Resolving the Feedback Bottleneck of Multi-Antenna Coded Caching

Eleftherios Lampiris, *Member, IEEE,* Antonio Bazco-Nogueras, *Member, IEEE,* Petros Elia, *Member, IEEE*

***Abstract*—Multi-antenna cache-aided wireless networks were thought to suffer from a severe feedback bottleneck, since achieving the maximal Degrees-of-Freedom (DoF) performance required feedback from all served users for the known transmission schemes. These feedback costs match the caching gains and thus scale with the number of users. In the context of the L-antenna Multiple-Input Single Output broadcast channel with K receivers, each having normalized cache size γ, we pair a fundamentally novel algorithm together with a new information-theoretic converse and identify the optimal tradeoff between feedback costs and DoF performance, by showing that having channel state information from only $C < L$ served users implies an optimal one-shot linear DoF of $C + K\gamma$. As a side consequence of this, we also now understand that the well known DoF performance $L + K\gamma$ is in fact exactly optimal. In practice, the above means that we are able to disentangle caching gains from feedback costs, thus achieving unbounded caching gains at the mere feedback cost of the multiplexing gain. This further solidifies the role of caching in boosting multi-antenna systems; caching now can provide unbounded DoF gains over multi-antenna downlink systems, at no additional feedback costs. The above results are extended to also include the corresponding multiple transmitter scenario with caches at both ends.**

***Index Terms*—Coded caching, multi-antenna transmission, Channel State Information, feedback cost, cache-aided MISO.**

I. Introduction

THE seminal work of Maddah-Ali and Niesen [1] revealed how caching modest amounts of content at the receivers has the potential to yield unprecedented reductions in the delivery delay of content-related traffic.

Specifically, the work in [1] considered a shared-link broadcast channel where a transmitter is tasked with serving content from a library of N files to K receiving users. Each user is endowed with a cache that can store a fraction $\gamma \in [0, 1]$ of the library, thus yielding a normalized cumulative cache size of $t \triangleq K\gamma$, which essentially means that each part of the library can appear t different times across the different caches. The approach of [1] was to design the cache placement algorithm in such a manner that desired content that resides in different caches could be combined together to form a single transmitted multicast signal that carries information for multiple users. In turn, these same users would then access their individual caches in order to remove all the unwanted interference from the multicast signal, and thus decode their desired message. In this shared-link (noiseless, wired) setting, with unitary link capacity, this strategy allows for a worst-case (normalized) delivery time of

$$\mathcal{T}_1(t) = \frac{K - t}{1 + t}, \tag{1}$$

which implies an ability to serve $1 + t$ users at a time. This performance is shown in [3] to be within a multiplicative gap of 2.01 of the optimal gain, while under the assumption of uncoded placement the above performance is exactly optimal [4], [5].

The direct extension of this result to the equivalent high Signal-to-Noise Ratio (SNR) single-antenna *wireless* Broadcast Channel (BC) — where the long-term capacity of each point-to-point link is similarly normalized to 1 file per unit of time — implies a Degrees-of-Freedom[1] (DoF) performance of

$$\mathcal{D}_1(t) \triangleq \frac{K - t}{\mathcal{T}_1(t)} = 1 + t, \tag{2}$$

which can be achieved without any Channel State Information at the Transmitter (CSIT).

This came in direct contrast with multi-antenna systems, which are known to also provide DoF gains but only with very high feedback costs that scale with these DoF gains. As it is known (cf. [6], [7]), such feedback costs are the reason for which most multi-antenna solutions fail to scale (cf. [8]–[21]). The huge impact of feedback on the network's performance has triggered a major research interest in understanding how imperfect, partial, or limited feedback can help improve system performance [22]. Among the vast literature that resulted from this interest, different works have focused, for example, on analyzing the impact of feedback in interference-limited multi-antenna cellular networks [23]–[25], the feasibility of Interference Alignment [26], the limited-feedback resource allocation in heterogeneous wireless networks [27], the capacity for Gaussian multiple access channels with feedback [28], and the effect of either rate-limited feedback in the interference channel [29] or SNR-dependent feedback in the Broadcast Channel [30]. Recently, the analysis of the significance of feedback has been extended also to secure communications [31], as well as to the capacity of burst noise-erasure channels [32].

A. Multi-antenna cache-aided channels

At the same time, there is substantial interest in combining the gains from caching with the traditional multiplexing gains of feedback-aided multi-antenna systems. Combining the two

This work was partially supported by the ANR project ECOLOGICAL-BITS-AND-FLOPS and by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929. (ERC project DUALITY) and by the European Research Council under the ERC grant agreement N. 789190 (project CARENET). Part of this work was presented in the 2018 IEEE International Symposium on Information Theory (ISIT) [2].

[1]We rigorously define the Degrees-of-Freedom in Section VI-A, Def. 4.

ingredients is only natural, given the promise of coded caching and the fact that multi-antenna technologies are currently the backbone of wireless systems. One can argue that coded caching stands a much better chance in becoming a pertinent ingredient of wireless systems if it properly accounts for the fact that the most powerful and omnipresent resource in current networks is the use of multi-antenna arrays.

This direction seeks to merge two seemingly opposing approaches, where traditional feedback-based multi-antenna systems work by creating parallel channels that separate users' signals, while coded caching fuses users' signals and counts on each user receiving maximum interference. In this context, the work in [33] analyzed the wired multi-server (L servers) setting, which can easily be seen to correspond to the high-SNR cache-aided Multiple-Input Single-Output (MISO) BC setting with L transmit antennas. An interesting outcome of this work is the revelation that multiplexing and caching gains can be combined additively, yielding an achievable DoF performance equal to

$$\mathcal{D}_L(t) = L + t. \tag{3}$$

In the same spirit, the work in [34] studied the K_T-transmitter, fully-connected network where the transmitters are equipped with caches that can each store a fraction $\gamma_T \in [1/K_T, 1]$ of the library, amounting to a normalized cumulative (transmitter-side) cache size of $t_T = K_T\gamma_T$. Under a normalized cumulative receiver-side cache size of t, the achievable DoF this time takes the form

$$\mathcal{D}_{t_T}(t) = t_T + t. \tag{4}$$

As shown in [34], under the assumption of uncoded placement, the performance in (3)-(4) is within a factor of at most 2 from the optimal one-shot linear DoF. Subsequently, many works such as [35]–[42] have developed different coded caching schemes for the multi-transmitter and the multi-antenna settings.

B. Scaling feedback costs in multi-antenna coded caching

While the single antenna case in [1] provides the near optimal caching gain t without requiring any CSIT, a significant feedback problem arises in the presence of multiple antennas. Specifically, all prior multi-antenna coded caching methods [33], [35], [36] that achieve the full DoF $L + t$ require each of the $L + t$ benefiting receivers to communicate feedback to the transmitter. To make matters worse, the problem extends to the dissemination of CSI at the receivers (CSIR), where now the transmitter is further forced to incorporate in this CSIR additional information on the CSIT-based precoders of all the $L + t$ benefiting users (*global CSIR*).

To demonstrate the structural origins of these CSI costs, we focus on a simple instance of the multi-server method in [33], which acts as a proxy to other methods with similar feedback requirements.

***Example* 1.** *Let us consider the $L = 2$-antenna MISO BC, with $K = 4$ receiving users and normalized cumulative cache size $t = 2$. In this setting, the algorithm of [33] can treat $L+t = 4$ users at a time. Assuming that users $1, 2, 3, 4$ request files A, B, C, D, respectively, each of the three transmissions of [33] takes the form*[2]

$$\begin{aligned}\mathbf{x} = {} & \mathbf{h}_4^{\perp}(A_{23} \oplus B_{13} \oplus C_{12}) + \mathbf{h}_3^{\perp}(A_{24} \oplus B_{14} \oplus D_{12}) \\ & + \mathbf{h}_2^{\perp}(A_{34} \oplus C_{14} \oplus D_{13}) + \mathbf{h}_1^{\perp}(B_{34} \oplus C_{24} \oplus D_{23}),\end{aligned} \tag{5}$$

where $\mathbf{h}_k^{\perp}$ denotes the precoder that is orthogonal to the channel of user k, and where W_{ij} denotes the part of file $W \in \{A, B, C, D\}$ that is cached at both users i and j. We can see that the transmitter must know all users' channel vectors, $\mathbf{h}_k$, $k \in \{1, 2, 3, 4\}$, in order to form the four precoders. In addition, each receiver must know the composite channel-precoder product for each precoder in order to be able to decode the desired subfile (e.g. receiver 1 must know $\mathbf{h}_1^{\dagger}\mathbf{h}_1^{\perp}$ as well as $\mathbf{h}_1^{\dagger}\mathbf{h}_2^{\perp}$, $\mathbf{h}_1^{\dagger}\mathbf{h}_3^{\perp}$ and $\mathbf{h}_1^{\dagger}\mathbf{h}_4^{\perp}$). This implies a feedback cost equal to $L + t = 4$ feedback-bearing users per transmission[3].

As we know (see for example [6], [43]), such scaling feedback costs[4] can consume a significant portion of the coherence time, thus resulting in diminishing DoF gains.

C. State of the art

Motivated by this feedback bottleneck, different works on multi-antenna (and multi-transmitter) coded caching have sought to reduce CSI costs. However, in all prior cases, any subsequent CSI reduction comes at the direct cost of substantially reduced DoF. For example, the works in [38], [44] consider reduced quality CSIT, but yield a maximum DoF that remains close to $1 + t$, while [45] considers that the transmitter has access to perfect CSI for only some fixed subset of the users, and shows that the optimal DoF is lower than $L + t$. Further, the works in [46]–[48] consider delayed or reduced quality CSIT at the expense though of lower DoF performance, while the work in [49] considers only statistical CSI, but again achieves significantly lower DoF. Moreover, the work in [50] uses ACK/NACK type CSIT to ameliorate the issue of unequal channel strengths (cf. [51], [52]), yet achieving no multiplexing gains. Similar results can be found in [37], [53]–[59], in more decentralized scenarios that involve multiple cache-aided transmitters.

As a conclusion, both for the cache-aided MISO BC [33] as well as for its multi-transmitter equivalent [34], the corresponding DoF $\mathcal{D}_L(t) = L + t$, has been known to require perfect feedback from all $L + t$ served users.

Remark 1. *Since the conference version of this work [2], there have been multiple algorithms for the setting here considered*

[2]For sake of readability, in the examples provided throughout the document, we will omit the commas between numbers belonging to a set, such that, for example, the part of file A stored at the users of set $\{3, 4\}$ will be denoted by A_{34}.

[3]In practical terms, this implies $L + t = 4$ uplink training slots for CSIT acquisition and $L + t = 4$ downlink training slots for global CSIR acquisition. We note that global CSIR acquisition can be performed by communicating each precoder to all users simultaneously, a process that is described in Appendix II-C.

[4]In general we note that, in the context of Frequency Division Duplexing (FDD), the previously mentioned feedback results in a CSIT cost of $L + t$ feedback vectors. On the other hand, in a Time Division Duplexing (TDD) environment, it leads to $L + t$ uplink training time slots for CSIT acquisition and an additional cost of $L + t$ downlink training time slots for global CSIR acquisition.

that exhibit low CSIT requirements. The interested reader is directed to [41], [42], [60]–[62].

D. Summary of contributions

The focus of this work is to establish and achieve the optimal relationship between feedback costs and DoF performance in multiple-antenna cache-aided settings.

As a consequence of our work, we now know that:

1) The optimal DoF of the cache-aided MISO BC, under the assumptions of uncoded placement and one-shot linear schemes, takes the form
$$\mathcal{D}_L(t) = L + t, \tag{6}$$
which tightens the previously known bound by a multiplicative factor of 2.
2) The optimal DoF (under the same assumptions) when feedback is limited to $C \geq 1$ participating users takes the form
$$\mathcal{D}_L(t, C) = \min(L, C) + t. \tag{7}$$
3) Similarly, in the multi-transmitter scenario, with transmitter-side normalized cumulative cache of size t_T and with each transmitter equipped with L_T antennas, the above optimal DoF performance takes the form
$$\mathcal{D}_{t_T \cdot L_T}(t, C) = \min(t_T \cdot L_T, C) + t. \tag{8}$$

The above are direct outcomes of a completely novel coded caching algorithm, which manages to achieve the optimal performance given any amount of available feedback. In particular, in the L-antenna MISO BC, and in the equivalent fully connected multi-antenna multi-transmitter setting with $t_T L_T = L$, we show that:

1) The algorithm manages to achieve the optimal DoF
$$\mathcal{D}_L(t) = L + t \tag{9}$$
and do so with a minimal feedback cost of
$$C = L, \tag{10}$$
which substantially diminishes the previously known cost of $L + t$.
2) The algorithm optimally degrades its DoF to
$$\mathcal{D}_L(t, C) = C + t \tag{11}$$
when feedback is reduced to $C \in \{2, ..., L-1\}$. This is an improvement over the state of the art, which, for the same DoF, would require a feedback cost of $C + t$.
The novelty of our scheme lies in the deviation from the traditional clique-based structure that most schemes are based on. Rather than requiring from each user to "cache-out" t subfiles in a XOR as is commonly done, we are able to design transmissions that can benefit from a two-pronged approach: some users cache-out $t + L - 1$ subfiles, and thus do not require the assistance of CSI-aided precoding, while others only cache-out $\frac{t}{L}$ subfiles but for that they rely on feedback. This allows our scheme to avoid the need to eventually "steer-away" subfiles from every active user, which had been the reason for the high feedback costs in all known prior designs.

Finally, an important contribution of this work can be found in the novel outer bound. This bound extends the effort in [34] in two crucial ways.

- A main contribution of the converse result is the incorporation of the limited feedback constraint. We integrate this new restriction by characterizing its impact on the number of users that we can serve simultaneously, which is obtained by exploiting the dimensionality of the linear system implicit in the multi-user transmission with constrained feedback.
- Furthermore, we are able to improve the converse result of [34] by leveraging on the following insights: First, we exploit the symmetry of the configuration, which allows us to express the objective function of the optimization problem only in terms of the number of transmitters and the number of receivers that are caching each packet — eliminating any dependence on the specific packet or node. Second, this symmetrization allows us to eventually produce an objective function that has monotonicity and convexity properties, and which in turn allows us to manipulate the solution to yield a tight bound. It might be worth noting that, as a result of this new approach, our converse also establishes exact optimality for a few subsequent works.

E. Notation

Symbols $\mathbb{N}, \mathbb{C}$ denote the sets of natural and complex numbers, respectively. For $n, k \in \mathbb{N},\ n \geq k$, we denote the binomial coefficient with $\binom{n}{k}$, while $[k]$ denotes set $\{1, 2, ..., k\}$. For the bitwise-XOR operator we use $\oplus$. Greek lowercase letters are mainly reserved for sets. We further assume that all sets are ordered, and we use $|\cdot|$ to denote the cardinality of a set. Bold lowercase letters are reserved for vectors, while for some vector $\mathbf{h}$, comprised of Q elements, we denote its elements as $\mathbf{h}(q)$, $q \in [Q]$, i.e., $[\mathbf{h}(1), \mathbf{h}(2), ..., \mathbf{h}(Q)] \triangleq \mathbf{h}^T$. Bold uppercase letters are used for matrices, while for some matrix $\mathbf{H}$ we denote its i-th row, j-th column element as $\mathbf{H}(i, j)$.

II. System Model

We consider the cache-aided MISO BC where an L-antenna transmitter serves K receiving, single-antenna, cache-aided users. The distributed version of this setting, with multiple cache-aided transmitters, is discussed in Section V.

