Section II, under centralized placement we achieve the GDoF given by*

$$\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta) = \frac{K}{\mathsf{GNDT}_{\mathrm{C}}(\mu, \alpha, \beta)}. \tag{11}$$

Moreover, the achievable GDoF in (11) *satisfies*

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

The proof of Theorem 1 is presented in Section V. As in [7], [9], the somewhat loose multiplicative gap of 12 in Theorem 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 GDoF characterized in Theorem 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 GDoF in (11) is expressed as

$$\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta) = (1 - (\alpha - \beta))\frac{K}{1 - \mu} + (\alpha - \beta)\frac{1 + K\mu}{1 - \mu}. \tag{13}$$

It is easily seen that $\mathsf{GDoF}_{\mathrm{C}}(\mu, \alpha, \beta)$ in (13) reduces to its classical counterpart in [23] 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 GDoF in (13), for any μ (in the set above), is analogous to the form of the classical GDoF in [23]. This is explained in more details next, where we use the terminology of signal power levels measured in terms of the exponent of P [27]. We start by looking at specialized cases from which we build our way towards the general case.

1) DoF Under Partial CSIT: Recall that 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 (13), we obtain

$$\mathsf{DoF}_{\mathrm{C}}(\mu, \beta) = \beta\frac{K}{1 - \mu} + (1 - \beta)\frac{1 + K\mu}{1 - \mu}. \tag{14}$$

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 DoF collapses to the one obtained in the original setting with a shared link [4]. 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 DoF collapse in the classical MISO BC [20]. Moreover, it is worth noting that since the DoF of the cache-aided MISO BC is an upper bound for the DoF of cache-aided interference networks, this collapse under finite precision CSIT also holds for the networks in [8]–[10].

For partial CSIT ($0 < \beta < 1$), the DoF takes the form $\beta \mathsf{DoF}_C(\mu, 1) + (1 - \beta)\mathsf{DoF}_C(\mu, 0)$, laying on the line connecting the two extremes. In this case, partial CSIT of level β allows (power-controlled) zero-forcing transmission in the bottom β signal power levels without leaking any interference above the noise floor at undesired users. 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. To facilitate such partitioned transmission, messages (or files) are split into private and common parts delivered through the zero-forcing and shared-link layers, respectively.

The scheme described above expands upon, and inherits the main features of, the rate-splitting scheme[4] used for the classical MISO BC with partial CSIT (alongside other networks) [23], [25], [26], [28]–[30]. Hence, it is not surprising to see that the cache-aided DoF takes the same weighted-sum form of the classical DoF in [26], recovered from the above by setting $\mu = 0$.

2) GDoF Under Finite Precision CSIT: This is recovered from (13) by setting $\beta = 0$ and corresponds to the achievable GDoF in [17]. It is easily checked that the GDoF in this case takes the form of the DoF in (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 [28]. 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.

3) The General Case: For arbitrary levels of β and α, the insights derived in [23] 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 (13).

B. 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 $\mathsf{GNDT}_D(\mu, \alpha, \beta)$, where

$$\mathsf{GNDT}_D(\mu, \alpha, \beta) \triangleq K \sum_{m=0}^{K-1} \frac{\binom{K-1}{m}\mu^m (1-\mu)^{K-m}}{K(1 - (\alpha - \beta)) + (1 + m)(\alpha - \beta)} \tag{15}$$

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

Theorem 2. *For the symmetric cache-aided MISO BC under partial CSIT described in Section II, under decentralized placement we achieve the GDoF given by*

$$\mathsf{GDoF}_D(\mu, \alpha, \beta) = \frac{K}{\mathsf{GNDT}_D(\mu, \alpha, \beta)}. \tag{16}$$

Moreover, the achievable GDoF in (16) *satisfies*

$$\mathsf{GDoF}_D(\mu, \alpha, \beta) \leq \mathsf{GDoF}(\mu, \alpha, \beta) \leq 12 \cdot \mathsf{GDoF}_D(\mu, \alpha, \beta). \tag{17}$$

The proof of Theorem 2 is presented in Section VI. The most significant consequence of Theorem 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 $\mathsf{GDoF}_C(\mu, \alpha, \beta) \leq 12 \cdot \mathsf{GDoF}_D(\mu, \alpha, \beta)$, obtained from (12) and (17). In Section VI-C, this multiplicative gap between the centralized GDoF and decentralized GDoF is tightened to 1.5.

In Section VI-B, we show that an upper bound on $\mathsf{GNDT}_D(\mu, \alpha, \beta)$ takes the form of the centralized delivery time in (10), yet with a lower coded-multicasting gain. It follows that the insights that follow Theorem 1, derived in the light of the centralized achievable GDoF, extend to the decentralized setting.

IV. OUTER BOUND

In this section, we obtain an outer bound (upper bound) for the GDoF. Since it is more convenient to work with the GNDT in (9), the outer bound is derived in terms of a lower bound on $\mathsf{GNDT}(\mu, \alpha, \beta)$.

Theorem 3. *For the symmetric cache-aided MISO BC under partial CSIT described in Section II, a lower bound on the optimal GNDT is given by*

$$\mathsf{GNDT}(\mu, \alpha, \beta) \geq \max_{s \in \{1, 2, ..., K\}} \mathsf{GNDT}_s^{\text{lb}}(\mu, \alpha, \beta), \tag{18}$$

---

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

[4]Also known as signal space partitioning [28].

where $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)$ *is defined as*[5]

$$\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta) \triangleq \left(\frac{s}{1 + (s-1)(1 - (\alpha - \beta))} \left(1 - \frac{M}{\left\lfloor \frac{N}{s} \right\rfloor} \right) \right)^+ . \tag{19}$$

In the above, for any subset of $s \leq K$ users, the corresponding $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)$ in (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 3 and (9), it is easily seen that an upper bound for the GDoF is given by

$$\mathsf{GDoF}(\mu, \alpha, \beta) \leq \min_{s \in \{1, 2, \ldots, K\}} \frac{K}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \alpha, \beta)}. \tag{20}$$

The outer bound in Theorem 3 is employed to prove the converse parts of Theorem 1 and Theorem 2 in the following sections. In the remainder of this section, we present a proof for Theorem 3. The proof relies on two main ingredients summarized as follows.

(a) 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.

(b) 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 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.

A. 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 $\left\lfloor \frac{N}{s} \right\rfloor$ distinct files and no file is requested by two different users. We denote the $\left\lfloor \frac{N}{s} \right\rfloor$ 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 $\left\lfloor \frac{N}{s} \right\rfloor$ 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 $\left\lfloor \frac{N}{s} \right\rfloor$ files. It is readily seen that $\mathsf{GNDT}_{\mathrm{md}}$ satisfies

$$\mathsf{GNDT}_{\mathrm{md}} \leq \left\lfloor \frac{N}{s} \right\rfloor \mathsf{GNDT}(\mu, \alpha, \beta) \tag{21}$$

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

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.

