Resolving a Feedback Bottleneck of Multi-Antenna Coded Caching

Eleftherios Lampiris
EURECOM
06410 Sophia Antipolis, France
Email: lampiris@eurecom.fr

Petros Elia
EURECOM
06410 Sophia Antipolis, France
Email: elia@eurecom.fr

***Abstract*— The work here makes substantial progress towards resolving the well known feedback bottleneck of multi-antenna coded caching, by introducing a fundamentally new algorithm that completely untangles caching gains from feedback costs. In the context of the K-user MISO broadcast channel with cache-aided receivers and an L-antenna transmitter, the algorithm achieves full multiplexing gains and unbounded near-optimal caching gains at the mere CSIT/CSIR costs associated to achieving multiplexing gains only. While prior state-of-art required feedback costs that scaled with K, the new algorithm requires only L uplink/downlink feedback acquisition slots, irrespective of how large the achieved caching gain is, and thus irrespective of how many users are served at a time. In the end, the result solidifies the role of coded caching as a method for reducing feedback requirements in multi-antenna environments.**

I. Introduction

In the context of communicating multimedia content, the recent breakthrough of *coded caching* [1] revealed that caching modest amounts of such content at the receivers, can yield unprecedented reductions in the content delivery delay.

This work in [1] considered a shared-link broadcast channel where a transmitter wishes to serve content from a library of N files, to K receivers, each endowed with a cache of size equal to the size of M files. In this context — where each user can store a fraction $\gamma \triangleq \frac{M}{N}$ of the library, and where each user can ask simultaneously for their own file from this library — the work in [1] employed a novel cache-placement technique and a novel multicasting transmission scheme, which jointly exploited the fact that, for each user, the undesired cached content can be used as side information to remove interference stemming from other users' requested files. As a result, in the original shared-link (noiseless, wired) setting where the link has capacity 1 file per unit of time, this allowed for a worst-case (normalized) delivery time of

$$\mathcal{T} = \frac{K(1-\gamma)}{1+K\gamma} < \frac{1}{\gamma} \tag{1}$$

which implied an ability to serve $1 + K\gamma$ users at a time, corresponding to an additive *caching gain* of $G = K\gamma$, i.e., a gain of being able to serve an additional $K\gamma$ users at a time, as a result of caching. This caching gain was shown in [3] (see also [4]) to be optimal under the basic assumption of uncoded cache placement.

The direct extension of this result to the equivalent high-SNR single-antenna *wireless* broadcast channel (BC) — where again the long-term capacity of each point-to-point link is normalized to 1 file per unit of time — implied a degrees-of-freedom (DoF) performance of

$$d_\Sigma(\gamma) \triangleq \frac{K(1-\gamma)}{\mathcal{T}} = 1 + K\gamma. \tag{2}$$

This ability of single-antenna coded caching to serve a scaling number of users at a time, without any channel state information at the transmitter (CSIT), offered an alternative to the high feedback-costs required by multi-antenna systems. As is known (cf. [5], [6]), such feedback costs are the reason most multi-antenna solutions fail to scale[1].

At the same time, there was substantial interest in combining the gains from caching with the traditional multiplexing gains of feedback-aided multi-antenna systems. This was an interesting direction that sought 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 [26] showed that — in a wired multi-server (L servers) setting which can easily be seen to correspond to the high-SNR cache-aided MISO BC setting with L transmit antennas — the two gains (multiplexing and caching gains) could be combined additively, yielding a sum-DoF equal to

$$d_\Sigma(\gamma) = L + K\gamma \tag{3}$$

which was later shown to have a gap of at most 2 from the one-shot linear optimal sum-DoF [27]. Since then, many works such as [27]–[33] have developed different coded caching schemes for the multi transmit-antenna setting.

This work was supported by the ANR project ECOLOGICAL-BITS-AND-FLOPS. Part of this work was presented in ISIT 2018 [2].