In our setting, the transmitter has access to a library of $N \geq K$ files, $\mathcal{F} = \{W^{(n)}\}_{n=1}^{N}$, of equal size. Each user has a cache that can fit a fraction $\gamma \in [0, 1]$ of the library and thus the users can collectively store $t = K\gamma$ times the entire library. We assume that during the delivery phase the users request their desired file simultaneously, and that each requested file is different. The users' file demand vector is denoted as $\mathbf{d} = \{d_1, ..., d_K\}$, implying that each user k will request file $W^{(d_k)}$. The received signal at user $k \in [K]$ takes the form
$$y_k = \mathbf{h}_k^{\dagger}\mathbf{x} + w_k, \tag{12}$$

where $\mathbf{x} \in \mathbb{C}^{L\times 1}$ denotes the transmitted signal-vector from the L-antenna transmitter satisfying the power constraint $\mathbb{E}\left\{\|\mathbf{x}\|^2\right\} \leq P$. In the above, $\mathbf{h}_k \in \mathbb{C}^{L\times 1}$ denotes the random-fading channel vector of user k, which is assumed to be drawn from a continuous non-degenerate distribution. This fading process is assumed to be statistically symmetric across users. Finally, the additive noise $w_k \sim \mathbb{C}\mathcal{N}(0,1)$, experienced at user k, is assumed to be Gaussian. The work focuses on the DoF performance, and thus the SNR is considered to be large. We also assume that the quality of CSIT is perfect, and we define the feedback amount required *in each (packet) transmission* as follows.

Definition 1 (Feedback Cost)**.** *A communication is said to induce* feedback cost C *if C users need to communicate their CSI at the transmitter, and at the same time the transmitter needs to communicate information on C precoders to the users.*

Structure of the paper: In Section III we present the main results of this work and provide a preliminary example of the achievable scheme. Further, in Section IV we fully describe the achievable scheme for the single transmitter (L antennas) case, and elaborate on the example of Section III. In Section V we extend the scheme to the multi-transmitter case. In Section VI we describe the proof of the converse result, while in Section VII we provide general conclusions. The subsequent appendices include proofs, as well as a discussion on the CSIT and global CSIR feedback acquisition process that conveys the precoder information to the users.

III. Main Results

We proceed with our main results, first by considering the single transmitter case (with L transmit antennas), and later by extending the result to the general K_T-transmitter setting. We remind the reader that the scheme's optimality is under the assumptions of one-shot linear precoding with uncoded cache placement, while we note that we directly omit the trivial bound $\mathcal{D}_L(t,C) \leq K$, and that we also do not consider the case of $C = 0$ as this corresponds to the well known result in [1]. We additionally recall that the setting asks that each of the K receiving users is equipped with an identically-sized cache of normalized size γ, thus corresponding to a normalized cumulative receiver-side cache size of $t = K\gamma$. Further, we assume that communicating a single subfile requires multiple coherence times.

Theorem 1. *In the K-user cache-aided MISO BC with L transmit antennas, normalized cumulative cache size t, and feedback cost C, the optimal DoF is*

$$\mathcal{D}_L(t,C) = \min(L,C) + t. \tag{13}$$

Proof. The achievability part is constructive and is described in Section IV, while the converse is proved in Section VI. □

Let us consider now the more general setting where the L-antenna transmitter is substituted by K_T cache-enabled transmitters. Each transmitter is equipped with L_T transmit antennas, and is able to store a fraction $\gamma_T \in [1/K_T, 1]$ of the library, inducing a normalized cumulative cache size of $t_T \triangleq K_T\gamma_T$.

Theorem 2. *In the K_T-transmitter wireless network, where each transmitter is equipped with L_T transmit antennas, with transmitter-side normalized cumulative cache size t_T, receiver-side normalized cumulative cache size t, and feedback cost C, the optimal DoF is*

$$\mathcal{D}_{L_Tt_T}(t,C) = \min(L_Tt_T, C) + t. \tag{14}$$

Proof. The achievability part of the proof is described in Section V, while the converse is described in Section VI. □

Remark 2. *Comparing Theorem 1 with Theorem 2, we can see that the cache-aided MISO BC and its multi-transmitter equivalent (corresponding to $L_Tt_T = L$) are akin not only in terms of DoF performance, but also in terms of the CSIT required to achieve this performance. Their behavior is the same, irrespective of the amount $C \geq 1$ of available feedback.*

The following corollary establishes the *exact* optimal DoF performance of the considered multi-antenna settings.

Corollary 1. *The optimal DoF of the L-antenna MISO BC with K users and normalized cumulative cache size t takes the form*

$$\mathcal{D}_L(t) = L + t. \tag{15}$$

Remark 3. *The DoF performance $\mathcal{D}_L(t,C) = L + t$ can be achieved by knowing the CSIT of only $C = L$ users at each transmission.*

Remark 4. *We can see from Theorem 1 and Theorem 2 that, in order to achieve the maximum DoF performance $\mathcal{D}_L(t,C) = L + t$ of the multi-antenna setting, condition $C \geq L$ is both sufficient and necessary.*

Remark 5. *In several scenarios such as in [63], [64], the best known bounds — which are built on the converse proof of [34] — endure a multiplicative gap to the optimal performance. Our converse proof improves the converse in [34] by tightening the lower bound of the solution to the linear program proposed in [34]. Consequently, our converse also closes the multiplicative gap of such subsequent works. For example, it follows directly from the results derived here that the achievable DoF presented in [63] for a cache-aided interference network with heterogeneous parallel channels and centralized cache placement is in fact exactly optimal. Similarly, the achievable DoF in [64] for cache-aided cellular networks again turns out to be exactly optimal.*

Intuition on the scheme and an example

Revisiting the previous optimal multi-antenna coded caching algorithms (cf. [33]–[36]) — which, as we noted, require CSIT from all $L + t$ "active" users — we remark that the main premise of these designs is that each transmitted subfile can be cached-out by some t users (as in the algorithm of [1]), and at the same time it can be zero-forced at some other $L - 1$ users. This, in turn, allows each of the $L + t$ active users to receive its desired subfile free of interference. This design, while achieving the maximum DoF, incurs very high CSIT costs. Notably, these costs are associated with the need to eventually "steer-away" subfiles from every active user.

The idea that we follow is different. In order to reduce the amount of CSIT to only L feedback-aided users, while retaining the full DoF performance, it follows that the t users whose CSI is unknown (hereon referred to as set π) will need to cache-out a total of $t+L-1$ subfiles each because they cannot be assisted by precoding. On the other hand, the L users whose CSI is known (hereon referred to as set λ) will be assisted by precoding, and thus they can more easily receive their desired subfile. Hence, the main design challenge is to transmit together subfiles that can be decoded by each user of set π. We proceed with a preliminary description of the proposed algorithm.

Algorithm overview: We first note that the cache placement draws directly from [1], both in terms of file partition as well as in terms of storing of subfiles in the users' caches.

On the other hand, the XOR generation method will be fundamentally different. The first step is to construct XORs composed of $\frac{t}{L}+1$ subfiles and to compose each transmit-vector with L such XORs. This allows each transmission to communicate $L+t$ different subfiles aimed at serving, simultaneously, a set of $L+t$ users. As discussed above, each such set of $L+t$ active users is partitioned into two sets; the first set, λ, consists of the L users that are assisted by precoding. The second set, π, has t users who are not assisted by precoding and who must compensate with their caches. The vector of XORs will be multiplied by $\mathbf{H}_{\lambda}^{-1}$, which represents the normalized inverse of the channel matrix between the transmitter and the users in set λ.

We will see that the above design guarantees that, during the decoding process, each of the users in λ only receives one of the XORs (because the rest will be nulled-out by the precoder), while the remaining t users, i.e., those in π, receive a linear combination of all L XORs. Hence, this means that each user in λ needs to cache out $\frac{t}{L}$ subfiles in order to decode its desired subfile, while the users in π need to cache out $t+L-1$ subfiles, i.e., all but one subfiles.

Algorithm demonstration through an example: Next, we will demonstrate a single transmission of our algorithm by considering the setting of Example 1. The goal is to achieve the same performance as before (delivery to four users at a time) while using CSIT from only two users at a time. The example in its entirety can be found in Section IV-C, Example 4.

Example 2. *In the same MISO BC setting of Example 1 with $L=2$ transmit antennas, $K=4$ users, and normalized cumulative cache size $t=2$, one transmitted vector of the proposed algorithm takes the form*[5]

$$\mathbf{x}=\mathbf{h}_2^{\perp}(A_{34}\oplus C_{14})+\mathbf{h}_1^{\perp}(B_{34}\oplus D_{23}), \tag{16}$$

where $\mathbf{h}_k^{\perp}$, $k\in\{1,2\}$, denotes the precoder-vector designed to be orthogonal to the channel of user k, files A, B, C, and D are requested by users $1,2,3$, and 4, respectively, and where W_{ij}, $W\in\{A,B,C,D\}$, represents the subfile of file W that can be found in the caches of users i and j.

[5]The reader is warned that there is a small notational discrepancy between the subfile indices of this example and the formal notation. In this example we have kept the notation as simple as possible in order to more easily provide a basic intuition on the structure of the scheme.

Assuming that user k receives y_k, $k\in[4]$, the message at each user takes the form

$$\begin{bmatrix}y_1\\y_2\\y_3\\y_4\end{bmatrix}=\begin{bmatrix}\mathbf{h}_1^{\dagger}(\mathbf{h}_2^{\perp}A_{34}\oplus C_{14}+\mathbf{h}_1^{\perp}B_{34}\oplus D_{23})\\\mathbf{h}_2^{\dagger}(\mathbf{h}_2^{\perp}A_{34}\oplus C_{14}+\mathbf{h}_1^{\perp}B_{34}\oplus D_{23})\\\mathbf{h}_3^{\dagger}(\mathbf{h}_2^{\perp}A_{34}\oplus C_{14}+\mathbf{h}_1^{\perp}B_{34}\oplus D_{23})\\\mathbf{h}_4^{\dagger}(\mathbf{h}_2^{\perp}A_{34}\oplus C_{14}+\mathbf{h}_1^{\perp}B_{34}\oplus D_{23})\end{bmatrix}=\begin{bmatrix}A_{34}\oplus C_{14}\\B_{34}\oplus D_{23}\\\mathbf{h}_3^{\dagger}(\mathbf{h}_2^{\perp}A_{34}\oplus C_{14}+\mathbf{h}_1^{\perp}B_{34}\oplus D_{23})\\\mathbf{h}_4^{\dagger}(\mathbf{h}_2^{\perp}A_{34}\oplus C_{14}+\mathbf{h}_1^{\perp}B_{34}\oplus D_{23})\end{bmatrix} \tag{17}$$

where we have ignored noise for simplicity.

Hence, we see that users 1 and 2 only receive the first and second XOR, respectively, due to the precoder design. This means that each of these two users can decode its desired subfiles, A_{34} and B_{34}, respectively, by caching-out the unwanted subfiles C_{14} and D_{23}, respectively.

On the other hand, looking at the decoding process for users 3 and 4, we see that user 3 can cache-out subfiles A_{34}, B_{34}, and D_{23} in order to decode the desired C_{14}. Similarly, user 4 can cache-out subfiles A_{34}, B_{34}, and C_{14} to decode the desired subfile D_{23}. In order to achieve this, users 3 and 4 need to employ their cached content, but they also need some CSI knowledge: user 3 needs products $\mathbf{h}_3^{\dagger}\mathbf{h}_2^{\perp}$ and $\mathbf{h}_3^{\dagger}\mathbf{h}_1^{\perp}$, while user 4 needs $\mathbf{h}_4^{\dagger}\mathbf{h}_2^{\perp}$ and $\mathbf{h}_4^{\dagger}\mathbf{h}_1^{\perp}$. This can be handled with the broadcasting of information for only two precoders. The reader is referred to Appendix II-C for an exposition of how the feedback acquisition here requires only $L=2$ training slots, which is simply because information on a precoder can be broadcast in a single shot, irrespective of how many users it is broadcast to.

IV. Description of the Scheme

We proceed with the presentation of the scheme's cache-placement and content-delivery phases. We focus on the single transmitter MISO BC setting, while the multi-transmitter scenario is presented in Section V. Furthermore, we also assume in the following that $C=L$, noting that the extension of the scheme to the case $C<L$ is trivial and it can be achieved by simply "shutting down" $L-C$ antennas. The scheme is described for the case where $\frac{t}{L}\in\mathbb{N}$, while the remaining cases can be achieved using memory sharing and, as shown in [39], would incur a small DoF reduction[6].

Communication happens in two phases, namely the placement and the delivery phases. The placement phase is responsible for populating the caches of the users with content, while the delivery phase is responsible for communicating to the users their desired files. Further, we assume that each transmitted signal as described by our algorithm (cf. (25)) either fits inside a single coherence period or requires multiple coherence periods to be successfully communicated.

[6]This DoF reduction as a result of non-integer values of t/L has been calculated in [39], and it is upper bounded by a multiplicative factor of 2 when $L>t$ and by a multiplicative factor of 1.5 when $L<t$. Very recent efforts [41], [60], subsequent to our work, have addressed this memory sharing issue through new designs that are able to retain the same optimal DoF and desirable low feedback cost without being constrained on the choice of L.

Precoder design: For some user set $\lambda \subset [K]$, $|\lambda| = L$, we denote as $\mathbf{H}_\lambda^{-1}$ the normalized inverse of the $L \times L$ channel matrix $\mathbf{H}_\lambda$ corresponding to the channel between the transmitter and the L users of set λ. Further, the ℓ-th column of $\mathbf{H}_\lambda^{-1}$, $\ell \in [L]$, is denoted by $\mathbf{h}^{\perp}_{\lambda\setminus\lambda(\ell)}$ and describes a vector that is orthogonal to the channels of the users of set $\lambda\setminus\{\lambda(\ell)\}$. Hence, for an arbitrary user $k \in [K]$ it holds

$$\mathbf{h}_k^\dagger \cdot \mathbf{h}^{\perp}_{\lambda\setminus\lambda(\ell)} \begin{cases} = 0, & \text{if } k \in \lambda \setminus \lambda(\ell) \\ \neq 0, & \text{else.} \end{cases} \tag{18}$$

A. Placement phase

The placement phase is executed without knowledge of the number of transmit antennas, and without knowledge of CSI. The placement follows the original scheme in [1] where each file $W^{(n)}, n \in [N]$, is initially split into $\binom{K}{t}$ subfiles

$$W^{(n)} \to \left\{ W_\tau^{(n)}, \quad \tau \subset [K], \ |\tau| = t \right\}, \tag{19}$$

each indexed by a t-length set $\tau \subset [K]$, such that the cache of user $k \in [K]$ takes the form

$$\mathcal{Z}_k = \left\{ W_\tau^{(n)} : \forall \tau \ni k, |\tau| = t, \forall n \in [N] \right\}. \tag{20}$$

B. Delivery phase

This phase begins with the request from each user of a single file from the library. To satisfy these demands, the transmitter will sequentially serve each one of the possible combinations of $L + t$ users. The communication to a particular subset of $L + t$ users is denoted as *transmission slot*.

The transmitter selects a subset of $L + t$ users for each transmission slot. Specifically, these users are divided into set $\lambda \subset [K]$, $|\lambda| = L$, who provide CSI, and set $\pi \subset [K]\setminus\lambda$, $|\pi| = t$, who need not provide CSI.