B. 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

$$Y_i(t) = \sqrt{P} G_{ii}(t) X_i(t) + \sum_{j=1, j \neq i}^{K} \sqrt{P^\alpha} G_{ij}(t) X_j(t) + Z_i(t) \tag{22}$$

$$B_i(t) = \sqrt{P^\gamma} A_i(t) + C_i(t) \tag{23}$$

where (22) and (23) describe the first and second sub-channels, respectively. All physical properties of (4), described in Section II-A, are inherited by the first sub-channel in (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 $\gamma \log(P) + o\left(\log(P) \right)$ 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 (22) and (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 (22) and (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 1. *For the s-user MISO BC, consisting of two parallel sub-channels, described in* (22) *and* (23)*, the optimal (sum) GDoF is bounded above as*

$$\mathsf{GDoF}_{\mathrm{P}}(\alpha, \beta, \gamma) \leq (\alpha - \beta) + s\left(1 - (\alpha - \beta) \right) + s\gamma. \tag{24}$$

It is evident that the bound on $\mathsf{GDoF}_{\mathrm{P}}(\alpha, \beta, \gamma)$ in (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 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 1 is relegated to Appendix A. Next, we argue that by setting γ such that

$$\gamma \cdot \mathsf{GNDT}_{\mathrm{md}} = M \tag{25}$$

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.

$$\frac{s\left\lfloor\frac{N}{s}\right\rfloor}{\mathsf{GDoF_P}(\alpha,\beta,\gamma)} \leq \mathsf{GNDT_{md}}. \qquad (26)$$

This follows by observing that (25) guarantees that for each user i, the content of the cache U_i in the original channel can be delivered over the second sub-channel in (23) using at most $\mathsf{GNDT_{md}}$ time-slots. Since this does not influence the GDoF achieved over the first sub-channel in (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 $\mathsf{GNDT_{md}}$ time-slots. We proceed while assuming that (25) holds.

By combining (26) with Lemma 1 and (25), followed by invoking (21), we obtain

$$\left\lfloor\frac{N}{s}\right\rfloor s \leq \mathsf{GNDT_{md}}\big(1 + (s-1)(1-(\alpha-\beta)) + s\gamma\big) \quad (27)$$

$$= \mathsf{GNDT_{md}}\big(1 + (s-1)(1-(\alpha-\beta))\big) + sM \quad (28)$$

$$\leq \mathsf{GNDT}(\mu,\alpha,\beta)\left\lfloor\frac{N}{s}\right\rfloor\big(1 + (s-1)(1-(\alpha-\beta))\big) + sM. \qquad (29)$$

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

$$\mathsf{GNDT}(\mu,\alpha,\beta) \geq \left(\frac{s}{1+(s-1)(1-(\alpha-\beta))}\left(1-\frac{M}{\left\lfloor\frac{N}{s}\right\rfloor}\right)\right)^+. \qquad (30)$$

The lower bound in (30) is further tightened by maximizing over all possible sizes of user subsets, i.e. $s \in [K]$, from which the result in (18) directly follows.

C. Insights and Relation to Prior Works

The multi-demand communication to a subset of users corresponds to the cut-set-based bound in [4], while the cache replacement is inspired by [16]. However, it is worthwhile highlighting that bounding the DoF under partial current and perfect delayed CSIT and side links (after cache replacement) in [16] is very different from bounding the GDoF under only partial current CSIT and side links in Lemma 1. In particular, the DoF upper bound in [16] follows the footsteps of [31], 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 1 is hence based on the outer bounds in [20], [22], [23], which rely on the aligned image sets approach under channel uncertainty.

It is also worthwhile highlighting that the GDoF upper bound in Lemma 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 [16], 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 1 that the considered cache-aided MISO BC collapses to the shared-link setting in [4] when $(\alpha - \beta) = 1$. However, even when current CSIT is completely absent in [16], the synergy between caches and delayed CSIT leads to an improved performance compared to the shared-link setting.

V. CENTRALIZED PLACEMENT

In this section, we treat the centralized setting and prove Theorem 1. We start with the achievability and then we prove order-optimality using the outer bound in Theorem 3.

A. Achievability scheme

Here we present a centralized scheme which achieves the delivery time given by $\mathsf{GNDT_C}(\mu,\alpha,\beta)$ in (10), and hence the GDoF given by $\mathsf{GDoF_C}(\mu,\alpha,\beta)$ in Theorem 1. This scheme builds upon and generalizes the one proposed for the cache-aided MISO BC in [17]. The key difference is that the scheme in [17] 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 2. *Consider the K-user MISO BC with signal model given by* (4) *and properties described in Section II-A. Further assume that the transmitter has a common message $W^{(\mathrm{c})}$, intended to all user, and private messages $W_1^{(\mathrm{p})}, \ldots, W_K^{(\mathrm{p})}$, where $W_i^{(\mathrm{p})}$ is intended only to user i. We achieve the GDoF*

$$\mathsf{GDoF}^{(\mathrm{c})} = (\alpha - \beta) \qquad (31)$$

$$\mathsf{GDoF}_i^{(\mathrm{p})} = 1 - (\alpha - \beta), \ \forall i \in [K] \qquad (32)$$

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 (31) and (32) is achieved using signal space partitioning [23], [28]. Using the terminology of signal power levels to explain this 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. The full proof of Lemma 2 is relegated to Appendix B.

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 [23]. For the other extreme of $\mu = 1$, we have $\mathsf{GNDT_C}(1,\alpha,\beta) = 0$

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 $\mu \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\}$, as pointed out in [4].

1) Placement phase: The placement is analogous to [4] 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 $U_i = \{W_{l,\mathcal{T}} : \mathcal{T} \in \Omega, i \in \mathcal{T}, l \in [N]\}$. 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.