[1]For a detailed view on the feedback requirements of the MISO (multiple input single output) BC the reader is directed in the work of [7] for a degrees of freedom characterization under perfect CSIT, in [8]–[11] for a no CSIT analysis, in [12]–[14] for a treatment of the compound CSIT scenario and in [15] for the exploitation of delayed CSIT knowledge. Further, [16] considers only the patterns of the channel coherence period to be known, the works in [17]–[20] allowed mixed CSI at the transmitters to be known, while the work in [21] considered alternating CSIT.

A. Scaling Feedback Costs in Multi-Antenna Coded Caching

While though in the single antenna case [1] one could achieve the near optimal (and under some assumptions, optimal [3], [4]) caching gain $K\gamma$ without requiring any channel state information at the transmitter, a significant feedback problem arose in the presence of multiple antennas. Specifically, all known multi-antenna coded caching methods [26]–[30] that achieve the full DoF $L+K\gamma$, incur scaling feedback costs as they require each of the $L+K\gamma$ benefiting receivers to send feedback to the transmitter. A similar *global CSIR* cost appears with respect to the required CSI at the receivers (CSIR) which was required to include information on the CSIT-based precoders of all the $L+K\gamma$ benefiting users.

The following example aims to demonstrate the aforementioned CSI costs, and it focuses on a simple instance of the original multiserver method in [26] which serves as a proxy to other methods with similar feedback requirements.

***Example* 1.** *Let us consider the cache-aided MISO BC setting with $K=4$ users, normalized cache size $\gamma=1/2$ and $L=2$ transmit antennas, where, using the multiserver approach, one can treat $L+K\gamma=4$ users at a time. Assuming that users $1,2,3,4$ respectively request files A,B,C,D, then each of the three transmissions takes the form*

$$\begin{aligned}\mathbf{x} =&\mathbf{h}_4^{\perp}(A_{23}\oplus B_{13}\oplus C_{12})+\mathbf{h}_3^{\perp}(A_{24}\oplus B_{14}\oplus D_{12})+ \\ &+\mathbf{h}_2^{\perp}(A_{34}\oplus C_{14}\oplus D_{13})+\mathbf{h}_1^{\perp}(B_{34}\oplus C_{24}\oplus D_{23})\end{aligned} \quad (4)$$

where $\mathbf{h}_k^{\perp}$ denotes the precoder orthogonal to the channel of user k, and where A_{ij} (respectively B_{ij},C_{ij},D_{ij}) denotes the part of file A (respectively of B,C,D) that is cached at users i and j. We clearly see that the transmitter must know all users' channel vectors (in order to form the four precoders), and at the same time — in order to be able to decode the desired subfile — each receiver must know the composite channel-precoder product for each precoder (e.g. receiver 1 must know $\mathbf{h}_1^T\mathbf{h}_1^{\perp}$ as well as $\mathbf{h}_1^T\mathbf{h}_2^{\perp}$, $\mathbf{h}_1^T\mathbf{h}_3^{\perp}$ and $\mathbf{h}_1^T\mathbf{h}_4^{\perp}$). This implies $L+K\gamma=4$ uplink training slots for CSIT acquisition, and $L+K\gamma=4$ downlink training slots for global CSIR acquisition[2].

In the context of frequency division duplexing (FDD), this feedback cost of existing methods, implies a CSIT cost of $L+K\gamma$ feedback vectors, while in the more interesting setting of Time Division Duplexing (TDD), this requires $L+K\gamma$ uplink training time slots for CSIT acquisition, and an extra cost of $L+K\gamma$ downlink training time slots for global CSIR acquisition. As we know, such scaling feedback costs can consume a significant portion of the coherence time, thus resulting in diminishing DoF gains, as this was shown in [5].

Motivated by this feedback bottleneck, different works on multi-antenna (multi-transmitter) coded caching have sought to reduce CSI costs, but in all known cases, any subsequent CSI reductions come at the direct cost of substantially reduced DoF. For example, the works in [31], [34] consider reduced quality CSIT, but yield a maximum DoF that is bounded close to $K\gamma+1$, while the works in [35], [36] consider only statistical CSI, but again achieve much lower DoF. Similarly, the work in [37] uses ACK/NACK type CSIT to ameliorate the issue of unequal channel strengths, but again, achieves no multiplexing gains.