Upon notification of the requests $\{W^{(d_k)}, \ k \in [K]\}$, and after the number of antennas is revealed to be L, each requested subfile $W_\tau^{(d_k)}$ is further split twice as follows:

$$W_\tau^{(d_k)} \to \{W_{\sigma,\tau}^{(d_k)}, \ \sigma \subseteq [K] \setminus (\tau \cup \{k\}), |\sigma| = L - 1\} \tag{21}$$

$$W_{\sigma,\tau}^{(d_k)} \to \{W_{\sigma,\tau}^{(r,d_k)}, \ r \in [L + t]\}. \tag{22}$$

Each subfile $W_{\sigma,\tau}^{r,(d_k)}$ is uniquely characterized by 4 indices. Index τ indicates the t users who have cached the subfile. Index σ indicates a set of $L - 1$ users from which this subfile will be steered-away via precoding. Superscript r is used for symmetrization, as will become evident later on. Finally, recall that index $d_k \in [N]$ corresponds to the file index of user k's demand.

In the following we describe how, for every transmission slot, the transmitter first creates a vector of L XORs, and then precodes each XOR with the appropriate precoder.

a) Individual XOR design: As previously mentioned, each transmitted XOR has $t/L + 1$ recipients, which we refer to as set μ. We recall that each subfile is cached at t receivers, and we consider the set ν to be the set of $t - t/L = t\frac{L-1}{L}$ users who have cached the set of files intended for users in set μ. In particular, these two sets $\mu, \nu \subset [K]$, are disjoint ($\mu \cap \nu = \emptyset$), and their sizes are $|\mu| = \frac{t}{L} + 1$ and $|\nu| = t\frac{L-1}{L}$ respectively. We also consider set $\sigma \subseteq \Big([K] \setminus (\mu \cup \nu)\Big)$, $|\sigma| = L - 1$, which will be later chosen more carefully. With these in place, we construct XOR[7] $X_\mu^{\nu,\sigma}$ as

$$X_\mu^{\nu,\sigma} = \bigoplus_{k\in\mu} W_{\sigma,(\nu\cup\mu)\setminus\{k\}}^{(d_k)} \tag{23}$$

which consists of $\frac{t}{L} + 1$ subfiles, where

- each subfile in the XOR is requested by one user in μ, and where
- all subfiles of the XOR are known by all users in ν.

The set $(\nu \cup \mu) \setminus \{k\}$ plays the role of τ from the placement phase, as it describes the set of users that have this subfile (labeled by τ) in their cache, while set σ is a selected subset of $L - 1$ users from set λ.

***Example* 3.** *Let us consider the MISO BC with $L = 2$ transmit antennas, $K \geq 6$ users and normalized cumulative cache size $t = 4$. Let the aforementioned sets be $\mu = \{1, 2, 3\}$, $\nu = \{4, 5\}$, and consider some arbitrary $\sigma \subseteq [K] \setminus \{1, 2, 3, 4, 5\}$, $|\sigma| = 1$. Then, the XOR of* (23) *takes the form*

$$X_{123}^{45,\sigma} = W_{\sigma,\underbrace{2345}_{\tau}}^{(d_1)} \oplus W_{\sigma,1345}^{(d_2)} \oplus W_{\sigma,1245}^{(d_3)}. \tag{24}$$

As we have described before, this XOR delivers subfiles desired by all the users of set μ, while each element of the XOR is cached at all users of set ν. It is easy to see that users $1, 2$, and 3 work in the traditional way to cache out the interfering subfiles in order to get their own desired subfile, such that for example user 1 caches out $W_{\sigma,1345}^{(d_2)} \oplus W_{\sigma,1245}^{(d_3)}$ to get its own $W_{\sigma,2345}^{(d_1)}$. In turn, users 4 and 5 are fully protected against this entire undesired XOR because they have cached all 3 subfiles of this XOR. As a quick verification, we see that each index τ has size $|\tau| = t = 4$, which adheres to the available cache-size constraint as each file can be stored at exactly $t = 4$ receivers.

b) Design of vector of XORs: Equipped with the design of each individual XOR, the goal is to select L such XORs in order to communicate them in a single transmission slot. Algorithm 1 forms a set of $L + t$ users and a set of L distinct such XORs to serve them with. Specifically, the steps that are followed are described below.

- In Step 1, a set λ of L users is chosen.
- In Step 2, a (ZF-type) precoder $\mathbf{H}_\lambda^{-1}$ is designed to spatially separate the L users in λ.
- In Step 3, another set $\pi \subseteq [K] \setminus \lambda$ of t users is selected from the remaining users.

To construct the L XORs and to properly place them in the vector, the following steps take place.

- In Step 4, set π of t users is arbitrarily partitioned into L non-overlapping sets $\phi_i, \ i \in [L]$, each having $\frac{t}{L}$ users.
- Steps 5 and 6 are responsible for forming the L different sets μ (cf. (23)), where each such set μ consists of $\frac{t}{L} + 1$ users. Specifically, in every iteration of Step 5, the

[7] In a small abuse of notation, we will henceforth refer to the segments of the original subfiles again as subfiles. We also note that, for clarity of exposition and to avoid many indices, index r of (22) will henceforth be suppressed, and thus any $W_{\sigma,\tau}^{(r,d_k)}$ will be denoted as $W_{\sigma,\tau}^{(d_k)}$ unless r is explicitly needed.

Algorithm 1: Delivery Phase

for $\lambda \subset [K], |\lambda| = L$ *(precoded users in* λ*)* **do**
  Calculate $\mathbf{H}_{\lambda}^{-1}$
  for $\pi \subseteq ([K] \setminus \lambda)\,,\ |\pi| = t$ **do**
    Break π into some $\phi_i\ i \in [L]:\ |\phi_i| = \frac{t}{L}$, $\bigcup_{i\in[L]} \phi_i = \pi,\ \phi_i \cap \phi_j = \emptyset, \forall i,j \in [L]$
    for $s \in \{0, 1, ..., L-1\}$ **do**
      $v_i = ((s + i - 1) \mod L) + 1, i \in [L]$
      Transmit

$$\mathbf{x}_{\lambda,\pi}^{s} = \mathbf{H}_{\lambda}^{-1} \cdot \begin{bmatrix} X_{\lambda(1)\cup\phi_{v_1}}^{\pi\setminus\phi_{v_1},\lambda\setminus\lambda(1)} \\ X_{\lambda(2)\cup\phi_{v_2}}^{\pi\setminus\phi_{v_2},\lambda\setminus\lambda(2)} \\ \vdots \\ X_{\lambda(L)\cup\phi_{v_L}}^{\pi\setminus\phi_{v_L},\lambda\setminus\lambda(L)} \end{bmatrix}. \quad (25)$$

algorithm associates a user from set λ with some set ϕ_{v_i}, in order to form set μ and such that after L iterations each user from λ would be associated with every set ϕ_{v_i}. For example, when $s = 0$, the first XOR of the vector will be intended for users in set $\{\lambda(1)\} \cup \phi_1$ (while completely known by all users in $\pi \setminus \phi_1$), the second XOR will be intended for the users in the set $\{\lambda(2)\} \cup \phi_2$ (while completely known by all users in $\pi \setminus \phi_2$), and so on. Further, when $s = 1$ the first XOR will be intended for users in $\{\lambda(1)\} \cup \phi_2$ (while completely known by all users in $\pi \setminus \phi_2$), the second XOR will be for users in $\{\lambda(2)\} \cup \phi_3$ (while completely known by all users in $\pi \setminus \phi_3$), and so on. In particular, Step 5 (and the operation in Step 6, as shown in Algorithm 1) allows us to iterate over all sets ϕ_i, associating every time a distinct set ϕ_i to a distinct user from group λ, until all users from set λ have been associated with all sets ϕ_i. The verification that this association does not leave behind any subfiles is performed later on in this section.

- Then, in the last step (Step 7), the vector of the L XORs is transmitted after being precoded by matrix $\mathbf{H}_{\lambda}^{-1}$.

c) Decoding at the users: By the very nature of the XOR design, as seen in (23), the vector of XORs we constructed in (25) guarantees that the users in λ can decode the single XOR that they receive (recall that for such users, all other XORs are steered away due to ZF precoding) and can thus subsequently proceed to decode their own file through the use of their cached content. Further, the design guarantees that each user in π has cached all subfiles that are found in the entire vector, apart from its desired subfile. Benefitting from their receiver-side CSI (see Appendix II-C), the users of set π are provided with all the necessary CSI estimates, which allows for the decoding of the linear combination of the transmitted vector.

To see the above more clearly, let us look at the signal received and the subsequent decoding process at some of the users. For some user $\ell \in \lambda$, the decoding process is simple. The received message takes the form

$$y_{\ell} = \mathbf{h}_{\ell}^{\dagger}\mathbf{H}_{\lambda}^{-1} \begin{bmatrix} X_{\lambda(1)\cup\phi_{v_1}}^{\pi\setminus\phi_{v_1},\lambda\setminus\lambda(1)} \\ X_{\lambda(2)\cup\phi_{v_2}}^{\pi\setminus\phi_{v_2},\lambda\setminus\lambda(2)} \\ \vdots \\ X_{\lambda(L)\cup\phi_{v_L}}^{\pi\setminus\phi_{v_L},\lambda\setminus\lambda(L)} \end{bmatrix} = X_{\{\ell\}\cup\phi_{v_k}}^{\pi\setminus\phi_{v_k},\lambda\setminus\{\ell\}},$$

where ϕ_{v_k}, $k \in [L]$, represents the subset of π, of size $|\phi_{v_k}| = \frac{t}{L}$, associated with ℓ (Step 5 of Algorithm 1). The selected precoders allow user ℓ to receive only one of the XORs (cf. (26)). Due to the design of this remaining XOR (see (23)), all but one subfiles have been cached by user ℓ, and thus the user can decode its desired subfile.

On the other hand, the decoding process at some user in set π requires, also, access to CSI. The received message at user $p \in \pi$ takes the form

$$y_p = \mathbf{h}_p^{\dagger}\mathbf{H}_{\lambda}^{-1} \begin{bmatrix} X_{\lambda(1)\cup\phi_{v_1}}^{\pi\setminus\phi_{v_1},\lambda\setminus\lambda(1)} \\ X_{\lambda(2)\cup\phi_{v_2}}^{\pi\setminus\phi_{v_2},\lambda\setminus\lambda(2)} \\ \vdots \\ X_{\lambda(L)\cup\phi_{v_L}}^{\pi\setminus\phi_{v_L},\lambda\setminus\lambda(L)} \end{bmatrix} \quad (26)$$

$$= \sum_{j=1}^{L} \mathbf{h}_p^{\dagger}\mathbf{h}_{\lambda\setminus\lambda(j)}^{\perp} X_{\lambda(j)\cup\phi_{v_j}}^{\pi\setminus\phi_{v_j},\lambda\setminus\lambda(j)}. \quad (27)$$

First, we observe that, due to the process described in Appendix II-C, user p has estimated all products $\mathbf{h}_p^{\dagger}\mathbf{h}_{\lambda\setminus\{\ell\}}^{\perp},\ \forall \ell \in \lambda$, that appear in (27). Then, by taking account of the fact that $\phi_{v_i} \cap \phi_{v_j} = \emptyset$ if $i \neq j$, we can see that user p belongs to one of the sets $\phi_{v_j} \subset \pi$. This means that user p has stored the content of all but one XORs (see (23)) and can thus remove them from (27). By removing the $L-1$ known XORs, the remaining message at user p is

$$\mathbf{h}_p^{\dagger}\mathbf{h}_{\lambda\setminus\lambda(j)}^{\perp} X_{\lambda(j)\cup\phi_{v_j}}^{\pi\setminus\phi_{v_j},\lambda\setminus\lambda(j)} \quad (28)$$

where $\phi_{v_j} \ni p$. Due to its structure (cf. (23)), the XOR can be successfully used by user p to decode its own desired message.

C. Evaluating the scheme's performance

In order to calculate the achievable DoF of the proposed scheme, we begin by showing that each desired subfile of set $\{W_{\sigma,\tau}^{r,(d_k)}\}_{r=1}^{L+t}$ is transmitted exactly once. Since each such collection of subfiles has the same sub-indices, it follows that there is no need to distinguish between them, as long as each appears exactly once.

a) Each desired subfile is transmitted exactly once: For any arbitrary subfile $W_{\sigma,\tau}^{(d_k)}$, the labeling (σ, τ, k) defines the set of active users $\lambda \cup \pi = \sigma \cup \tau \cup \{k\}$. Let us recall that $\lambda \cap \pi = \emptyset$, $\sigma \cap \tau = \emptyset$, that $\sigma \subset \lambda$ and that $|\sigma| = L-1$, $|\lambda| = L$,

$|\pi| = |\tau| = t$. For our fixed σ, τ, k, let us consider the two complementary cases; case i) $k \in \lambda$, and case ii) $k \notin \lambda$.

In case i), $\lambda = \sigma \cup \{k\}$, since $\tau \cap \lambda = \emptyset$. Moreover,

$$\pi = (\sigma \cup \tau \cup \{k\}) \setminus \lambda = \tau$$

means that a fixed (σ, τ, k) corresponds to a single (λ, π). For any fixed (λ, π) in Algorithm 1, Step 5 iterates L times, thus identifying L specific component subfiles which are defined by the same (σ, τ, k) and therefore can be differentiated by L different $r \in [t + L]$; these L component subfiles of $W_{\sigma,\tau}^{(d_k)}$ will appear in transmissions $\mathbf{x}_{\lambda,\pi}^{s}$, $s = 0, 1, \ldots, L-1$.

In case ii), the fact that $k \notin \lambda$ implies that for a given (σ, τ, k) (which also defines the set of active users) there can be t different sets λ which take the form

$$\lambda = \sigma \cup \{\tau(i)\}, \quad i \in [t].$$

This means that any fixed triplet (σ, τ, k) corresponds to t different possible sets λ. Since for a fixed (σ, τ, k), the union of $\lambda \cup \pi$ is fixed, we can conclude that each fixed (σ, τ, k) is associated to t different pairs (λ, π).

Now, having chosen a specific pair (λ, π), where we remind that $k \in \pi$, we can see from Step 5 of Algorithm 1 that user k should belong to exactly one set $\phi_{v_i}, i \in [L]$. Let that set be ϕ_{v_j}. This means that from all L transmissions of Step 5, a component subfile of the form $W_{\sigma,\tau}^{(d_k)}$ will be transmitted in exactly one transmission, and in particular, in the single transmission which includes XOR

$$X_{\tau(i)\cup\phi_{v_j}}^{\pi\setminus\phi_{v_j},\sigma}.$$

In total, for all the different (λ, π) sets, subfile $W_{\sigma,\tau}^{(d_k)}$ will be transmitted $L + t$ times. Finally, since we showed that an arbitrary subfile, $W_{\sigma,\tau}^{(d_k)}$, will be transmitted exactly $L+t$ times, this implies that all subfiles of interest will be transmitted once we go over all possible λ, π sets.

b) DoF calculation: The resulting DoF can now easily be seen to be $L+t$, simply because each transmission slot includes $L + t$ different subfiles, and because each file was indeed transmitted exactly once. A quick verification, accounting for the subpacketization

$$\mathcal{S}_L = \binom{K}{t}\binom{K-t-1}{L-1}(L+t),$$

and accounting for the number of iterations in each step, tells us that the worst-case delivery time takes the form

$$\mathcal{T}_L(t) = \frac{\overbrace{\binom{K}{L}}^{\text{Step 1}}\overbrace{\binom{K-L}{t}}^{\text{Step 3}}\cdot\overbrace{L}^{\text{Step 5}}}{\binom{K}{t}\binom{K-t-1}{L-1}(L+t)} = \frac{K-t}{L+t}, \tag{29}$$

which in turn directly implies a DoF of

$$\mathcal{D}_L(t) = \frac{K(1-\gamma)}{\mathcal{T}_L(t)} = L + t$$

which is achieved with CSI from only $C = L$ users per transmission. □

To illustrate the above algorithm, we proceed to present the delivery phase for the setting of Example 2.