2) 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 $W_{d_i,\mathcal{T}}^{(\mathrm{c})}$ and a private mini-subfile $W_{d_i,\mathcal{T}}^{(\mathrm{p})}$ such that $W_{d_i,\mathcal{T}} = \left(W_{d_i,\mathcal{T}}^{(\mathrm{c})}, W_{d_i,\mathcal{T}}^{(\mathrm{p})}\right)$. The two mini-subfiles $W_{d_i,\mathcal{T}}^{(\mathrm{c})}$ and $W_{d_i,\mathcal{T}}^{(\mathrm{p})}$ 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

$$q = \frac{(1+m_\mathrm{C})(\alpha-\beta)}{K(1-(\alpha-\beta)) + (1+m_\mathrm{C})(\alpha-\beta)}. \tag{33}$$

All common mini-subfiles are coded using the techniques in the original coded-multicasting scheme in [4]. In particular, subsets of $1+m_\mathrm{C}$ common mini-subfiles $W_{d_i,\mathcal{T}}^{(\mathrm{c})}$ are combined together using a bitwise XOR operation to generate multicasting messages intended for subsets of $1+m_\mathrm{C}$ users as follows

$$W_{\mathcal{S}}^{(\mathrm{c})} = \oplus_{i \in \mathcal{S}} W_{d_i,\mathcal{S}\setminus\{i\}}^{(\mathrm{c})} \tag{34}$$

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

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 [4]. 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, 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.

3) 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 2 and $(1+m_\mathrm{C})$ is the gain due to coded-multicasting. Hence, the delivery time for the shared-link layer is

$$\frac{Kq(1-\mu)}{(\alpha-\beta)(1+m_\mathrm{C})} = \frac{K(1-\mu)}{K(1-(\alpha-\beta)) + (1+m_\mathrm{C})(\alpha-\beta)}. \tag{35}$$

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 2. Hence, the delivery time for this layer is

$$\frac{K(1-q)(1-\mu)}{K(1-(\alpha-\beta))} = \frac{K(1-\mu)}{K(1-(\alpha-\beta)) + (1+m_\mathrm{C})(\alpha-\beta)}. \tag{36}$$

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

$$\mathsf{GNDT}_\mathrm{C}(\mu, \alpha, \beta) = \frac{K(1-\mu)}{K(1-(\alpha-\beta)) + (1+m_\mathrm{C})(\alpha-\beta)}. \tag{37}$$

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

B. Converse

Here we prove the converse in (12), which is equivalent to showing order-optimality of $\mathsf{GNDT}_\mathrm{C}(\mu, \alpha, \beta)$, i.e.

$$\frac{\mathsf{GNDT}_\mathrm{C}(\mu, \alpha, \beta)}{\mathsf{GNDT}(\mu, \alpha, \beta)} \leq 12.$$

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

$$\mathsf{GNDT}_\mathrm{C}(\mu, \delta) \triangleq \frac{K(1-\mu)}{K(1-\delta) + (1+K\mu)\delta} \tag{38}$$

and

$$\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta) \triangleq \left(\frac{s}{1+(s-1)(1-\delta)}\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)\right)^+. \tag{39}$$

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 (38), where the lower convex envelope is taken for the remaining points in $\mu \in [0,1]$. From the above, the lower bound in (18) is rewritten as

$$\mathsf{GNDT}(\mu, \alpha, \beta) \geq \max_{s \in \{1,2,\ldots,K\}} \mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta = \alpha - \beta) \tag{40}$$

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 C-A 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 (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.

VI. DECENTRALIZED PLACEMENT

In this section, we prove Theorem 2 which considers the decentralized setting. As in Section V, we start with the achievability and then proceed to prove order-optimality.

A. Achievability Scheme

Here we propose a decentralized scheme which achieves the delivery time given by $\mathsf{GNDT}_{\mathrm{D}}(\mu, \alpha, \beta)$ in (15), and hence the GDoF given by $\mathsf{GDoF}_{\mathrm{D}}(\mu, \alpha, \beta)$ in Theorem 2. We start with the placement phase.

1) Placement Phase: This is similar to the procedure in the original decentralized coded-caching paper [5], 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[6] $\tilde{\mathcal{T}} \in 2^{[K]}$, 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 $\{W_{l, \tilde{\mathcal{T}}} : \tilde{\mathcal{T}} \in 2^{[K]}\}$.

2) 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 $|\tilde{\mathcal{T}}| = m$. 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 V-A2 for the centralized setting, while replacing m_{C} in Section V-A2 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 (34) for the decentralized setting, all subfiles are assumed to be zero-padded to the length of the longest subfile [5]. 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 [23].

[6]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.

3) 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

$$\mu^m(1 - \mu)^{K-m}F + o(F) \tag{41}$$

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

$$K\binom{K-1}{m}\mu^m(1 - \mu)^{K-m}. \tag{42}$$

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

$$\frac{K\binom{K-1}{m}\mu^m(1 - \mu)^{K-m}}{K(1 - (\alpha - \beta)) + (1 + m)(\alpha - \beta)}. \tag{43}$$

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

$$\mathsf{GNDT}_{\mathrm{D}}(\mu, \alpha, \beta) = K\sum_{m=0}^{K-1}\frac{\binom{K-1}{m}\mu^m(1 - \mu)^{K-m}}{K(1 - (\alpha - \beta)) + (1 + m)(\alpha - \beta)}. \tag{44}$$

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

B. Converse

In this part, we prove the converse in (17), which is achieved by showing order-optimality of $\mathsf{GNDT}_{\mathrm{D}}(\mu, \alpha, \beta)$, i.e.

$$\frac{\mathsf{GNDT}_{\mathrm{D}}(\mu, \alpha, \beta)}{\mathsf{GNDT}(\mu, \alpha, \beta)} \leq 12.$$

As in the centralized setting, $\mathsf{GNDT}_{\mathrm{D}}(\mu, \alpha, \beta)$ only depends on the difference $\delta = (\alpha - \beta)$. Therefore, we work with

$$\mathsf{GNDT}_{\mathrm{D}}(\mu, \delta) \triangleq K\sum_{m=0}^{K-1}\frac{\binom{K-1}{m}\mu^m(1 - \mu)^{K-m}}{K(1 - \delta) + (1 + m)\delta} \tag{45}$$

where $\mathsf{GNDT_D}(\mu, \delta = \alpha - \beta) = \mathsf{GNDT_D}(\mu, \alpha, \beta)$. Unlike $\mathsf{GNDT_C}(\mu, \delta)$ in (38), $\mathsf{GNDT_D}(\mu, \delta)$ does not have the desirable form which allows comparing it to the bound in (40) directly. Hence, the first (key) step of the converse is to derive an upper bound on $\mathsf{GNDT_D}(\mu, \delta)$, denoted by $\mathsf{GNDT_D^{ub}}(\mu, \delta)$, which takes the form of the centralized achievable delivery time in (38). This is given in the following result.