In this work here, we will introduce a fundamentally new algorithm that achieves the desired sum-DoF $d_\Sigma=L+K\gamma$, with a feedback cost associated to having only L users. For example, in TDD, the developed scheme requires only L uplink and L downlink training slots, irrespective of how large the achieved caching gain is, and thus irrespective of how many users are served at a time. This reveals that coded caching in multi-antenna settings can use CSI to first provide the maximal multiplexing gain, and then use the caches to provide an additional full caching gain without any additional CSI costs.

[2]The process of feedback acquisition will be described in Algorithm 1 found in Section IV, where we will recall that global CSIR acquisition can be performed by broadcasting a set of training symbols to all users.

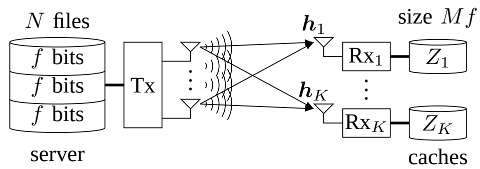

Fig. 1. An L-antenna transmitter having access to a library of N files and communicating with K receivers, each with cache size equal to $M=\gamma N$ files.

II. System Model

We assume K single-antenna receiving users served by an L-antenna transmitter. The received signal at user $k\in\{1,2,\dots K\}\triangleq[K]$ takes the form

$$y_k=\mathbf{h}_k^T\mathbf{x}+w_k,\quad \forall k\in[K] \quad (5)$$

where $\mathbf{x}\in\mathbb{C}^{L\times 1}$ denotes the transmitted vector from the L-antenna array satisfying the power constraint $\mathbb{E}\left\{\|\mathbf{x}\|^2\right\}\le P$, where $\mathbf{h}_k^T\in\mathbb{C}^{1\times L}$ denotes the random-fading channel vector of user k, and where $w_k\sim\mathcal{N}(0,1)$ is the AWGN noise experienced at user k. The work focuses on the DoF performance, thus the signal to noise ratio is considered to be large. We also assume that the fading process is statistically symmetric across users.

Communication happens in two main phases. First, in the *cache-placement phase*, the caches are filled with content from a library of N files $\{W^{(n)},n\in[N]\}$. Then in the *feedback-acquisition and content delivery phase*, each user simultaneously requests a single file from the library, and then — while acquiring and disseminating the required CSI — the base station serves the requested files by considering the requests and the cached content. The aim is to reduce the worst case (over all possible demands) delivery time $\mathcal{T}$. We will first assume that $K\gamma$ is an integer multiple of L, while for the other cases, we will use standard memory sharing[3]. Further, we assume that the system operates in TDD mode[4], and for simplicity the precoder method of choice will be Zero-Forcing (ZF)[5].

a) Notation: For some set $\lambda \subset [K]$ of $|\lambda| = L$ users[6], we will denote with H_λ^{-1} the normalized inverse of the $L \times L$ channel matrix H_λ corresponding to the channel from the transmitter to the L users in set λ, while the kth column of H_λ^{-1} will be denoted by $\mathbf{h}_{\lambda\setminus\{k\}}^{\perp}$, where

$$\mathbf{h}_p^T \cdot \mathbf{h}_{\lambda\setminus\{k\}}^{\perp} = \begin{cases} 1, & \text{if } p = k \\ 0, & \text{if } p \in \lambda \setminus \{k\} \\ \neq 0, & \text{if } p \notin \lambda. \end{cases} \tag{6}$$

We will use Z_k to denote the cache content of user $k \in [K]$, and $d_k \in [N]$ to denote the index of the file[7] requested by user k. Finally, $\oplus$ denotes the bitwise-XOR operator and $\binom{n}{k}$ denotes the n-choose-k operator ($n \geq k$, n and $k \in \{1, 2, ...\}$).