***Example* 4** (Example of scheme)**.** *Consider a transmitter with $L = 2$ antennas, serving $K = 4$ users with normalized cumulative cache size $t = 2$. Each file is split into*

$$\mathcal{S}_L = \overbrace{(t+L)}^{r}\overbrace{\binom{K-t-1}{L-1}}^{\sigma}\overbrace{\binom{K}{t}}^{\tau} = 24$$

subfiles. The $\binom{K}{L}\binom{K-L}{t}L = 12$ transmission slots that satisfy all the users' requests are

$$\mathbf{x}_{12,34}^1 = \mathbf{H}_{12}^{-1}\begin{bmatrix} A_{2,34}^{(1)} \oplus C_{2,14}^{(1)} \\ B_{1,34}^{(1)} \oplus D_{1,23}^{(1)} \end{bmatrix}, \mathbf{x}_{12,34}^2 = \mathbf{H}_{12}^{-1}\begin{bmatrix} A_{2,34}^{(2)} \oplus D_{2,13}^{(1)} \\ B_{1,34}^{(2)} \oplus C_{1,24}^{(1)} \end{bmatrix}$$

$$\mathbf{x}_{34,12}^1 = \mathbf{H}_{34}^{-1}\begin{bmatrix} B_{4,13}^{(1)} \oplus C_{4,12}^{(1)} \\ A_{3,24}^{(1)} \oplus D_{3,12}^{(1)} \end{bmatrix}, \mathbf{x}_{34,12}^2 = \mathbf{H}_{34}^{-1}\begin{bmatrix} A_{4,23}^{(1)} \oplus C_{4,12}^{(2)} \\ B_{3,14}^{(1)} \oplus D_{3,12}^{(2)} \end{bmatrix}$$

$$\mathbf{x}_{24,13}^1 = \mathbf{H}_{24}^{-1}\begin{bmatrix} A_{4,23}^{(2)} \oplus B_{4,13}^{(2)} \\ C_{2,14}^{(2)} \oplus D_{2,13}^{(2)} \end{bmatrix}, \mathbf{x}_{24,13}^2 = \mathbf{H}_{24}^{-1}\begin{bmatrix} B_{4,13}^{(3)} \oplus C_{4,12}^{(3)} \\ A_{2,34}^{(3)} \oplus D_{2,13}^{(3)} \end{bmatrix}$$

$$\mathbf{x}_{13,24}^1 = \mathbf{H}_{13}^{-1}\begin{bmatrix} A_{3,24}^{(2)} \oplus B_{3,14}^{(2)} \\ C_{1,24}^{(2)} \oplus D_{1,23}^{(2)} \end{bmatrix}, \mathbf{x}_{13,24}^2 = \mathbf{H}_{13}^{-1}\begin{bmatrix} A_{3,24}^{(3)} \oplus D_{3,12}^{(2)} \\ B_{1,34}^{(3)} \oplus C_{1,24}^{(3)} \end{bmatrix}$$

$$\mathbf{x}_{14,23}^1 = \mathbf{H}_{14}^{-1}\begin{bmatrix} A_{4,23}^{(3)} \oplus B_{4,13}^{(4)} \\ D_{1,23}^{(3)} \oplus C_{1,24}^{(4)} \end{bmatrix}, \mathbf{x}_{14,23}^2 = \mathbf{H}_{14}^{-1}\begin{bmatrix} A_{4,23}^{(4)} \oplus C_{4,12}^{(4)} \\ B_{1,34}^{(4)} \oplus D_{1,23}^{(4)} \end{bmatrix}$$

$$\mathbf{x}_{23,14}^1 = \mathbf{H}_{23}^{-1}\begin{bmatrix} A_{3,24}^{(4)} \oplus B_{3,14}^{(3)} \\ C_{2,14}^{(3)} \oplus D_{2,13}^{(4)} \end{bmatrix}, \mathbf{x}_{23,14}^2 = \mathbf{H}_{23}^{-1}\begin{bmatrix} B_{3,14}^{(4)} \oplus D_{3,12}^{(4)} \\ C_{2,14}^{(4)} \oplus A_{2,34}^{(4)} \end{bmatrix}.$$

As we see, the delay is $\mathcal{T}_2 = \frac{12}{24} = \frac{1}{2}$ and the DoF is $\mathcal{D}_2 = \frac{K(1-\gamma)}{\mathcal{T}_2} = 4$. This performance is optimal.

V. Extension to the Multi-Transmitter Environment

We now consider the multiple-transmitter case, where each of the K_T transmitters is equipped with $L_T \geq 1$ antennas, and each has a cache capacity equal to a fraction $\gamma_T \in [1/K_T, 1]$ of the library, such that the transmitter-side normalized cumulative cache size is $t_T = K_T\gamma_T$. As we have seen, by denoting $L \triangleq L_T t_T$ we can draw a direct comparison between the two settings (the cache-aided MISO BC, and the corresponding cache-aided multi-transmitter setting) showing that they share the same DoF performance $\mathcal{D}_L(t, C) = t + C$, under the same feedback requirement C.

The scheme for the multi-transmitter setting closely resembles the scheme presented in Algorithm 1, with the difference being that precoding vectors $\mathbf{h}_{\lambda\setminus\{\lambda(\ell)\}}^{\perp}$ are formed in a distributed manner. In particular, for each transmitted subfile, the t_T transmitters who have access to that subfile must cooperate to form (each using its own L_T antennas) a distributed precoder vector of length L, which possesses the attributes described in (18). The only modification to Algorithm 1 is in the precoder design (Step 2) where the transmission vector (cf. (25)) now takes the form

$$\mathbf{x}_{\lambda,\pi}^{s} = \sum_{\ell=1}^{L}\sum_{k\in\{\lambda(\ell)\}\cup\phi_{u_\ell}} \mathbf{h}_{\lambda\setminus\{\lambda(\ell)\}}^{\perp} W_{\{\lambda(\ell)\}\cup\pi\setminus\{k\}}^{(d_k)}. \tag{30}$$

It is important to notice that for a specific $\ell \in [L]$, the respective precoder vector $\mathbf{h}_{\lambda\setminus\{\lambda(\ell)\}}^{\perp}$ is designed at the $t_T = \frac{L}{L_T}$

transmitters which have stored subfile $W^{(d_k)}_{\{\lambda(\ell)\}\cup\pi\setminus\{k\}}$. This further means that the precoding vectors $\mathbf{h}^{\perp}_{\lambda\setminus\{\lambda(\ell)\}}$ are subfile-dependent and thus potentially different.

Placement at the transmitters: To guarantee that each subfile is stored at exactly t_T transmitters, we use the approach of [39] which does not require an increase of the subpacketization, and which we include here for completeness. The placement algorithm starts from the first transmitter and caches the first $M_T = \gamma_T N$ files in their entirety, while the second transmitter caches the next set of M_T files, and so on. Specifically, transmitter $k_T \in [K_T]$ caches

$$\mathcal{Z}^{\text{Tx}}_{k_T} = \left\{W^{(n)}, \quad n \in \{1 + (k_T - 1)M_T, ..., k_T M_T\}\right\}, \quad (31)$$

where we note that the index of each file is calculated using the modulo operation, i.e., each file index $n \in [N]$ appearing in (31) takes the form $n = (n-1) \mod (N) + 1$. All the other steps remain the same.

VI. Converse

In this section, we prove the converse part of Theorem 2, corresponding to the multi-transmitter environment, and, by extension, the converse part of Theorem 1, which can be deduced by setting the problem parameters as $K_T = 1$, $L_T = L$, and $t_T = 1$.

The bound draws partly from [34], mainly for the initial steps, but we introduce new ideas that allow us to capture the CSI-availability effect as well as to introduce a new bounding solution for the optimization problem that directly tightens the converse. Similarly to [34], we are constrained to i) placement done under the assumption of uncoded prefetching, and ii) linear delivery schemes that have the one-shot property, where no data is transmitted more than once.

Specifically, the steps that we implement to prove the converse part of Theorem 2 are as follows:

1) We bound the number of messages that can be simultaneously transmitted under feedback constraints.
2) We rewrite the problem as an integer optimization problem that seeks to minimize the delivery time for a given prefetching policy and file demand vector.
3) We obtain a novel solution of the optimization problem by leveraging the cache-size constraints and the convexity of the problem.

The second step, i.e., the formulation of the optimization problem, follows from [34, Sections V.B, V.C], and we include it here for completeness. On the other hand, the novelty lies on the first and the third steps, which are instrumental in obtaining the converse.

We begin by introducing some additional definitions and notation. In the following, we consider a slightly different channel model with respect to the one described in Section II for the achievable scheme. Let us remark that these modifications do not impact our results, and indeed they are irrelevant for the description of the achievable scheme. On this basis, we have omitted these considerations before for the sake of clarity, and they are incorporated only in this section.

A. Preliminary definitions

We denote the superset of all the sets of caches at the users as ζ^{Rx}, such that $\zeta^{\text{Rx}} \triangleq \{\mathcal{Z}_1, \ldots, \mathcal{Z}_K\}$. Similarly, the superset of cached content stored at the transmitters is denoted by $\zeta^{\text{Tx}} \triangleq \{\mathcal{Z}^{\text{Tx}}_1, \ldots, \mathcal{Z}^{\text{Tx}}_{K_T}\}$. We consider that every file $W^{(n)}$ in the library $\mathcal{F}$ is divided into F packets, $\{W^{(n),f}\}_{f=1}^{F}$, each of size B bits. The caching is done at the level of packets and we do not allow breaking the packets into smaller sub-packets[8]. As is standard, we consider that the transmitters encode each packet $W^{(d_k),f}$ into a coded packet[9] $\tilde{W}_s^{(d_k)} \triangleq g(W_s^{(d_k)})$ of $\tilde{B}$ complex symbols using a random Gaussian coding scheme $g : \mathbb{F}_2^B \to \mathbb{C}^{\tilde{B}}$ of rate $\log P + o(\log P)$. We introduce in the following some definitions that are instrumental to the proof.

Definition 2 (Communication Block)**.** *A Communication Block is defined as the time required to transmit a packet — which has size equal to the atomic unit — to a single user, in the absence of caching and of interference. A block consists of $\frac{\tilde{B}}{\log P}$ time instants.*

Hence, for a certain demand vector $\mathbf{d}$, we consider that the transmission lasts for a set β of communication blocks, where each block $b \in \beta$ has duration $\tilde{B}$ time slots. During a given communication block b, the transmitters send a set of packets, denoted as ρ_b, to a subset of users $\kappa_b \subseteq [K]$ such that every user in κ_b desires only one packet from ρ_b. The file requested by user k is denoted as $W^{(d_k)}$, and the specific packet of $W^{(d_k)}$ that is transmitted in this communication block is denoted as $W^{(d_k),f_k}$. Note that, for the sake of readability, we omit the reference to the specific communication block in which the packet is scheduled. Thus, the set of transmitted packets is explicitly given by $\rho_b = \left\{W^{(d_k),f_k}\right\}_{k\in\kappa_b}$, $f_k \in [F]$, $d_k \in [N]$. Furthermore, the transmitters must transmit every packet of the file $W^{(d_k)}$ that is not cached by user k throughout the $|\beta|$ communication blocks. The transmission will last until all the required packets are correctly received.

The goal of the converse is to bound the minimum number of communication blocks required to transmit the demanded files $\{W^{(d_1)}, \ldots, W^{(d_K)}\}$ assuming the worst-case demand vectors. To this end, we first consider the optimal delivery time for a given placement. Specifically, for a given prefetching policy $(\zeta^{\text{Tx}}, \zeta^{\text{Rx}})$, the minimum one-shot linear delivery time achievable for the worst-case demand is defined as

$$\mathcal{T}(\zeta^{\text{Tx}}, \zeta^{\text{Rx}}) \triangleq \sup_{\{d_1,\ldots,d_K\}} \inf_{\substack{\beta \\ \{\rho_b\}_{b\in\beta}}} \frac{1}{F}|\beta|. \quad (32)$$

Note that the delivery time is normalized with respect to the file-size, such that a single unit of the delivery time corresponds to F communication blocks. By the same token, we can define the worst-case optimal delivery time as follows.

Definition 3 (Worst-case Delivery Time)**.** *In a K-user fully-connected wireless network with K_T transmitters, with L_T*

[8]Packets are considered to be the atomic unit of size in place of bits, such that they are big enough for the laws of Shannon to apply and the probability of decoding error to vanish as B increases [34], [48]. Regarding the description of the achievable scheme in Section IV, it can be assumed, without loss of generality, that F is an integer multiple of the number of subfiles.

[9]Here, "coded" refers to the channel coding strategy, and should not be confused with coded prefetching, which is not considered in this paper.

antennas per transmitter, with normalized cumulative cache size t_T at the transmitters side and t at the receivers side, and upon defining $L = L_T t_T$, the worst-case optimal delivery time is defined as the minimum achievable one-shot linear delivery time over all caching realizations:

$$\mathcal{T}_L^\star(t) \triangleq \inf_{\zeta^{Tx},\zeta^{Rx}} \mathcal{T}(\zeta^{Tx},\zeta^{Rx}). \tag{33}$$

Further, we define the optimal (one-shot linear) DoF using the previous definition.

Definition 4 (Optimal Degrees-of-Freedom)**.** *In the cache-aided network of Definition 3, the optimal one-shot linear DoF performance takes the form*

$$\mathcal{D}_L^\star(t) \triangleq \frac{K(1-\gamma)}{\mathcal{T}_L^\star(t)}. \tag{34}$$

We recall that we seek to minimize the delivery time (or, equivalently, maximize the DoF performance) under constrained feedback resources where, in each communication block, the transmitters acquire feedback only for a subset of C users in total. Let us consider a particular communication block b. We denote the set of users for which there exists CSIT at communication block b as η_b, $\eta_b \subseteq \kappa_b$, $|\eta_b| = C$, and its complementary set as $\eta_b^{\mathsf{c}} \triangleq \kappa_b \backslash \eta_b$. Furthermore, we denote the sets of transmitters and receivers who have cached the packet $W^{(d_k),f_k}$ intended to receiver k as ϵ_k and δ_k, respectively.

For some set α, the indicator function is denoted by $\mathbb{1}_\alpha(k)$, such that $\mathbb{1}_\alpha(k) = 1$ if $k \in \alpha$ and 0 otherwise. Accordingly, we introduce

$$C'_k \triangleq C + \mathbb{1}_{\eta_b^{\mathsf{c}}}(k), \quad \forall k \in \kappa_b, \tag{35}$$

such that $C'_k = C$ if $k \in \eta_b$ and $C'_k = C+1$ if $k \notin \eta_b$. We will also use

$$L_k \triangleq \min(C'_k,\, L_T|\epsilon_k|), \tag{36}$$

such that L_k represents the minimum between the number of transmit antennas that have cached the packet intended to receiver k ($W^{(d_k),f_k}$) and the number of users for which there is CSIT available excluding user k. In other words, C'_k indicates the number of users for which the transmitters can use the CSIT so as to benefit from spatial multiplexing for packet $W^{(d_k),f_k}$. Further, we introduce the following definition.

Definition 5 (Packet Order)**.** *A packet is said to be of "*order (u,v)*" if it is stored in the cache of u different transmitters and v different users.*

B. Bounding the number of simultaneous packets

Now, we aim to bound the number of users that can be simultaneously served during a given communication block. This bound is presented in the following lemma.