Lemma 3. *The decentralized delivery time* $\mathsf{GNDT_D}(\mu, \delta)$ *is bounded above as*

$$\mathsf{GNDT_D}(\mu, \delta) \leq \mathsf{GNDT_D^{ub}}(\mu, \delta) = \frac{K(1-\mu)}{K(1-\delta) + (1+u)\delta} \tag{46}$$

where u *is given by*

$$u = \frac{K(1-\mu)}{\mathsf{GNDT_D}(\mu, 1)} - 1. \tag{47}$$

The proof of Lemma 3 is given in Appendix D. One important consequence of Lemma 3 is that the expression in (46) allows us to show order-optimality of $\mathsf{GNDT_D^{ub}}(\mu, \delta)$ to within a factor of 12 using similar techniques to the ones used for the centralized setting. The details are relegated to Appendix C-B. The order-optimality of $\mathsf{GNDT_D}(\mu, \delta)$ to within a factor of 12 follows, which concludes the converse.

C. Gap Between Decentralized and Centralized Schemes

From a straightforward inspection of (38) and (46), it can be seen that for integer values of $K\mu$ (for which a close form of $\mathsf{GNDT_C}(\mu, \delta)$ is obtained), we have

$$\frac{\mathsf{GDoF_C}(\mu, \delta)}{\mathsf{GDoF_D}(\mu, \delta)} = \frac{\mathsf{GNDT_D}(\mu, \delta)}{\mathsf{GNDT_C}(\mu, \delta)} \leq \frac{K(1-\delta) + (K\mu+1)\delta}{K(1-\delta) + (u+1)\delta}. \tag{48}$$

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 [4], [5]. Hence, it follows from the observations in [5] (and then the proof in [32]) that for $\delta = 1$, there is a small price to pay due to decentralization, making the ratio in (48) small. By further examining the bound on the right-most side of (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 (48) is bounded above as follows.

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

$$\frac{\mathsf{GDoF_C}(\mu, \delta)}{\mathsf{GDoF_D}(\mu, \delta)} \leq 1.5. \tag{49}$$

The above corollary is obtained by employing the results in Theorem 1, Lemma 3 and [32]. The full proof is relegated to Appendix E.

VII. CONCLUSIONS

In this paper, we characterized the optimal GDoF of the symmetric cache-aided MISO BC under partial CSIT up to a constant multiplicative factor. Moreover, we showed that such GDoF characterization is robust to decentralization, i.e. we proposed a decentralized caching strategy which attains an order-optimal GDoF performance. In order to derive the 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 GDoF, and is generally easier to deal with when characterizing achievable and optimal performances.

At the heart of our converse proof is a GDoF outer 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 GDoF takes a familiar weighted-sum form, often observed in classical networks (with no caches) under partial CSIT. Achieving such GDoF relies on a key interplay between spatial multiplexing and coded-multicasting gains.

This work opens the door for a number of interesting extensions. An intriguing direction is to consider a setting in which each transmitter can only store part of the library, hence enabling only partial transmitter cooperation as opposed to the full cooperation assumed in this work (e.g. general cache-aided interference networks). This setting generalizes the works in [9], [10] to the GDoF framework under partial CSIT considered in this paper. Progress along these lines is reported in [17], while limiting to absent CSIT and considering only achievability. As observed in [9], [10], under such partial cooperating (through caching at the transmitters), the underlying physical channel is modeled by the X channel. Hence, it is worthwhile highlighting in this context that for the X channel, the GDoF under partial CSIT is still an open problem. Another interesting direction is relaxing the symmetry in the channel. However, one 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 [23], [30]. Last but no least, reducing the constant multiplicative factor of 12 is also of significant interest. For the original shared-link setting, recent efforts managed to reduce the constant multiplicative factor [33], [34]. Our observations through numerical simulations, which show that the gap is much smaller than 12, provide hope that such tightening may also be possible for the order-optimal characterizations presented in this paper.

APPENDIX A
PROOF OF LEMMA 1

The proof is based on the approach in [20], [22], [23], 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 [20], [22]. We consider $s = K$ users. For general $s \leq K$, the exact same steps follow while considering only the corresponding s rate bounds.

A. 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

$$\bar{Y}_i(t) = \lfloor G_{ii}(t)\bar{X}_i(t) \rfloor + \sum_{j=1, j \neq i}^{K} \lfloor \bar{P}^{\alpha-1} G_{ij}(t)\bar{X}_j(t) \rfloor \quad (50)$$

$$\bar{B}_i(t) = \bar{A}_i(t) \quad (51)$$

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 GDoF upper bound for the deterministic channel is also a GDoF upper bound for the original channel using the same steps in [20]. Therefore we focus on the deterministic channel henceforth.

B. 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

$$nR_k \leq I\left(M_k; \bar{Y}_k^n, \bar{B}_k^n \mid M_{[k+1:K]}, \mathcal{G} \right) + o(n) \quad (52)$$
$$\leq H\left(\bar{Y}_k^n, \bar{B}_k^n \mid M_{[k+1:K]}, \mathcal{G} \right) - H\left(\bar{Y}_k^n, \bar{B}_k^n \mid M_{[k:K]}, \mathcal{G} \right) + o(n). \quad (53)$$

After omitting $o(n)$ and $o\left(\log(P) \right)$ terms, we obtain

$$n \sum_{k=1}^{K} R_k \leq n(1 + \gamma) \log(\bar{P}) +$$
$$\sum_{k=2}^{K} \underbrace{H\left(\bar{Y}_{k-1}^n, \bar{B}_{k-1}^n \mid M_{[k:K]}, \mathcal{G} \right) - H\left(\bar{Y}_k^n, \bar{B}_k^n \mid M_{[k:K]}, \mathcal{G} \right)}_{H_k^\Delta}. \quad (54)$$

Hence, the focus becomes to bound the differences of entropies $H_2^\Delta, \ldots, H_K^\Delta$.

C. Bounding the Differences of Entropies

Focusing on the term H_k^Δ, for any $k \in [2 : K]$, we proceed as follows:

$$H_k^\Delta =$$
$$H\left(\bar{Y}_{k-1}^n, \bar{B}_{k-1}^n \mid M_{[k:K]}, \mathcal{G} \right) - H\left(\bar{Y}_k^n, \bar{B}_k^n \mid M_{[k:K]}, \mathcal{G} \right) \quad (55)$$
$$= H\left(\bar{Y}_{k-1}^n \mid M_{[k:K]}, \mathcal{G} \right) - H\left(\bar{Y}_k^n \mid M_{[k:K]}, \mathcal{G} \right)$$
$$+ H\left(\bar{B}_{k-1}^n \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_{k-1}^{[n]} \right) - H\left(\bar{B}_k^n \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_k^n \right) \quad (56)$$
$$\leq H\left(\bar{Y}_{k-1}^n \mid M_{[k:K]}, \mathcal{G} \right) - H\left(\bar{Y}_k^n \mid M_{[k:K]}, \mathcal{G} \right) + n\log\left(\bar{P}^{\gamma} + 1 \right). \quad (57)$$

In the above, (56) is obtained from the chain rule, while (57) follows from $H\left(\bar{B}_k^n \mid M_{[k:K]}, \mathcal{G}, \bar{Y}_k^n \right) \geq 0$ and $H\left(\bar{B}_{k-1}^n \mid \right.$

$M_{[k:K]}, \mathcal{G}, \bar{Y}_{k-1}^n \left. \right) \leq H\left(\bar{B}_{k-1}^n \right) \leq \sum_{t=1}^{n} H\left(\bar{B}_{k-1}(t) \right) \leq n \log \left(\bar{P}^{\gamma} + 1 \right)$. 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 II-A2. This difference is bounded above as

$$H\left(\bar{Y}_{k-1}^n \mid M_{[k:K]}, \mathcal{G} \right) - H\left(\bar{Y}_k^n \mid M_{[k:K]}, \mathcal{G} \right) \leq$$
$$n\left(1 - (\alpha - \beta) \right) \log(\bar{P}) + o\left(\log(\bar{P}) \right). \quad (58)$$

The inequality in (58) follows directly from [23] (see the proofs of [23, Th. 1] and [23, Th. 2]), and is obtained using the aligned image sets approach [20]. 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 (58) and (57), we obtain

$$H_k^\Delta \leq n\left(\gamma + 1 - (\alpha - \beta) \right) \log(\bar{P}) + o\left(\log(\bar{P}) \right). \quad (59)$$

The bound in (59) holds for all $k \in [2 : K]$. By plugging (59) into (54), the result in (24) directly follows.

APPENDIX B
PROOF OF LEMMA 2

First, we rewrite the signal model in (4) in vector form as

$$Y_i = \sqrt{P}[\hat{G}_{i1} \cdots \hat{G}_{iK}]\mathbf{Q}_i\mathbf{X} + \sqrt{P^{1-\beta}}[\tilde{G}_{i1} \cdots \tilde{G}_{iK}]\mathbf{Q}_i\mathbf{X} + Z_i \quad (60)$$

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 $W^{(c)}$ and $W_1^{(p)}, \ldots, W_K^{(p)}$ are encoded into unit power independent Gaussian codewords $X^{(c)}$ and $X_1^{(p)}, \ldots, X_K^{(p)}$, respectively. The transmitted signal is then constructed as

$$\mathbf{X} = \mathbf{D}\left(\sqrt{1 - P^{\beta-\alpha}}\mathbf{V}^{(c)}X^{(c)} + \sqrt{P^{\beta-\alpha}} \sum_{k=1}^{K} \mathbf{V}_k^{(p)}X_k^{(p)} \right). \quad (61)$$

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. $\mathbf{V}^{(c)}$ is a generic (random) unit vector and $\mathbf{V}_k^{(p)} \triangleq \left[V_{k1}^{(p)} \cdots V_{kK}^{(p)} \right]^\mathsf{T}$ is a zero-forcing unit vector designed using the channel estimates such that

$$\sqrt{P^\alpha}\left(\hat{G}_{i1}V_{k1}^{(p)} + \cdots + \sqrt{P^{1-\alpha}}\hat{G}_{ii}V_{ki}^{(p)} + \cdots + \hat{G}_{iK}V_{kK}^{(p)} \right)$$
$$= 0, \ \forall i \neq k. \quad (62)$$

It is simple to verify from the zero-forcing condition that $V_{ki}^{(p)}$ cannot scale faster than $O(\sqrt{P^{\alpha-1}})$ for all $k \neq i$. Hence, the

received signal of user i is rewritten as

$$Y_i = \sqrt{P} a_i^{(c)} X^{(c)} + \sqrt{P^{1+\beta-\alpha}} a_{ii}^{(p)} X_i^{(p)} + \sum_{k=1, k\neq i}^{K} a_{ik}^{(p)} X_k^{(p)} + Z_i \tag{63}$$

where $a_i^{(c)}$ and $a_{ik}^{(p)}$, for all $i, k \in [K]$, are all $O(1)$.

Each user i decodes $X^{(c)}$ by treating interference as noise and recovers $W^{(c)}$. As $X^{(c)}$ is received with power $O(P)$, while interference plus noise has power $O(P^{1+\beta-\alpha})$, it follows that $X^{(c)}$ supports a rate of $(\alpha - \beta) \log(P) + o(\log(P))$. Then, each user i proceeds to remove the contribution of $X^{(c)}$ from the received signal and decodes its own $X_i^{(p)}$ while treating the remaining interference as noise, from which $W_i^{(p)}$ is recovered. As $X_i^{(p)}$ is received with power $O(P^{1+\beta-\alpha})$, while the remaining interference plus noise has power $O(1)$, it follows that $X_i^{(p)}$ supports a rate of $(1 + \beta - \alpha) \log(P) + o(\log(P))$.

Remark 2. *It is worthwhile highlighting that the achievable GDoF in Lemma 2 (shown in this appendix) can be inferred from [23]. One key difference, however, is that the MISO BC considered in [23] 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 [25].*

APPENDIX C
PROOFS OF ORDER OPTIMALITY

Here we provide proofs for the order-optimality parts of Theorem 1 and Theorem 2. We start with an instrumental lemma used throughout the proofs in the following subsections.

Lemma 4. *For parameters K, μ and s defined previously, if $\frac{K}{s(1+K\mu)} \geq 1$, then the function given by*

$$f(\delta; K, \mu, s) = \frac{1 + (s-1)(1-\delta)}{K(1-\delta) + (1+K\mu)\delta} \tag{64}$$

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 $K \geq s(1+K\mu)$. ∎

A. Order Optimality of $\mathsf{GNDT}_C(\mu, \delta)$

We show here that for any μ, there exists a particular $s \in [K]$ such that $\mathsf{GNDT}_C(\mu, \delta)/\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta) \leq 12$ 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 (19), we get that $\mathsf{GNDT}_1^{\mathrm{lb}}(\mu, \delta) = 1 - \mu$. On the other hand, $\mathsf{GNDT}_C(\mu, \delta) \leq \mathsf{GNDT}_C(\mu, 1) \leq K(1-\mu)$. Hence, $\mathsf{GNDT}_C(\mu, \delta)/\mathsf{GNDT}_1^{\mathrm{lb}}(\mu, \delta) \leq 12$.