Lemma 1. *Let us consider a single communication block $b \in \beta$, where each packet of set ρ_b is scheduled to be simultaneously transmitted to one of the users of set κ_b, such that $|\rho_b| = |\kappa_b| = K_b$. Assume that each transmitter has only access to the CSIT of every user of set $\eta_b \subseteq \kappa_b$, $|\eta_b| = C$, and that for every user k, $k \in \kappa_b$, the set of users that have cached the packet intended to user k is given by δ_k. For each intended packet to be successfully decoded at the appropriate receiver, the number of simultaneously transmitted packets must satisfy*

$$K_b \leq \min_{k\in\kappa_b} \left(L_k + |\delta_k|\right). \tag{37}$$

Proof. The proof is relegated to Appendix I. □

Corollary 2. *Consider some communication block b. A packet of order (u,v) can be simultaneously transmitted with* at most $\min(C, L_T u) + v - 1$ *other packets of the same order in order to be successfully decoded.*

Proof. The proof follows directly after substituting $|\delta_k|$ for v and $|\epsilon_k|$ for u in (37) of Lemma 1 for every $k \in \kappa_b$. Therefore, we obtain that

$$K_b \leq \min_{k\in\kappa_b} \min(C'_k,\, L_T u) + v = \min(C,\, L_T u) + v,$$

which proves Corollary 2. □

Next, we present the definition of the *feasible set of packets*, which is based on Lemma 1.

Definition 6 (Feasible Sets)**.** *Let a communication block b be characterized by the set ρ_b of packets to be transmitted, by the set κ_b of users for whom the packets are intended, and by the set η_b of users for whom there is CSIT. A set of packets ρ_b selected to be transmitted at communication block b is said to be* feasible *if it satisfies* (37) *in Lemma 1, i.e., if for every $k \in \kappa_b$ it holds that*

$$K_b \leq L_k + |\delta_k|. \tag{38}$$

Consider a subset of users $\delta \subseteq [K]$ and a subset of transmitters $\epsilon \subseteq [K_T]$. We define

$$\omega_{\epsilon,\delta}^{(n)} \triangleq \bigcap_{\substack{f\in[F]\\ e\in\epsilon, c\in\delta}} \{W^{(n),f} \cap \mathcal{Z}_e^{\mathrm{Tx}} \cap \mathcal{Z}_c\} \tag{39}$$

to be the set of packets of file $W^{(n)}$, $n \in [N]$, that are exclusively stored in the caches of the transmitters in ϵ and the users in δ. Further, the number of packets in the set $\omega_{\epsilon,\delta}^{(n)}$ is denoted by $a_{\epsilon,\delta}^{(n)}$.

C. Lower-bound on the number of communication blocks

In this section, we lower-bound the number of communication blocks that are required for a successful transmission. This lower bound is based on a linear program that was first stated in [34]. The formulation of the linear program matches that of [34], and it is presented in Appendix II-A for completeness.

Let us consider first a given demand vector $\mathbf{d}$ and cache-placement strategies ζ^{Tx}, ζ^{Rx}. The minimum number of communication blocks $|\beta|$ required to successfully transmit all the requested files in $\mathbf{d}$ for the specific strategies ζ^{Tx}, ζ^{Rx} is denoted as $T_\beta^\star(\zeta^{\mathrm{Tx}},\zeta^{\mathrm{Rx}},\mathbf{d})$ and is rigorously defined in Appendix II-A.

We are interested in lower-bounding the value of $T_\beta^\star(\zeta^{\mathrm{Tx}},\zeta^{\mathrm{Rx}},\mathbf{d})$ for any worst-case demand $\mathbf{d}$. As shown in [34] (see also [5], [48]), the solution to the optimization problem can be lower-bounded by averaging over all the possible

permutations of the demand vector $\mathbf{d}$. Hence, for a given cache-placement strategy ζ^{Tx}, ζ^{Rx}, let us define

$$\bar{T}^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}}) \triangleq \frac{1}{|\psi(N,K)|}\sum_{\mathbf{d}\in\psi(N,K)} T^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}},\mathbf{d}) \quad (40)$$

to be the average number of required communication blocks over the set of all possible worst-case demand-vectors $\mathbf{d}$. In the above, $\psi(N,K)$ denotes the set of all K-permutations of the library files (N indices), and recall that $|\psi(N,K)| = \frac{N!}{(N-K)!}$.

We focus now on lower-bounding $\bar{T}^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}})$. Recalling Corollary 2, a packet of order (u,v) can be scheduled with at most $\min(L_T u, C) + v - 1$ packets of the same order. Consequently, for any $\epsilon \subseteq [K_T]$, $\delta \subseteq [K]$, such that $|\epsilon| = u$ and $|\delta| = v$, the maximum possible DoF for any packet in any set $\omega^{(n)}_{\epsilon,\delta}$, $n \in [N]$, is $\min\big(\min(L_T u, C)+v, K\big)$. From this bound over the maximum DoF, we can bound the minimum number of communication blocks needed for a specific demand and cache-placement strategy. Specifically, let us first note that, in order to transmit all the packets in a set $\omega^{(d_j)}_{\epsilon,\delta}$ satisfying that $|\epsilon| = u$ and $|\delta| = v$, we need at least $a^{(d_j)}_{\epsilon,\delta}/\min(\min(C, L_T u) + v, K)$ communication blocks. Upon defining

$$g_{u,v} \triangleq \min(C, L_T u) + v \quad (41)$$

for the sake of compactness, we obtain the lower bound

$$T^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}},\mathbf{d}) \geq \sum_{v=0}^{K}\sum_{u=1}^{K_T}\sum_{j=1}^{K}\sum_{\substack{\epsilon\subseteq[K_T]\\|\epsilon|=u}}\sum_{\substack{\delta\subseteq[K]\\|\delta|=v\\ \delta\not\ni j}}\frac{a^{(d_j)}_{\epsilon,\delta}}{g_{u,v}}. \quad (42)$$

Incorporating (42) in (40) yields

$$\bar{T}^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}}) \geq \sum_{\mathbf{d}\in\psi(N,K)}\frac{1}{|\psi(N,K)|} \times\left(\sum_{v=0}^{K}\sum_{u=1}^{K_T}\sum_{j=1}^{K}\sum_{\substack{\epsilon\subseteq[K_T]\\|\epsilon|=u}}\sum_{\substack{\delta\subseteq[K]\\|\delta|=v\\ \delta\not\ni j}}\frac{a^{(d_j)}_{\epsilon,\delta}}{g_{u,v}}\right) \quad (43)$$

$$\geq \sum_{v=0}^{K}\sum_{u=1}^{K_T}\sum_{j=1}^{K}\sum_{\substack{\epsilon\subseteq[K_T]\\|\epsilon|=u}}\sum_{\substack{\delta\subseteq[K]\\|\delta|=v\\ \delta\not\ni j}}\frac{1}{N}\sum_{n=1}^{N}\frac{a^{(n)}_{\epsilon,\delta}}{g_{u,v}} \quad (44)$$

$$= \frac{1}{N}\sum_{v=0}^{K}\sum_{u=1}^{K_T}\frac{1}{g_{u,v}}\sum_{j=1}^{K}\sum_{\substack{\epsilon\subseteq[K_T]\\|\epsilon|=u}}\sum_{\substack{\delta\subseteq[K]\\|\delta|=v\\ \delta\not\ni j}}\sum_{n=1}^{N}a^{(n)}_{\epsilon,\delta}, \quad (45)$$

where (44) follows since, over the set of demand-vector permutations $\psi(N,K)$, every file $W^{(n)}$ is requested by every user j the same number of times. The last equality is obtained from a simple re-ordering of terms.

D. Tightening the lower-bound

The lower-bound in (45) is obtained by combining the approach in [34] with the novel outcome of Lemma 1 that accounts for the limited feedback constraint. Henceforth, we deviate from the approach in [34] so as to tighten the lower-bound. Let us consider the total number $a_{\epsilon,\delta}$ of packets stored exclusively at the transmitters in $\epsilon \subseteq [K_T]$ and at the receivers in set $\delta \subseteq [K]$. This number satisfies

$$a_{\epsilon,\delta} \triangleq \sum_{n=1}^{N} a^{(n)}_{\epsilon,\delta}. \quad (46)$$

Similarly, let $b_{\epsilon,v}$ denote the size of the set of packets stored exclusively by all transmitters in ϵ and at a total of v receivers. Then,

$$b_{\epsilon,v} \triangleq \sum_{\substack{\delta\subseteq[K]\\|\delta|=v}} a_{\epsilon,\delta}. \quad (47)$$

For a given set of transmitters ϵ and a given user-set size $|\delta| = v$, it follows that

$$\sum_{j=1}^{K}\sum_{\substack{\delta\subseteq[K]\\|\delta|=v\\ \delta\not\ni j}} a_{\epsilon,\delta} = (K-v)b_{\epsilon,v}. \quad (48)$$

In order to prove (48), let us consider a specific subset $\delta' \subseteq [K]$, $|\delta'| = v$. The number of packets cached at the transmitters of set ϵ and the users of set δ' is given by $a_{\epsilon,\delta'}$. For a given $j \in [K]$, the term $a_{\epsilon,\delta'}$ is included in the summation $\sum_{\delta\subseteq[K],\,|\delta|=v,\,\delta\not\ni j} a_{\epsilon,\delta}$ if and only if $j \notin \delta'$. Since (48) sums over all $j \in [K]$ and $|\delta'| = v$, the term $a_{\epsilon,\delta'}$ appears $K - v$ times in (48), one for each j satisfying that $j \notin \delta'$. From the fact that this holds for any $\delta' \subseteq [K]$ with $|\delta'| = v$, and from the definition of $b_{\epsilon,v}$ in (47), we obtain (48). Applying (48) into (45) yields

$$\bar{T}^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}}) \geq \frac{1}{N}\sum_{v=0}^{K}\sum_{u=1}^{K_T}\frac{K-v}{g_{u,v}}\sum_{\substack{\epsilon\subseteq[K_T]\\|\epsilon|=u}} b_{\epsilon,v} \quad (49)$$

$$= \frac{1}{N}\sum_{v=0}^{K}\sum_{u=1}^{K_T}\frac{K-v}{\min(C, L_T u)+v}b_{u,v}, \quad (50)$$

where in (50) we have applied (41) and we have introduced $b_{u,v}$ to denote the number of packets cached at u transmitters and v receivers. It is direct that

$$b_{u,v} \triangleq \sum_{\substack{\epsilon\subseteq[K_T]\\|\epsilon|=u}} b_{\epsilon,v}. \quad (51)$$

Note that $\bar{T}^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}})$ represents the necessary number of communication blocks to complete the transmission. From the definition of delivery time in (32), it follows that

$$\mathcal{T}(\zeta^{\text{Tx}},\zeta^{\text{Rx}}) = \frac{1}{F}\bar{T}^{\star}_{\beta}(\zeta^{\text{Tx}},\zeta^{\text{Rx}}), \quad (52)$$

where (52) simply translates the unit of measure to consider normalization by the file size instead of the packet size. From (50) and (52) we have that

$$\mathcal{T}(\zeta^{\text{Tx}},\zeta^{\text{Rx}}) \geq \frac{1}{FN}\sum_{v=0}^{K}\sum_{u=1}^{K_T}\frac{(K-v)}{\min(C, L_T u)+v}b_{u,v}. \quad (53)$$

Consequently, we have obtained a lower bound that depends only on the portion of the library that is cached at a specific

number of transmitters and the number of receivers, irrespectively of who has stored which packet.

For some function $c(\cdot,\cdot)$, we denote the lower convex envelope of the points

$$\{\big(t_1, t_2, c(t_1,t_2)\big) | t_1, t_2 \in \{0,\, 1,\, \ldots,\, K\}\},$$

by $\text{conv}(c(t_1,t_2))$. Let us introduce the notation $c(u,v) \triangleq \frac{K-v}{\min(C, L_T u)+v}$. Since $c(u,v)$ is a decreasing sequence in v and non-increasing in u, $\text{conv}(c(u,v))$ is a non-increasing and convex function [5]. Furthermore, we define the number of packets cached at u transmitters (resp. v receivers) as b_u^t (resp. b_v^r), i.e.,

$$b_u^t \triangleq \sum_{v=0}^{K} b_{u,v}, \tag{54}$$

$$b_v^r \triangleq \sum_{u=1}^{K_T} b_{u,v}. \tag{55}$$

Therefore, the cache-size constraints of the considered setting can be written as

$$\sum_{u=1}^{K_T}\sum_{v=0}^{K} b_{u,v} = \sum_{u=1}^{K_T} b_u^t = \sum_{v=0}^{K} b_v^r = NF, \tag{56}$$

$$\sum_{v=0}^{K} v b_v^r \leq FK\gamma N, \tag{57}$$

$$\sum_{u=1}^{K_T} u b_u^t \leq FK_T\gamma_T N. \tag{58}$$

The constraint in (56) ensures that every packet of the library is cached at some node (transmitter or receiver) in the network, while (57) corresponds to the cache size constraint at the users, and (58) corresponds to the cache size constraint at the transmitters. From the above, (53) can be lower-bounded as

$$\mathcal{T}(\zeta^{\text{Tx}}, \zeta^{\text{Rx}}) \geq \frac{1}{FN}\sum_{v=0}^{K}\sum_{u=1}^{K_T} \frac{(K-v)}{\min(C, L_T u)+v} b_{u,v} \tag{59}$$

$$\geq \text{conv}\big(c(K_T\gamma_T,\, t)\big) \tag{60}$$

$$= \text{conv}\left(\frac{K(1-\gamma)}{t+\min(C, K_T\gamma_T L_T)}\right), \tag{61}$$

where (60) comes from exploiting the convexity of the problem and from applying Jensen's Inequality. The detailed proof of how to reach (60) from (59) is relegated to Appendix II-B. Since $\mathcal{T}_L^\star(t,C) \triangleq \inf_{\zeta^{\text{Tx}},\zeta^{\text{Rx}}} \mathcal{T}(\zeta^{\text{Tx}}, \zeta^{\text{Rx}})$, the converse proof of Theorem 2 is concluded. □

VII. Conclusion

We have characterized the optimal one-shot linear DoF of the multi-antenna cache-aided broadcast channel and its multi-transmitter equivalent under limited feedback resources and uncoded placement, and we have provided a novel multi-antenna coded caching algorithm which we proved to be optimal. Our converse applies to a variety of other works, allowing the identification of their exact DoF performance.

Our results showed that achieving the maximum DoF performance only requires feedback from a limited number of users equal to the number of antennas. This further allows non-scaling feedback costs with respect to the number of users, as compared to previously known methods.

Various benefits of reducing feedback

This feedback reduction has multiple beneficial effects. Firstly, reducing the feedback requirements will allow for an increase of the effective DoF, simply because a bigger fraction of the coherence period is dedicated to communicating data rather than to feedback training. Secondly, the proposed algorithm allows for the increase of the overall number of users without a subsequent increase in feedback costs.

At the end of the day, our result makes a strong argument that caching can substantially ameliorate the well known feedback bottleneck of multi-antenna high-rate environments.

Appendix I
Proof of Lemma 1

We split the proof in two disjoint cases. We begin with the assumption that each transmitter has access to CSIT from every user scheduled to be served in the considered communication block[10]. Afterwards, we will introduce the CSIT constraint enforcing that the transmitter can only receive feedback from some C users.

Let us consider a single communication block. Without loss of generality, we assume that the K_b served users are the first K_b users, from 1 to K_b, and that the packets to be transmitted are $\{W^{(n),1}\}_{n=1}^{K_b}$. Under the one-shot and linear precoding assumptions, it follows that each transmitter sends a linear combination of the scheduled packets. In particular, the transmitted signal from a given transmitter j only carries information of the packets that it has cached, i.e., it only includes the users $i \in [K_b]$ for which $j \in \epsilon_i$. We define the global beamforming vector applied to the packet intended for user i as $\mathbf{p}_i \in \mathbb{C}^{L_T|\epsilon_i| \times 1}$, since only $|\epsilon_i|$ transmitters have cached $W^{(i),1}$.