Next, we consider $K \geq 13$. As in [4], 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 $\mathsf{GNDT}_C(\mu, 1) \leq$

$\mathsf{GNDT}_C(0, 1) = K$. By setting $s = \lfloor 0.275K \rfloor$, we know from [4] 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

$$\frac{\mathsf{GNDT}_C(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_C(0, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)}$$
$$= \underbrace{\frac{1 + (s-1)(1-\delta)}{K(1-\delta) + \delta}}_{f(\delta; K, 0, s)} \cdot \frac{K}{s\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)}. \tag{65}$$

Since $\frac{K}{s} \geq \frac{1}{0.275} > 1$, from Lemma 4 it follows that $f(\delta; K, 0, s)$ is non-decreasing in $\delta \in [0, 1]$. Hence,

$$\frac{\mathsf{GNDT}_C(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_C(0, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_C(0, 1)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, 1)} \leq 12. \tag{66}$$

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 [4] that $\mathsf{GNDT}_C(\mu, 1) \leq \mathsf{GNDT}_C(\tilde{\mu}, 1) \leq \frac{1}{\mu}$. By setting $s = \lfloor \frac{0.3}{\mu} \rfloor$, we also know from [4] that $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, 1) \geq \frac{1}{12\mu}$. Considering a generic $\delta \in [0, 1]$, we write

$$\frac{\mathsf{GNDT}_C(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_C(\tilde{\mu}, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)}$$
$$= \underbrace{\frac{1 + (s-1)(1-\delta)}{K(1-\delta) + (1+K\tilde{\mu})\delta}}_{f(\delta; K, \tilde{\mu}, s)} \cdot \frac{K(1-\tilde{\mu})}{s\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)}. \tag{67}$$

As $\frac{K}{s(1+K\tilde{\mu})} \geq \frac{1}{0.3} \frac{K\mu}{1+K\mu} > 1$, Lemma 4 implies that $f(\delta; K, \tilde{\mu}, s)$ is non-decreasing in $\delta \in [0, 1]$. Hence,

$$\frac{\mathsf{GNDT}_C(\mu, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_C(\tilde{\mu}, \delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta)} \leq \frac{\mathsf{GNDT}_C(\tilde{\mu}, 1)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, 1)} \leq 12. \tag{68}$$

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 [4], we know that $\mathsf{GNDT}_C(\mu, \delta) \leq \mathsf{GNDT}_C(\mu, 1) \leq 12(1 - \mu)$. Hence $\mathsf{GNDT}_C(\mu, \delta)/\mathsf{GNDT}_1^{\mathrm{lb}}(\mu, \delta) \leq 12$. This concludes the proof.

B. Order Optimality of $\mathsf{GNDT}_D(\mu, \delta)$

As for the centralized setting, we show that for any μ, there exists a particular $s \in [K]$ such that $\mathsf{GNDT}_D^{\mathrm{ub}}(\mu, \delta)/\mathsf{GNDT}_s^{\mathrm{lb}}(\mu, \delta) \leq 12$ for all $\delta \in [0, 1]$. We start with the following lemma.

Lemma 5. *The value u, defined in (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 (47), we have

$$\frac{K(1-\mu)}{K\mu}\left(1 - (1-\mu)^K\right) = \frac{K(1-\mu)}{1+u} \tag{69}$$

which follows from

$$\mathsf{GNDT}_D(\mu, 1) = \frac{K(1-\mu)}{K\mu}\left(1 - (1-\mu)^K\right)$$

as shown in [5]. Hence, showing that $u \leq K\mu$ it is equivalent to showing that

$$\frac{K(1-\mu)}{K\mu}\left(1-(1-\mu)^K\right) \geq \frac{K(1-\mu)}{1+K\mu} \quad (70)$$

$$\Rightarrow (K\mu+1)\left(1-(1-\mu)^K\right) \geq K\mu \quad (71)$$

$$\Rightarrow 1 \geq (K\mu+1)(1-\mu)^K. \quad (72)$$

The inequality in (72) 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 5, the remainder of the proof follows the same procedures in Appendix C-A. 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 (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 [5], 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 [5]. On the other hand, for a generic $\delta \in [0,1]$, we have

$$\frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta)} = \underbrace{\frac{1+(s-1)(1-\delta)}{K(1-\delta)+(1+u)\delta}}_{f(\delta;K,u/K,s)} \cdot \frac{K(1-\mu)}{s\left(1-\frac{M}{\lfloor\frac{N}{s}\rfloor}\right)}. \quad (73)$$

By applying Lemma 5 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 4, the function $f(\delta; K, u/K, s)$ is non-decreasing in $\delta \in [0,1]$. It follows that

$$\frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta)} \leq \frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,1)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,1)} \leq 12. \quad (74)$$

Next, we consider the sub-case $\frac{1}{K} < \mu \leq \frac{1}{12}$. From [5], we have $\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,1) \leq \frac{1}{\mu}$, and by setting $s = \lfloor\frac{1}{4\mu}\rfloor$, we have $\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,1) \geq \frac{1}{12\mu}$. For a generic $\delta \in [0,1]$, we have

$$\frac{\mathsf{GNDT}_{\mathrm{D}}^{\mathrm{ub}}(\mu,\delta)}{\mathsf{GNDT}_s^{\mathrm{lb}}(\mu,\delta)} = \underbrace{\frac{1+(s-1)(1-\delta)}{K(1-\delta)+(1+u)\delta}}_{f(\delta;K,u/K,s)} \cdot \frac{K(1-\mu)}{s\left(1-\frac{M}{\lfloor\frac{N}{s}\rfloor}\right)}. \quad (75)$$

By applying Lemma 5, it follows that $\frac{K}{s(1+u)} \geq 4 \cdot \frac{K\mu}{1+K\mu} > 1$. Hence, from Lemma 4, $f(\delta; K, u/K, s)$ is non-decreasing in $\delta \in [0,1]$. Therefore, the statement in (74) 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 [5], 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.

APPENDIX D
PROOF OF LEMMA 3

It readily seen from the definition of u in (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

$$b_m = \frac{K\binom{K-1}{m}\mu^m(1-\mu)^{K-m}}{K(1-\mu)}. \quad (76)$$

It can be shown that $\sum_{m=0}^{K-1} b_m = 1$ as follows

$$\sum_{m=0}^{K-1} b_m = \frac{1}{K(1-\mu)}\sum_{m=0}^{K-1} K\binom{K-1}{m}\mu^m(1-\mu)^{K-m} \quad (77)$$

$$= \sum_{m=0}^{K-1}\binom{K-1}{m}\mu^m(1-\mu)^{K-1-m} = 1 \quad (78)$$

where (78) follows from the binomial identity[7]. Hence, the inequality in (46) is equivalently written as

$$\sum_{m=0}^{K-1}\frac{b_m}{K(1-\delta)+(1+m)\delta} \leq \frac{1}{K(1-\delta)+(1+u)\delta} \quad (79)$$

$$\Rightarrow \sum_{m=0}^{K-1}\frac{b_m}{c_m+v} \leq \frac{1}{\tilde{c}+v}. \quad (80)$$

where $v \triangleq K(1-\delta)$, $c_m \triangleq (1+m)\delta$ and $\tilde{c} \triangleq (1+u)\delta$. By rearrangement of (80), we obtain

$$(\tilde{c}'+1)\sum_{m=0}^{K-1}\frac{b_m}{c_m'+1} \leq 1. \quad (81)$$

where $\tilde{c}' = \tilde{c}/v$ and $c_m' = c_m/v$ By the definition of u in (47), for any $\delta \in (0,1)$, we have

$$\sum_{m=0}^{K-1}\frac{b_m}{(1+m)\delta} = \frac{1}{(1+u)\delta} \quad (82)$$

$$\Rightarrow \sum_{m=0}^{K-1}\frac{b_m}{c_m} = \frac{1}{\tilde{c}} \quad (83)$$

$$\Rightarrow \sum_{m=0}^{K-1}\frac{b_m}{c_m'} = \frac{1}{\tilde{c}'}. \quad (84)$$

By plugging $\tilde{c}'$ from (84) into (81), we obtain

$$\left(\frac{1}{\sum_{m=0}^{K-1}\frac{b_m}{c_m'}}+1\right)\sum_{m=0}^{K-1}\frac{b_m}{c_m'+1} \leq 1. \quad (85)$$

Hence, showing that (85) holds implies that (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 (78), and by applying Jensen's inequality, we have

$$\sum_{m=0}^{K-1} b_m f\left(\frac{1}{c_m'}\right) \leq f\left(\sum_{m=0}^{K-1}\frac{b_m}{c_m'}\right) \quad (86)$$

---

[7]Note that the binomial identity is given by $(a+b)^n = \sum_{r=0}^n \binom{n}{r}a^r b^{n-r}$.

$$\Rightarrow \sum_{m=0}^{K-1} b_m \frac{1}{c'_m + 1} \leq \frac{\sum_{i=1}^{n} \frac{b_m}{c'_m}}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m} + 1} \qquad (87)$$

$$\Rightarrow \left(\frac{\sum_{m=0}^{K-1} \frac{b_m}{c'_m} + 1}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m}} \right) \left(\sum_{m=0}^{K-1} \frac{b_m}{c'_m + 1} \right) \leq 1 \qquad (88)$$

$$\Rightarrow \left(\frac{1}{\sum_{m=0}^{K-1} \frac{b_m}{c'_m}} + 1 \right) \sum_{m=0}^{K-1} \frac{b_m}{c'_m + 1} \leq 1 \qquad (89)$$

which is the inequality in (85). This concludes the proof.

APPENDIX E
PROOF OF COROLLARY 1

First, for $\mu = 0$ we have $\mathsf{GDoF_C}(0, \delta) = \mathsf{GDoF_D}(0, \delta) = K(1 - \delta) + \delta$, while for $\mu = 1$ we have $\mathsf{GNDT_C}(1, \delta) = \mathsf{GNDT_D}(1, \delta) = 0$. Therefore, we focus on $\mu \in (0, 1)$ in what follows. The multiplicative factor of 1.5 in (49) can be shown by considering the three following cases:

1) $K \geq 3$: From Theorem 1, it follows that $\mathsf{GDoF_C}(\mu, \delta)$ is bounded above by

$$\mathsf{GDoF_C}(\mu, \delta) \leq (1 - \delta) \frac{K}{1 - \mu} + \delta \frac{1 + K\mu}{1 - \mu} \qquad (90)$$

where (90) holds with equality for $\mu \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}\}$, as expressed in (13). For the remaining points in $\mu \in [0, 1]$, the achievable GDoF upper bound in (90) follows from

$$\mathsf{GNDT_C}(\mu, \delta) \geq \frac{K(1 - \mu)}{K(1 - \delta) + (1 + K\mu)\delta} \qquad (91)$$

which in turn holds as $\frac{K(1-\mu)}{K(1-\delta)+(1+K\mu)\delta}$ is convex in μ and $\mathsf{GNDT_C}(\mu, \delta)$ is the lower convex envelope (see (10)). From Lemma 3, a lower bound for $\mathsf{GDoF_D}(\mu, \delta)$ is given by

$$\mathsf{GDoF_D}(\mu, \delta) \geq (1 - \delta) \frac{K}{1 - \mu} + \delta \frac{1 + u}{1 - \mu} \qquad (92)$$

where $1 + u = \frac{K(1-\mu)}{\mathsf{GNDT_D}(\mu,1)}$ from (47). From [5], we know that $\mathsf{GNDT_D}(\mu, 1)$ can be written as

$$\mathsf{GNDT_D}(\mu, 1) = \frac{1 - \mu}{\mu} \left(1 - (1 - \mu)^K \right). \qquad (93)$$

It follows that $1 + u$ is given by

$$1 + u = \frac{K\mu}{1 - (1 - \mu)^K}. \qquad (94)$$

From (90) and (92), the ratio between $\mathsf{GDoF_C}(\mu, \delta)$ and $\mathsf{GDoF_D}(\mu, \delta)$ is bounded above as

$$\frac{\mathsf{GDoF_C}(\mu, \delta)}{\mathsf{GDoF_D}(\mu, \delta)} \leq \frac{(1-\delta)K + \delta(1+K\mu)}{(1-\delta)K + \delta(1+u)} \leq \frac{1 + K\mu}{1 + u}. \qquad (95)$$

where the rightmost inequality in (95) follows from $u \leq K\mu$, which in turn is obtained from Lemma 5 in Appendix C-B. By plugging (94) into (95), we obtain

$$\frac{1 + K\mu}{1 + u} = \frac{1 + K\mu}{K\mu} \left(1 - (1 - \mu)^K \right) \leq 1.5 \qquad (96)$$

where the bound by 1.5 follows from [32, 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 (95) and (96) we obtain

$$\frac{\mathsf{GDoF_C}(\mu, \delta)}{\mathsf{GDoF_D}(\mu, \delta)} \leq \frac{1 + 2\mu}{2\mu} \left(1 - (1 - \mu)^2 \right). \qquad (97)$$

It is readily seen that the right-hand-side of (97), 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 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 1 that $\mathsf{GDoF_C}(\mu, \delta) = \frac{2}{1-\mu}$. Combining this with the upper bound for $\mathsf{GDoF_D}(\mu, \delta)$ in (92), we obtain

$$\frac{\mathsf{GDoF_C}(\mu, \delta)}{\mathsf{GDoF_D}(\mu, \delta)} \leq \frac{2}{2(1 - \delta) + (1 + u)\delta} \leq \frac{2}{1 + u} \qquad (98)$$

where the rightmost inequality in (98) follows from the fact that $1 + u \leq 2$, which can be easily shown. By plugging (94) into (98), we obtain

$$\frac{2}{1 + u} = \frac{1}{\mu} \left(1 - (1 - \mu)^2 \right) = 2 - \mu. \qquad (99)$$

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 $\mathsf{GNDT_C}(\mu, \delta) = \mathsf{GNDT_D}(\mu, \delta) = 1 - \mu$, hence (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 $\mathsf{GDoF_C}(\mu, \delta)/\mathsf{GDoF_D}(\mu, \delta) = 1.5$. Therefore, 1.5 is in fact the tightest possible upper bound for $\mathsf{GDoF_C}(\mu, \delta)/\mathsf{GDoF_D}(\mu, \delta)$.