A. Transmission of packets with CSIT from all scheduled users

We proceed in a similar way to [34, Lemma 3] by converting the MISO BC setting into a new MISO interference channel with K_b virtual transmitters, $\{\widehat{\text{TX}}_i\}_{i=1}^{K_b}$. $\widehat{\text{TX}}_i$ has $L_T|\epsilon_i|$ antennas and aims to transmit $W^{(i),1}$ to user $i \in [K_b]$. Note that the channel of different virtual transmitters is correlated because in the real physical channel the same antenna of a certain transmitter belongs to several virtual transmitters [34].

Let us denote the channel coefficients from virtual transmitter $\widehat{\text{TX}}_i$ to user k by $\mathbf{g}_{k,i} \in \mathbb{C}^{L_T|\epsilon_i| \times 1}$. Then, in an analogous way to the approach in [34], [65], it follows that the decodability conditions that must be satisfied[11] are

$$\mathbf{g}_{k,i}^\dagger \mathbf{p}_i = 0 \quad \forall i, k \in [K_b] : k \notin \delta_i \tag{62}$$

$$\mathbf{g}_{k,k}^\dagger \mathbf{p}_k \neq 0 \quad \forall k \in [K_b], \tag{63}$$

[10] While this setting is already studied in [34], we recall it here because it is an initial step towards the general feedback-constrained bound that we present.

[11] Since we are restricted to linear transmission schemes, the transmission block is not successful if these conditions are not satisfied, simply because the signal-to-interference ratio would not be enough to decode the intended message (cf. [34], [63]).

where we recall that δ_i is the subset of users that have cached the packet intended to user i.

Under the assumption that the transmitters have access to the CSI of all the K_b served users, we can rewrite the conditions in (62) as follows: First, we know from (63) that vectors $\mathbf{p}_i$ must have at least a non-zero coefficient, denoted by q_i. We can rotate the vector such that $\mathbf{p}_i = q_i \mathbf{P}_i \begin{bmatrix} 1 \\ \hat{\mathbf{v}}_i \end{bmatrix}$, where $\mathbf{P}$ is a permutation matrix and $\hat{\mathbf{v}}_i$ has size $(L|\epsilon_i| - 1 \times 1)$. Upon defining similarly $\hat{\mathbf{g}}_{k,i} \triangleq \mathbf{P}_i^{-1}\mathbf{g}_{k,i} = \begin{bmatrix} \hat{g}_{k,i}^{(1)} \\ \hat{\mathbf{g}}_{k,i}^{(2:)} \end{bmatrix}$, it follows that

$$\mathbf{g}_{k,i}^{\dagger}\mathbf{p}_i = (\mathbf{P}_i\hat{\mathbf{g}}_{k,i})^{\dagger} q_i \mathbf{P}_i \begin{bmatrix} 1 \\ \hat{\mathbf{v}}_i \end{bmatrix} \tag{64}$$

$$= q_i \left(\hat{g}_{k,i}^{(1)\dagger} + \hat{\mathbf{g}}_{k,i}^{(2:)\dagger} \hat{\mathbf{v}}_i \right) = 0, \tag{65}$$

where the last equality follows from (62).

Consider now the set $\delta_i^{\mathsf{c}} \triangleq [K_b] \setminus \big(\delta_i \cup \{i\}\big)$ of served users that neither cache nor desire $W^{(i),1}$. Let $m_i \triangleq |\delta_i^{\mathsf{c}}|$ and let $\delta_i^{\mathsf{c}}(n)$ denote the n-th user of δ_i^{c}. Since (65) has to hold for any $k \in \delta_i^{\mathsf{c}}$, we obtain the following linear system:

$$\underbrace{\begin{bmatrix} \hat{\mathbf{g}}_{\delta_i^{\mathsf{c}}(1),i}^{(2:)\dagger} \\ \hat{\mathbf{g}}_{\delta_i^{\mathsf{c}}(2),i}^{(2:)\dagger} \\ \vdots \\ \hat{\mathbf{g}}_{\delta_i^{\mathsf{c}}(m_i),i}^{(2:)\dagger} \end{bmatrix}}_{\mathbf{A}_{\delta_i^{\mathsf{c}}}} \hat{\mathbf{v}}_i = \underbrace{\begin{bmatrix} \hat{g}_{\delta_i^{\mathsf{c}}(1),i}^{(1)\dagger} \\ \hat{g}_{\delta_i^{\mathsf{c}}(2),i}^{(1)\dagger} \\ \vdots \\ \hat{g}_{\delta_i^{\mathsf{c}}(m_i),i}^{(1)\dagger} \end{bmatrix}}_{\boldsymbol{b}_{\delta_i^{\mathsf{c}}}} \tag{66}$$

where $\mathbf{A}_{\delta_i^{\mathsf{c}}} \in \mathbb{C}^{m_i \times (L_T|\epsilon_i|-1)}$ and $\boldsymbol{b}_{\delta_i^{\mathsf{c}}} \in \mathbb{C}^{m_i \times 1}$. Because (62) needs to be satisfied, the linear system in (66) must be solvable almost surely in order to guarantee the successful reception of all the messages. Since $\operatorname{rank}(\mathbf{A}_{\delta_i^{\mathsf{c}}}) = \min(m_i,\, L_T|\epsilon_i| - 1)$, it follows that $m_i \leq L_T|\epsilon_i| - 1$, implying that

$$K_b - |\delta_i| - 1 \leq L_T|\epsilon_i| - 1 \tag{67}$$

$$\implies \qquad K_b \leq L_T|\epsilon_i| + |\delta_i|. \tag{68}$$

Let us consider now an arbitrary communication block b during which a set of users κ_b, $|\kappa_b| = K_b$, is served. Given that (67) must hold for any $i \in \kappa_b$, we obtain that

$$K_b \leq \min_{i \in \kappa_b} L_T|\epsilon_i| + |\delta_i|, \tag{69}$$

from which we obtain Lemma 1 for the case without feedback constraints.

B. Transmission of packets with CSIT from only C users

Let us now consider the case in which the transmitters have CSIT only from a subset η of $|\eta| = C$ users, and recall that $\eta^{\mathsf{c}} = \kappa_b \backslash \eta$.

Since the transmitters do not know the channel towards the users belonging to the set η^{c}, the condition in (62) can not be satisfied with a high probability. In consequence, the transmitters can only use the CSIT from the users belonging to η, such that the solvable linear system becomes

$$\mathbf{A}_{\delta_i^{\mathsf{c}} \cap \eta} \hat{\mathbf{v}}_i = \boldsymbol{b}_{\delta_i^{\mathsf{c}} \cap \eta}, \tag{70}$$

where $\mathbf{A}_{\delta_i^{\mathsf{c}} \cap \eta}$ and $\boldsymbol{b}_{\delta_i^{\mathsf{c}} \cap \eta}$ are defined just like $\mathbf{A}_{\delta_i^{\mathsf{c}}}$ and $\boldsymbol{b}_{\delta_i^{\mathsf{c}}}$ in (66) except that now we only consider the users in $\{\delta_i^{\mathsf{c}} \cap \eta\}$ rather than in δ_i^{c}. Note that the set $\delta_i^{\mathsf{c}} \cap \eta$ is comprised of the users that have not cached the message for user i *and* for whom the transmitter has acquired CSIT.

We focus now on the required conditions that allow the successful reception of packets by each user in κ_b. From (70), it follows that the set δ_i^{c} must be a subset of the users for which there is CSIT available ($\delta_i^{\mathsf{c}} \subseteq \eta$) for any user $i \in \kappa_b$. This is due to the fact that, for any user i such that $i \notin \eta$, the lack of CSIT implies the impossibility of satisfying (62) and thus the impossibility of correctly decoding at user i [63]. Following the same reasoning as in (67), having $\delta_i^{\mathsf{c}} \subseteq \eta$ implies that $m_i \leq C$. However, note that it holds that $i \notin \delta_i^{\mathsf{c}}$ for any user i. Thus, $|\delta_i^{\mathsf{c}} \cap \eta| \leq C - 1 \;\; \forall i \in \eta$. Let $m_i^{\eta} \triangleq |\delta_i^{\mathsf{c}} \cap \eta|$ be the size of the intersection between the set of users not caching the packet intended for user i and the set of users for whom there is CSIT available. We then have

$$m_i^{\eta} \leq C - 1 \quad \Rightarrow \quad K_b \leq C + |\delta_i| \qquad \text{if } i \in \eta, \tag{71}$$

$$m_i^{\eta} \leq C \quad \Rightarrow \quad K_b \leq C + |\delta_i| + 1 \qquad \text{if } i \notin \eta. \tag{72}$$

We recall the notation introduced in (35), where for any $i \in \kappa_b$ we define $C_i' \triangleq C + \mathbb{1}_{\eta^{\mathsf{c}}}(i)$. Furthermore, the bound in (67) also holds, as it suffices to consider a genie that provides the CSIT of every user to the transmitters. Since (67) must hold for any $i \in \kappa_b$, we obtain that

$$K_b \leq \min_{i \in \kappa_b} \Big(\min(C_i',\, L_T|\epsilon_i|) + |\delta_i| \Big), \tag{73}$$

which concludes the proof of Lemma 1. □

APPENDIX II
ADDITIONAL PROOFS AND MATERIAL

A. Integer Program Formulation

Let us consider a given demand vector $\mathbf{d}$ and cache-placement strategies ζ^{Tx}, ζ^{Rx} at the transmitters and the receivers, respectively. Equipped with the definition of a feasible set of packets (cf. Definition 6), we write an integer program that seeks to minimize the number of required communication blocks for a specific ζ^{Tx}, ζ^{Rx}, $\mathbf{d}$, as follows.

$$\min_{\substack{\rho_b \\ b \in \beta}} \quad |\beta| \tag{P1-a}$$

$$\text{s.t.} \bigcup_{b \in \beta} \rho_b = \bigcup_{k \in [K]} \Big(W^{(d_k)} \backslash \mathcal{Z}_k \Big) \tag{P1-b}$$

$$\rho_b \text{ is feasible } \forall b \in \beta, \tag{P1-c}$$

where (P1-b) imposes the necessary condition that all the demanded packets that are not in the cache of the intended user must be transmitted. The solution to (P1) is denoted by $T_{\beta}^{\star}(\zeta^{\text{Tx}}, \zeta^{\text{Rx}}, \mathbf{d})$.

B. Transition from (59) to (60)

We have that

$$\frac{1}{FN}\sum_{v=0}^{K}\sum_{u=1}^{K_T} b_{u,v}\text{conv}\left(\frac{K-v}{\min(C, L_T u)+v}\right)$$

$$= \frac{1}{NF}\sum_{v=0}^{K}\sum_{u=1}^{K_T} b_{u,v}\text{conv}\left(c(u,v)\right) \tag{74}$$

$$\overset{(a)}{=} \frac{\sum_{v=0}^{K}\sum_{u=1}^{K_T} b_{u,v}\text{conv}\left(c(u,v)\right)}{\sum_{v=0}^{K}\sum_{u=1}^{K_T} b_{u,v}} \tag{75}$$

$$\overset{(b)}{\geq} \text{conv}\left(c\left(\frac{\sum_{u=1}^{K_T} u b_u^t}{\sum_{u=1}^{K_T} b_u^t}, \frac{\sum_{v=0}^{K} v b_v^r}{\sum_{v=0}^{K} b_v^r}\right)\right) \tag{76}$$

$$\overset{(c)}{=} \text{conv}\left(c\left(\frac{\sum_{u=1}^{K_T} u b_u^t}{NF}, \frac{\sum_{v=0}^{K} v b_v^r}{NF}\right)\right), \tag{77}$$

where (a) comes from (56), (b) from Jensen's Inequality, and (c) from (56) again. The monotonically decreasing nature of $c(u,v)$, combined with (57)-(58), yield

$$\text{conv}\left(c\left(\frac{\sum_{u=1}^{K_T} u b_u^t}{NF}, \frac{\sum_{v=0}^{K} v b_v^r}{NF}\right)\right) \tag{78}$$

$$\geq \text{conv}\left(c\left(\frac{FK_T\gamma_T N}{NF}, \frac{FK\gamma N}{NF}\right)\right) \tag{79}$$

$$= \text{conv}\left(c(t_T, t)\right). \tag{80}$$

By recovering (59), we can write that

$$\mathcal{T}(\zeta^{\text{Tx}}, \zeta^{\text{Rx}}) \geq \frac{1}{FN}\sum_{v=0}^{K}\sum_{u=1}^{K_T} b_{u,v}\text{conv}\left(\frac{(K-v)}{\min(C, L_T u)+v}\right)$$

$$\geq \text{conv}\left(\frac{K(1-\gamma)}{\min(C, L_T t_T)+t}\right), \tag{81}$$

which concludes the proof. □

C. Discussion on the CSI acquisition

The CSIT acquisition phase can be done in a standard way such that the L users (set λ) communicate pilots, which will allow the transmitter to estimate the channels of these users. As such, here we focus on the process of CSIR acquisition where the goal is to communicate at each user of set $\pi \cup \lambda$ the channel-precoder products. The process requires 1 training slot for each precoder, which amounts to L training slots per transmission.

Communication of precoder $\mathbf{h}_\mu^\perp$, where $\mu \subset [L]$ and $|\mu| = L-1$, takes the form

$$\mathbf{x}_\mu = \begin{bmatrix} \mathbf{h}_\mu^\perp(1)\mathbf{s}(1) \\ \vdots \\ \mathbf{h}_\mu^\perp(L)\mathbf{s}(L) \end{bmatrix} \tag{82}$$

where $\mathbf{s}$ denotes a single training vector.

The received message, ignoring the noise for simplicity, at some user $k \in [K]$, takes the form

$$y_k = \mathbf{h}_k^\dagger \mathbf{x}_\mu = \sum_{\ell=1}^{L} \mathbf{h}_k^\dagger(\ell)\mathbf{h}_\mu^\perp(\ell)\mathbf{s}(\ell), \tag{83}$$

from which the composite channel-precoder product $\mathbf{h}_k^\dagger \mathbf{h}_\mu^\perp$ can be calculated.

To summarize, CSIT requires L slots because only the L users need to transmit their channel state, and global CSIR requires L slots because, for each fixed precoder, one training symbol suffices to communicate the composite channel-precoder product to any number of users.

D. Extensive example of the proposed scheme

We conclude with two more examples that aim to help the reader gain a deeper understanding of the mechanics of our algorithm.

Example 5. *Let us consider the $L = 2$ MISO BC with $K = 6$ users and normalized cumulative cache of size $t = 4$. The required 30 transmissions to satisfy all users' demands are:*

$$\mathbf{x}_{12,3456}^1 = \mathbf{H}_{12}^{-1}\begin{bmatrix} A_{2,3456}^{(1)} \oplus C_{2,1456}^{(1)} \oplus D_{2,1356}^{(1)} \\ B_{1,3456}^{(1)} \oplus E_{1,2346}^{(1)} \oplus F_{1,2345}^{(1)} \end{bmatrix},$$

$$\mathbf{x}_{12,3456}^2 = \mathbf{H}_{12}^{-1}\begin{bmatrix} A_{2,3456}^{(2)} \oplus E_{2,1346}^{(1)} \oplus F_{2,1345}^{(1)} \\ B_{1,3456}^{(2)} \oplus C_{1,2456}^{(1)} \oplus D_{1,2356}^{(1)} \end{bmatrix},$$

$$\mathbf{x}_{13,2456}^1 = \mathbf{H}_{13}^{-1}\begin{bmatrix} A_{3,2456}^{(1)} \oplus B_{3,1456}^{(1)} \oplus D_{3,1256}^{(1)} \\ C_{1,2456}^{(2)} \oplus E_{1,2346}^{(2)} \oplus F_{1,2345}^{(2)} \end{bmatrix},$$

$$\mathbf{x}_{13,2456}^2 = \mathbf{H}_{13}^{-1}\begin{bmatrix} A_{3,2456}^{(2)} \oplus E_{3,1246}^{(1)} \oplus F_{3,1245}^{(1)} \\ C_{1,2456}^{(3)} \oplus B_{1,3456}^{(3)} \oplus D_{1,2356}^{(2)} \end{bmatrix},$$

$$\mathbf{x}_{14,2356}^1 = \mathbf{H}_{14}^{-1}\begin{bmatrix} A_{4,2356}^{(1)} \oplus B_{4,1356}^{(1)} \oplus C_{4,1256}^{(1)} \\ D_{1,2356}^{(3)} \oplus E_{1,2346}^{(3)} \oplus F_{1,2345}^{(3)} \end{bmatrix},$$

$$\mathbf{x}_{14,2356}^2 = \mathbf{H}_{14}^{-1}\begin{bmatrix} A_{4,2356}^{(2)} \oplus E_{4,1236}^{(1)} \oplus F_{4,1235}^{(1)} \\ D_{1,2356}^{(4)} \oplus B_{1,3456}^{(4)} \oplus C_{1,2456}^{(4)} \end{bmatrix},$$

$$\mathbf{x}_{15,2346}^1 = \mathbf{H}_{15}^{-1}\begin{bmatrix} A_{5,2346}^{(1)} \oplus B_{5,1346}^{(1)} \oplus C_{5,1246}^{(1)} \\ E_{1,2346}^{(4)} \oplus D_{1,2356}^{(5)} \oplus F_{1,2345}^{(4)} \end{bmatrix},$$

$$\mathbf{x}_{15,2346}^2 = \mathbf{H}_{15}^{-1}\begin{bmatrix} A_{5,2346}^{(2)} \oplus D_{5,1236}^{(1)} \oplus F_{5,1234}^{(1)} \\ E_{1,2346}^{(5)} \oplus B_{1,3456}^{(5)} \oplus C_{1,2456}^{(5)} \end{bmatrix},$$

$$\mathbf{x}_{16,2345}^1 = \mathbf{H}_{16}^{-1}\begin{bmatrix} A_{6,2345}^{(1)} \oplus B_{6,1345}^{(1)} \oplus C_{6,1245}^{(1)} \\ F_{1,2345}^{(5)} \oplus D_{1,2356}^{(6)} \oplus E_{1,2346}^{(6)} \end{bmatrix},$$

$$\mathbf{x}_{16,2345}^2 = \mathbf{H}_{16}^{-1}\begin{bmatrix} A_{6,2345}^{(2)} \oplus D_{6,1235}^{(1)} \oplus E_{6,1234}^{(1)} \\ F_{1,2345}^{(6)} \oplus B_{1,3456}^{(6)} \oplus C_{1,2456}^{(6)} \end{bmatrix},$$

$$\mathbf{x}_{23,1456}^1 = \mathbf{H}_{23}^{-1}\begin{bmatrix} B_{3,1456}^{(2)} \oplus A_{3,2456}^{(3)} \oplus D_{3,1256}^{(2)} \\ C_{2,1456}^{(2)} \oplus E_{2,1346}^{(2)} \oplus F_{2,1345}^{(2)} \end{bmatrix},$$

$$\mathbf{x}_{23,1456}^2 = \mathbf{H}_{23}^{-1}\begin{bmatrix} B_{3,1456}^{(3)} \oplus E_{3,1246}^{(2)} \oplus F_{3,1245}^{(2)} \\ C_{2,1456}^{(3)} \oplus A_{2,3456}^{(3)} \oplus D_{2,1356}^{(2)} \end{bmatrix},$$

$$\mathbf{x}_{24,1356}^1 = \mathbf{H}_{24}^{-1}\begin{bmatrix} B_{4,1356}^{(2)} \oplus A_{4,2356}^{(3)} \oplus C_{4,1256}^{(2)} \\ D_{2,1356}^{(3)} \oplus E_{2,1346}^{(3)} \oplus F_{2,1345}^{(3)} \end{bmatrix},$$

$$\mathbf{x}_{24,1356}^2 = \mathbf{H}_{24}^{-1}\begin{bmatrix} B_{4,1356}^{(3)} \oplus E_{4,1236}^{(2)} \oplus F_{4,1235}^{(2)} \\ D_{2,1356}^{(4)} \oplus A_{2,3456}^{(4)} \oplus C_{2,1456}^{(4)} \end{bmatrix},$$

$$\mathbf{x}_{25,1346}^1 = \mathbf{H}_{25}^{-1}\begin{bmatrix} B_{5,1346}^{(2)} \oplus A_{5,2346}^{(3)} \oplus C_{5,1246}^{(2)} \\ E_{2,1346}^{(4)} \oplus D_{2,1356}^{(5)} \oplus F_{2,1345}^{(4)} \end{bmatrix},$$

$$\mathbf{x}^2_{25,1346}=\mathbf{H}^{-1}_{25}\begin{bmatrix}B^{(3)}_{5,1346}\oplus D^{(2)}_{5,1236}\oplus F^{(2)}_{5,1234}\\ E^{(5)}_{2,1346}\oplus A^{(5)}_{2,3456}\oplus C^{(5)}_{2,1456}\end{bmatrix},$$

$$\mathbf{x}^1_{26,1345}=\mathbf{H}^{-1}_{26}\begin{bmatrix}B^{(2)}_{6,1345}\oplus A^{(3)}_{6,2345}\oplus C^{(2)}_{6,1245}\\ F^{(5)}_{2,1345}\oplus D^{(6)}_{2,1356}\oplus E^{(6)}_{2,1346}\end{bmatrix},$$

$$\mathbf{x}^2_{26,1345}=\mathbf{H}^{-1}_{26}\begin{bmatrix}B^{(3)}_{6,1345}\oplus D^{(2)}_{6,1235}\oplus E^{(2)}_{6,1234}\\ F^{(6)}_{2,1345}\oplus A^{(6)}_{2,3456}\oplus C^{(6)}_{2,1456}\end{bmatrix},$$

$$\mathbf{x}^1_{34,1256}=\mathbf{H}^{-1}_{34}\begin{bmatrix}C^{(3)}_{4,1256}\oplus A^{(4)}_{4,2356}\oplus B^{(4)}_{4,1356}\\ D^{(3)}_{3,1256}\oplus E^{(3)}_{3,1246}\oplus F^{(3)}_{3,1245}\end{bmatrix},$$

$$\mathbf{x}^2_{34,1256}=\mathbf{H}^{-1}_{34}\begin{bmatrix}C^{(4)}_{4,1256}\oplus E^{(3)}_{4,1236}\oplus F^{(3)}_{4,1235}\\ D^{(4)}_{3,1256}\oplus A^{(4)}_{3,2456}\oplus B^{(4)}_{3,1456}\end{bmatrix},$$

$$\mathbf{x}^1_{35,1246}=\mathbf{H}^{-1}_{35}\begin{bmatrix}C^{(3)}_{5,1246}\oplus A^{(4)}_{5,2346}\oplus B^{(4)}_{5,1346}\\ E^{(4)}_{3,1246}\oplus D^{(5)}_{3,1256}\oplus F^{(4)}_{3,1245}\end{bmatrix},$$

$$\mathbf{x}^2_{35,1246}=\mathbf{H}^{-1}_{35}\begin{bmatrix}C^{(4)}_{5,1246}\oplus D^{(3)}_{5,1236}\oplus F^{(3)}_{5,1234}\\ E^{(5)}_{3,1246}\oplus A^{(5)}_{3,2456}\oplus B^{(5)}_{3,1456}\end{bmatrix},$$

$$\mathbf{x}^1_{36,1245}=\mathbf{H}^{-1}_{36}\begin{bmatrix}C^{(3)}_{6,1245}\oplus A^{(4)}_{6,2345}\oplus B^{(4)}_{6,1345}\\ F^{(5)}_{3,1245}\oplus D^{(6)}_{3,1256}\oplus E^{(6)}_{3,1246}\end{bmatrix},$$

$$\mathbf{x}^2_{36,1245}=\mathbf{H}^{-1}_{36}\begin{bmatrix}C^{(4)}_{6,1245}\oplus D^{(3)}_{6,1235}\oplus E^{(3)}_{6,1234}\\ F^{(6)}_{3,1245}\oplus A^{(6)}_{3,2456}\oplus B^{(6)}_{3,1456}\end{bmatrix},$$

$$\mathbf{x}^1_{45,1236}=\mathbf{H}^{-1}_{45}\begin{bmatrix}D^{(4)}_{5,1236}\oplus A^{(5)}_{5,2346}\oplus B^{(5)}_{5,1346}\\ E^{(4)}_{4,1236}\oplus C^{(5)}_{4,1256}\oplus F^{(4)}_{4,1235}\end{bmatrix},$$

$$\mathbf{x}^2_{45,1236}=\mathbf{H}^{-1}_{45}\begin{bmatrix}D^{(5)}_{5,1236}\oplus C^{(5)}_{5,1246}\oplus F^{(4)}_{5,1234}\\ E^{(5)}_{4,1236}\oplus A^{(5)}_{4,2356}\oplus B^{(5)}_{4,1356}\end{bmatrix},$$

$$\mathbf{x}^1_{46,1235}=\mathbf{H}^{-1}_{46}\begin{bmatrix}D^{(4)}_{6,1235}\oplus A^{(5)}_{6,2345}\oplus B^{(5)}_{6,1345}\\ F^{(5)}_{4,1235}\oplus C^{(6)}_{4,1256}\oplus E^{(6)}_{4,1236}\end{bmatrix},$$

$$\mathbf{x}^2_{46,1235}=\mathbf{H}^{-1}_{46}\begin{bmatrix}D^{(5)}_{6,1235}\oplus C^{(5)}_{6,1245}\oplus E^{(4)}_{6,1234}\\ F^{(6)}_{4,1235}\oplus A^{(6)}_{4,2356}\oplus B^{(6)}_{4,1356}\end{bmatrix},$$

$$\mathbf{x}^1_{56,1234}=\mathbf{H}^{-1}_{56}\begin{bmatrix}E^{(5)}_{6,1234}\oplus A^{(6)}_{6,2345}\oplus B^{(6)}_{6,1345}\\ F^{(5)}_{5,1234}\oplus C^{(6)}_{5,1246}\oplus D^{(6)}_{5,1236}\end{bmatrix},$$

$$\mathbf{x}^2_{56,1234}=\mathbf{H}^{-1}_{56}\begin{bmatrix}E^{(6)}_{6,1234}\oplus C^{(6)}_{6,1245}\oplus D^{(6)}_{6,1235}\\ F^{(6)}_{5,1234}\oplus A^{(6)}_{5,2346}\oplus B^{(6)}_{5,1346}\end{bmatrix}.$$

By examining any of the above transmitted vectors, we can deduce that each transmission serves a total of 6 *users, with a feedback cost of* $C=2$.

***Example* 6.** *Let us consider the* $L=2$ *MISO BC with* $K=5$ *users and normalized cumulative cache of size* $t=2$. *The first* 18 *out of the total* 60 *transmissions are:*

$$x^1_{12,34}=\mathbf{H}^{-1}_{12}\begin{bmatrix}A^{(1)}_{34,2}\oplus C^{(1)}_{14,2}\\ B^{(1)}_{34,1}\oplus D^{(1)}_{23,1}\end{bmatrix}, x^2_{12,34}=\mathbf{H}^{-1}_{12}\begin{bmatrix}A^{(2)}_{34,2}\oplus D^{(1)}_{13,2}\\ B^{(2)}_{34,1}\oplus C^{(1)}_{24,1}\end{bmatrix}$$

$$x^1_{12,35}=\mathbf{H}^{-1}_{12}\begin{bmatrix}A^{(1)}_{35,2}\oplus C^{(1)}_{15,2}\\ B^{(1)}_{35,1}\oplus E^{(1)}_{23,1}\end{bmatrix}, x^2_{12,35}=\mathbf{H}^{-1}_{12}\begin{bmatrix}A^{(2)}_{35,2}\oplus E^{(1)}_{13,2}\\ B^{(2)}_{35,1}\oplus C^{(1)}_{25,1}\end{bmatrix}$$

$$x^1_{12,45}=\mathbf{H}^{-1}_{12}\begin{bmatrix}A^{(1)}_{45,2}\oplus D^{(1)}_{15,2}\\ B^{(1)}_{45,1}\oplus E^{(1)}_{24,1}\end{bmatrix}, x^2_{12,45}=\mathbf{H}^{-1}_{12}\begin{bmatrix}A^{(2)}_{45,2}\oplus E^{(1)}_{14,2}\\ B^{(2)}_{45,1}\oplus D^{(1)}_{25,1}\end{bmatrix}$$

$$x^1_{13,24}=\mathbf{H}^{-1}_{13}\begin{bmatrix}A^{(1)}_{24,3}\oplus B^{(1)}_{14,3}\\ C^{(2)}_{24,1}\oplus D^{(2)}_{23,1}\end{bmatrix}, x^2_{13,24}=\mathbf{H}^{-1}_{13}\begin{bmatrix}A^{(2)}_{24,3}\oplus D^{(1)}_{12,3}\\ C^{(3)}_{24,1}\oplus B^{(3)}_{34,1}\end{bmatrix}$$

$$x^1_{13,25}=\mathbf{H}^{-1}_{13}\begin{bmatrix}A^{(1)}_{25,3}\oplus B^{(1)}_{15,3}\\ C^{(2)}_{25,1}\oplus E^{(2)}_{23,1}\end{bmatrix}, x^2_{13,25}=\mathbf{H}^{-1}_{13}\begin{bmatrix}A^{(2)}_{25,3}\oplus E^{(1)}_{12,3}\\ C^{(3)}_{25,1}\oplus B^{(3)}_{35,1}\end{bmatrix}$$

$$x^1_{13,45}=\mathbf{H}^{-1}_{13}\begin{bmatrix}A^{(1)}_{45,3}\oplus D^{(1)}_{15,3}\\ C^{(2)}_{45,1}\oplus E^{(1)}_{34,1}\end{bmatrix}, x^2_{13,45}=\mathbf{H}^{-1}_{13}\begin{bmatrix}A^{(2)}_{45,3}\oplus E^{(1)}_{14,3}\\ C^{(3)}_{45,1}\oplus D^{(1)}_{35,1}\end{bmatrix}$$

$$x^1_{14,23}=\mathbf{H}^{-1}_{14}\begin{bmatrix}A^{(1)}_{23,4}\oplus B^{(1)}_{13,4}\\ D^{(3)}_{23,1}\oplus C^{(4)}_{24,1}\end{bmatrix}, x^2_{14,23}=\mathbf{H}^{-1}_{14}\begin{bmatrix}A^{(2)}_{23,4}\oplus C^{(1)}_{12,4}\\ D^{(4)}_{23,1}\oplus B^{(4)}_{34,1}\end{bmatrix}$$

$$x^1_{14,25}=\mathbf{H}^{-1}_{14}\begin{bmatrix}A^{(1)}_{25,4}\oplus B^{(1)}_{15,4}\\ D^{(2)}_{25,1}\oplus E^{(2)}_{24,1}\end{bmatrix}, x^2_{14,25}=\mathbf{H}^{-1}_{14}\begin{bmatrix}A^{(2)}_{25,4}\oplus E^{(1)}_{12,4}\\ D^{(3)}_{25,1}\oplus B^{(3)}_{45,1}\end{bmatrix}$$

$$x^1_{14,35}=\mathbf{H}^{-1}_{14}\begin{bmatrix}A^{(1)}_{35,4}\oplus C^{(1)}_{15,4}\\ D^{(2)}_{35,1}\oplus E^{(2)}_{34,1}\end{bmatrix}, x^2_{14,35}=\mathbf{H}^{-1}_{14}\begin{bmatrix}A^{(2)}_{35,4}\oplus E^{(1)}_{13,4}\\ D^{(3)}_{35,1}\oplus C^{(1)}_{45,1}\end{bmatrix}$$

APPENDIX III
FEEDBACK AS A FUNCTION OF THE COHERENCE PERIOD

The model of our work was based on the assumption that each transmission slot spans one or more coherence periods. In this appendix we consider a different scenario where now multiple transmission slots can fit inside one coherence period and we show the generated feedback costs for this setting.