ACKNOWLEDGMENT

REFERENCES

[1] E. Piovano, H. Joudeh, and B. Clerckx, "Robust cache-aided interference management under full transmitter cooperation," in *Proc. IEEE ISIT*, Jun. 2018, pp. 1540–1544.

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

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

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

[5] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Networking*, vol. 23, no. 4, pp. 1029–1040, Aug 2015.

[6] M. A. Maddah-Ali and U. Niesen, "Coding for caching: fundamental limits and practical challenges," *IEEE Commun. Magazine*, vol. 54, no. 8, pp. 23–29, Aug. 2016.

[7] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE ISIT*, Jun. 2015, pp. 809–813.

[8] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[9] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5359–5380, Jul. 2018.

[10] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[11] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[12] J. Kakar, S. Gherekhloo, and A. Sezgin, "Fundamental limits on delivery time in cloud-and cache-aided heterogeneous networks," *arXiv:1706.07627*, 2017.

[13] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[14] X. Yi and G. Caire, "Topological coded caching," in *Proc. IEEE ISIT*, Jul. 2016, pp. 2039–2043.

[15] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *Proc. Allerton*, Sep. 2015, pp. 1099–1105.

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

[17] E. Lampiris, J. Zhang, and P. Elia, "Cache-aided cooperation with no CSIT," in *Proc. IEEE ISIT*, Jun. 2017, pp. 2960–2964.

[18] E. Piovano, H. Joudeh, and B. Clerckx, "On coded caching in the overloaded MISO broadcast channel," in *Proc. IEEE ISIT*, Jun. 2017, pp. 2795–2799.

[19] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *arXiv:1711.05969*, 2017.

[20] 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 Trans. Inf. Theory*, vol. 62, no. 10, pp. 5603–5618, Oct. 2016.

[21] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5534–5562, Dec. 2008.

[22] A. G. Davoodi and S. A. Jafar, "Transmitter cooperation under finite precision CSIT: A GDoF perspective," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 6020–6030, Sep. 2017.

[23] 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 Trans. Inf. Theory*, vol. 64, no. 11, pp. 7208–7217, Nov. 2018.

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

[25] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 315–328, Jan 2013.

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

[27] 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.

[28] B. Yuan and S. A. Jafar, "Elevated multiplexing and signal space partitioning in the 2 user MIMO IC with partial CSIT," in *Proc. IEEE SPAWC*, Jul. 2016.

[29] C. Hao and B. Clerckx, "MISO networks with imperfect CSIT: A topological rate-splitting approach," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2164–2179, May 2017.

[30] A. G. Davoodi and S. A. Jafar, "Generalized degrees of freedom of the symmetric K user interference channel under finite precision CSIT," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6561–6572, Oct. 2017.

[31] J. Chen, S. Yang, A. Özgür, and A. Goldsmith, "Achieving full DoF in heterogeneous parallel broadcast channels with outdated CSIT," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4154–4171, Jul. 2016.

[32] Q. Yan, X. Tang, and Q. Chen, "On the gap between decentralized and centralized coded caching schemes," *arXiv:1605.04626*, 2016.

[33] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4388–4413, Jul. 2017.

[34] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 647–663, Jan. 2019.

Enrico Piovano received the B.Sc. and M.Sc. degrees (summa cum laude) in Telecommunications Engineering from Politecnico di Torino, Italy, in 2012 and 2015, respectively, and the M.Sc. degree in Telecommunications Engineering from Télécom ParisTech, France, in 2015, as part of a double degree program. In 2014, he was an intern at Qualcomm Research in Nuremberg, Germany. He is currently a Ph.D. student in the Department of Electrical and Electronic Engineering at Imperial College London, U.K. His research interests include coded caching, wireless communications, information theory and machine learning.

Hamdi Joudeh received the M.Sc. degree in Communications and Signal Processing and the Ph.D. degree in Electrical Engineering from Imperial College London, U.K., in 2011 and 2016, respectively. He is currently a Research Associate at the Department of Electrical and Electronic Engineering, Imperial College London. His research interests include wireless communications and multiuser information theory. He is currently serving as an Associate Editor for the EURASIP Journal on Wireless Communications and Networking.

Bruno Clerckx (SM'17) received the M.S. and Ph.D. degrees in applied science from the Université Catholique de Louvain, Louvain-la-Neuve, Belgium, in 2000 and 2005, respectively. From 2006 to 2011, he was with Samsung Electronics, Suwon, South Korea, where he actively contributed to 3GPP LTE/LTE-A and IEEE 802.16m and acted as the Rapporteur for the 3GPP Coordinated Multi-Point Study Item. From 2014 to 2016, he was an Associate Professor with Korea University, Seoul, South Korea. He also held visiting research appointments at Stanford University, EURECOM, the National University of Singapore, and The University of Hong Kong. Since 2011, he has been with Imperial College London, first as a Lecturer from 2011 to 2015, and then as a Senior Lecturer from 2015 to 2017. He is currently a Reader (Associate Professor) with the Electrical and Electronic Engineering Department, Imperial College London, London, U.K.

Dr. Clerckx has authored two books, 150 peer-reviewed international research papers, and 150 standards contributions, and is the inventor of 75 issued or pending patents among which 15 have been adopted in the specifications of 4G (3GPP LTE/LTE-A and IEEE 802.16m) standards. His research area is communication theory and signal processing for wireless networks. He has been a TPC member, a symposium chair, or a TPC chair of many symposia on communication theory, signal processing for communication and wireless communication for several leading international IEEE conferences. He is an Elected Member of the IEEE Signal Processing Society SPCOM Technical Committee. He served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2011 to 2015 and is currently an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He has also been a (lead) guest editor for special issues of the EURASIP Journal on Wireless Communications and Networking, IEEE ACCESS and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He was an Editor for the 3GPP LTE-Advanced Standard Technical Report on CoMP.