Compared to a non-cache-aided multi-antenna system, the feedback costs of the cache-aided system are dependent on the size of the coherence period. This is because, while in the absence of caching one could fix a set of users and communicate to them repeatedly for the whole duration of the coherence period, thus avoiding the increase of the feedback costs, this is not an option for the cache-aided case. This inability to reuse CSI can be attributed to the requirement of coded caching to introduce new subsets of users after each transmission slot, which leads to what we call the "combine harvester" effect, where the number of users who need to communicate their CSI is rapidly increasing.

In order to explore the feedback costs associated with longer coherence periods we use the following example.

***Example* 7.** *Let us assume a MISO BC system with* $L=5$ *transmit antennas, serving* K *users, where each user is equipped with a cache of normalized size* $\gamma=\frac{1}{10}$. *For this setting, we plot in Figure 1 the fraction of the total communication (in units of file-size) that can be completed as a function of the available feedback for our proposed algorithm. Furthermore, we perform the same analysis for the algorithms in [33], [34].*

In particular, by denoting the feedback cost[12] *with* C *(*C *users communicate CSI, and the transmitter communicates* C *precoders to all the users) we can first observe that our algorithm requires CSI cost of only* $C\geq L=5$ *in order to start communicating with the maximum DoF of* $L+t$, *while*

[12]For simplicity we assume that the coherence block is long enough to exactly fit the transmission of this particular portion that we aim to complete. In other words, the time frame of this comparison here is such that we do not have to worry about users having to renew their CSI because the coherence period has elapsed.

the state-of-the-art algorithms require $C \geq L + K\gamma = 5 + \frac{K}{10}$ *CSI in order to achieve the same DoF.*

Let us recall from [33], [34] that, once feedback is acquired for some set of $C \geq L + K\gamma$ *users, then one can have*

$$(K\gamma + L)\binom{C}{L + K\gamma}\binom{L + K\gamma - 1}{K\gamma} \tag{84}$$

transmission slots[13] *without need for additional feedback.*

For our algorithm, we can similarly calculate the maximum possible transmission slots when having CSI from $C \geq L$ *users to be*

$$L \cdot \binom{C}{L}\binom{K - L}{K\gamma}. \tag{85}$$

Now comparing (84) *with* (85)*, and taking into account that each transmission slot in both cases carries the same amount of information, we can see that our algorithm serves a much larger portion of the delivery compared to [33], [34] for the same feedback cost* C*. Specifically, for some arbitrary cost* $C \geq L + K\gamma$*, the ratio of the two algorithms gives*

$$\frac{L \cdot \binom{C}{L}\binom{K-L}{K\gamma}}{(K\gamma + L)\binom{C}{L+K\gamma}\binom{L+K\gamma-1}{K\gamma}} \overset{(\star)}{\approx} \left(\frac{L + K\gamma}{L}\right)^{L}\left(\frac{K - L}{C}\right)^{K\gamma}$$

where in $(\star)$ *we used the approximation* $\binom{n}{k} \approx \left(\frac{n}{k}\right)^{k}$*. The comparison of the algorithms is illustrated in Figure 1, where we display the CSI cost needed to complete any fraction of the entire delivery.*

For example, we can observe that, in order to successfully communicate a fraction 10^{-5}*, in the case where* $K = 50$ *our algorithm requires feedback cost* $C = 7$*, while the state-of-the-art scheme requires a cost of approximately* 18*. This difference is further amplified when we focus on cases with higher number of users. As an example, in the case where* $K = 100$ *and the same fraction of the whole transmission, the respective costs of the new algorithm compared to the old algorithms are* 12 *and* 52*, while in the case where* $K = 200$ *these costs rise to* 22 *and* 131*, respectively.*

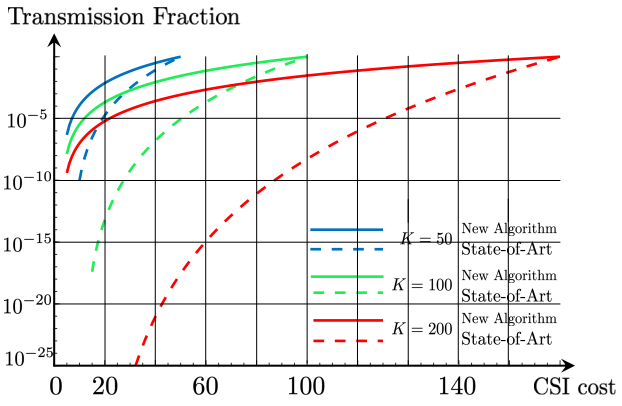

Fig. 1. The total CSI cost that is required to complete a fraction of the delivery phase inside a single coherence period. We compare the costs required by the proposed algorithm with the state-of-the-art algorithms [33], [34], where the later exhibit the same CSI requirements. The cost represents the number of users that need to send feedback. Parameter γ is fixed at value $\gamma = \frac{1}{10}$.

REFERENCES

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[2] E. Lampiris and P. Elia, "Achieving full multiplexing and unbounded caching gains with bounded feedback resources," *2018 IEEE International Symposium of Information Theory (ISIT)*, pp. 1440–1444, 2018.

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

[4] K. Wan, D. Tuninetti, and P. Piantanida, "An index coding approach to caching with uncoded cache placement," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1318–1332, 2020.

[5] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1281–1296, 2018.

[13] We added the first term in (84) in order to equate the subpacketization of the corresponding algorithm with that of our algorithm. Hence, any transmission of either the state-of-the-art algorithms or our algorithm carries the same amount of information.

[6] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, 2002.

[7] A. Lozano, R. W. Heath, and J. G. Andrews, "Fundamental limits of cooperation," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5213–5226, Sep. 2013.

[8] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[9] S. A. Jafar and A. J. Goldsmith, "Isotropic fading vector broadcast channels: The scalar upper bound and loss in degrees of freedom," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 848–857, 2005.

[10] C. Huang, S. A. Jafar, S. Shamai, and S. Vishwanath, "On degrees of freedom region of MIMO networks without channel state information at transmitters," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 849–857, 2012.

[11] A. Lapidoth, S. Shamai, and M. Wigger, "On the capacity of fading MIMO Broadcast Channels with imperfect transmitter side-information," *arXiv preprint cs/0605079*, 2006.

[12] C. S. Vaze and M. K. Varanasi, "The degree-of-freedom regions of MIMO Broadcast, Interference, and cognitive radio channels with no CSIT," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5354–5374, 2012.

[13] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4418–4431, 2012.

[14] A. Bazco-Nogueras, P. de Kerret, D. Gesbert, and N. Gresset, "On the degrees-of-freedom of the k-user distributed broadcast channel," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5642–5659, 2020.

[15] S. A. Jafar, "Blind interference alignment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 3, pp. 216–227, 2012.

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

[17] T. Gou and S. A. Jafar, "Optimal use of current and outdated channel state information: Degrees of freedom of the MISO BC with mixed CSIT," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1084–1087, 2012.

[18] J. Chen and P. Elia, "Degrees-of-freedom region of the MISO broadcast channel with general mixed-CSIT," *arXiv preprint arXiv:1205.3474*, 2012.

[19] ——, "Toward the performance vs. feedback tradeoff for the two-user MISO broadcast channel," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8336–8356, 2013.

[20] R. Tandon, S. A. Jafar, S. Shamai, and H. V. Poor, "On the synergistic benefits of alternating CSIT for the MISO broadcast channel," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4106–4128, 2013.

[21] J. Chen, P. Elia, and S. A. Jafar, "On the two-user MISO Broadcast Channel with alternating CSIT: A topological perspective." *IEEE*

Transactions on Information Theory, vol. 61, no. 8, pp. 4345–4366, 2015.

[22] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341–1365, October 2008.

[23] N. Lee and W. Shin, "Adaptive feedback scheme on K-cell MISO Interfering Broadcast Channel with limited feedback," *IEEE Transactions on Wireless Communications*, vol. 10, no. 2, pp. 401–406, 2011.

[24] J. Park, N. Lee, J. G. Andrews, and R. W. Heath, "On the optimal feedback rate in interference-limited multi-antenna cellular systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5748–5762, Aug 2016.

[25] M. Min, "Bounds on the optimal feedback rate for multi-antenna systems in interference-limited cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4845–4860, 2018.

[26] O. El Ayach, A. Lozano, and R. W. Heath, "On the overhead of Interference Alignment: Training, feedback, and cooperation," *IEEE Transactions on Wireless Communications*, vol. 11, no. 11, pp. 4192–4203, 2012.

[27] N. Mokari, F. Alavi, S. Parsaeefard, and T. Le-Ngoc, "Limited-feedback resource allocation in heterogeneous cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2509–2521, 2016.

[28] E. Sula, M. Gastpar, and G. Kramer, "Sum-rate capacity for symmetric gaussian multiple access channels with feedback," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2860–2871, 2020.

[29] A. Vahid, C. Suh, and A. S. Avestimehr, "Interference channels with rate-limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2788–2812, 2012.

[30] A. G. Davoodi and S. A. Jafar, "Gdof of the MISO BC: Bridging the gap between finite precision CSIT and perfect CSIT," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 1297–1301.

[31] G. Bassi, P. Piantanida, and S. Shamai Shitz, "The wiretap channel with generalized feedback: Secure communication and key generation," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2213–2233, 2019.

[32] L. Song, F. Alajaji, and T. Linder, "Capacity of burst noise-erasure channels with and without feedback and input cost," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 276–291, 2019.

[33] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server Coded Caching," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7253–7271, 2016.

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

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

[36] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2091–2106, 2020.

[37] J. S. Pujol Roig, D. Gündüz, and F. Tosato, "Interference networks with caches at both ends," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[38] E. Piovano, H. Joudeh, and B. Clerckx, "On Coded Caching in the overloaded MISO Broadcast Channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2795–2799.

[39] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1176–1188, 2018.

[40] S. P. Shariatpanahi and B. H. Khalaj, "On multi-server Coded Caching in the low memory regime," *arXiv preprint arXiv:1803.07655*, 2018.

[41] E. Lampiris, P. Elia, and G. Caire, "Bridging the gap between multiplexing and diversity in finite SNR multiple antenna coded caching," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1272–1277.

[42] E. Lampiris and P. Elia, "Bridging two extremes: Multi-antenna coded caching with reduced subpacketization and CSIT," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.

[43] M. Kobayashi and G. Caire, "On the net DoF comparison between ZF and MAT over time-varying MISO broadcast channels," in *IEEE International Symposium on Information Theory (ISIT)*, 2012, pp. 2286–2290.

[44] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015, pp. 1099–1105.

[45] A. Bazco-Nogueras and P. Elia, "Rate-memory trade-off for the cache-aided MISO Broadcast Channel with hybrid CSIT," in *IEEE Information Theory Workshop (ITW)*, 2021.

[46] M. A. T. Nejad, S. P. Shariatpanahi, and B. H. Khalaj, "On storage allocation in cache-enabled interference channels with mixed CSIT," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2017, pp. 1177–1182.

[47] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of Coded-Caching and CSIT feedback," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3142–3160, 2017.

[48] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized Degrees of Freedom of the symmetric cache-aided MISO Broadcast Channel with partial CSIT," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5799–5815, Sep. 2019.

[49] E. Lampiris, J. Zhang, and P. Elia, "Cache-aided cooperation with no CSIT," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2960–2964.

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

[51] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, "Fundamental limits of wireless caching under uneven-capacity channels," in *International Zurich Seminar*, February 2020.

[52] H. Joudeh, E. Lampiris, P. Elia, and G. Caire, "Fundamental limits of wireless caching under mixed cacheable and uncacheable traffic," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1693-1698.

[53] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 809–813.

[54] F. Xu, K. Liu, and M. Tao, "Cooperative Tx/Rx caching in interference channels: A storage-latency tradeoff study," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 2034–2038.

[55] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5359–5380, July 2018.

[56] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Annual Conference on Information Science and Systems (CISS)*, 2016.

[57] J. Zhang and O. Simeone, "Fundamental limits of cloud and cache-aided interference management with multi-antenna edge nodes," *IEEE Transactions on Information Theory*, vol. 65, no. 8, pp. 5197–5214, 2019.

[58] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5061–5076, 2017.

[59] Y. Cao and M. Tao, "Treating content delivery in multi-antenna coded caching as general message sets transmission: A DoF region perspective," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3129–3141, 2019.

[60] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7635–7651, 2020.

[61] I. Bergel and S. Mohajer, "Practical scheme for miso cache-aided communication," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[62] M. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tölli, "Low-complexity high-performance cyclic caching for large MISO systems," 2020.

[63] E. Piovano, H. Joudeh, and B. Clerckx, "Centralized and decentralized cache-aided interference management in heterogeneous parallel channels," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1881–1896, 2020.

[64] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Cache-Aided Interference Management in Wireless Cellular Networks," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3376–3387, 2019.

[65] M. Razaviyayn, G. Lyubeznik, and Z.-Q. Luo, "On the degrees of freedom achievable through interference alignment in a MIMO interference channel," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 812–821, Feb. 2012.

Eleftherios Lampiris obtained the B.Sc. and M.Sc. degrees, respectively, in Physics and Radio-Electrology from the University of Athens, Greece and the PhD in Electrical Engineering from Sorbonne University, France while working at EURECOM. He has worked as a Post-Doctoral Researcher at the Technical University of Berlin, Germany and EURECOM, Sophia Antipolis, France. His latest research interests include practical aspects of Coded Caching, Game Theory, Network Optimization and Computer Vision.

Antonio Bazco-Nogueras (M'20) received the B.S. and M.S. degrees in Telecommunications Engineering, both from University of Zaragoza, Spain, in 2014 and 2016, respectively. He obtained the Ph.D. degree from Sorbonne Université, Paris, France, in collaboration with the Mitsubishi Electric R&D Centre Europe, Rennes, France, in 2019. He was a post-doctoral researcher at EURECOM, Sophia-Antipolis, France, from 2020 to 2021. He is currently a post-doctoral researcher at IMDEA Networks Institute, Madrid, Spain. His research interests include multi-user information theory, intelligent networks, decentralized systems, content delivery networks, and cooperative wireless networks.

Petros Elia received the B.Sc. degree from the Illinois Institute of Technology, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 2001 and 2006 respectively. He is now a professor with the Department of Communication Systems at EURECOM in Sophia Antipolis, France. His latest research deals with distributed computing as well as with the intersection of caching and communications in multiuser settings. He has also worked in the area of complexity-constrained communications, MIMO, queueing theory and cross-layer design, coding theory, information theoretic limits in cooperative communications, and surveillance networks. He is a Fulbright scholar, the co-recipient of the NEWCOM++ distinguished achievement award 2008-2011 for a sequence of publications on the topic of complexity in wireless communications, and the recipient of the ERC Consolidator Grant 2017-2022 on cache-aided wireless communications.