III. Main result and an example

We proceed with the main result, which is based on the algorithm that we will describe in the next section.

Theorem 1. *In the K-user cache-aided MISO-BC with L antennas and normalized cache size γ, the DoF $d_\Sigma = K\gamma + L$ can be achieved with CSIT from only L users at a time, and thus with CSIT cost of L uplink training time-slots, and global CSIR cost of L downlink training time-slots.*

Proof. The proof is constructive and is found in Section IV-C which describes the scheme, while Section IV-B describes the CSI acquisition phase. □

a) Intuition and an example: Before fully describing the scheme, we proceed with some intuition on the design.

We first note that for the cache placement, the partition of files into subfiles and the storing of subfiles in the users' caches, will draw directly from [1].

On the other hand, the XOR generation method will be fundamentally different. The first step is to construct XORs composed of $\frac{K\gamma}{L} + 1$ subfiles, and to then have each transmission communicate L such XORs, thus allowing each transmission to communicate $L + K\gamma$ different subfiles aimed at simultaneously serving a set of $L + K\gamma$ users. Each such set of $L + K\gamma$ served ("active") users will be divided into two sets; the first set λ will consist of the L users that will be assisted by precoding, while the second set π will have $K\gamma$ users who will not be assisted by precoding and who must thus compensate with their caches. Finally, the vector of XORs will be multiplied by the inverse H_λ^{-1} of the channel matrix corresponding to the users in λ.

What we will see is that the design will guarantee that, during the decoding process, each of the users in λ will only receive one of the XORs (the rest will be nulled-out by the precoder), while the remaining $K\gamma$ users (i.e., those in π) will receive a linear combination of all L XORs. Hence this will mean that the users in λ will have to each cache out $\frac{K\gamma}{L}$ subfiles[8] in order to decode their desired subfile, while the users in π will have to cache out $K\gamma + L - 1$ subfiles i.e., all but one subfiles.

Next, we will demonstrate a single transmission of our algorithm using the setting of Example 1. The goal is to achieve the same performance as before (all 4 users being able to decode their subfiles in each transmission), while using CSIT from only two users at a time, thus requiring no more than $L = 2$ training slots in the uplink and $L = 2$ training slots in the downlink (this example in its entirety can be found in Example 4 in Section IV-D).

Example 2. *In the same MISO BC setting of Example 1, with $L = 2$ transmit antennas, $K = 4$ users, and a fractional cache size of $\gamma = 1/2$, a transmitted vector in the proposed algorithm, will take the form*[9]

$$\mathbf{x} = \mathbf{h}_2^{\perp}(A_{34} \oplus C_{14}) + \mathbf{h}_1^{\perp}(B_{34} \oplus D_{23}) \tag{7}$$

where, as before, files A, B, C and D are requested by users $1, 2, 3$ and 4, respectively, and where A_{ij} represents the part of A that can be found in the caches of users i and j (similarly for B_{ij}, C_{ij} and D_{ij}).

Assuming that user k receives y_k, $k \in \{1, 2, 3, 4\}$, then the message at each user takes the form:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}^T = \begin{bmatrix} \mathbf{h}_1^T(\mathbf{h}_2^{\perp} A_{34} \oplus C_{14} + \mathbf{h}_1^{\perp} B_{34} \oplus D_{23}) \\ \mathbf{h}_2^T(\mathbf{h}_2^{\perp} A_{34} \oplus C_{14} + \mathbf{h}_1^{\perp} B_{34} \oplus D_{23}) \\ \mathbf{h}_3^T(\mathbf{h}_2^{\perp} A_{34} \oplus C_{14} + \mathbf{h}_1^{\perp} B_{34} \oplus D_{23}) \\ \mathbf{h}_4^T(\mathbf{h}_2^{\perp} A_{34} \oplus C_{14} + \mathbf{h}_1^{\perp} B_{34} \oplus D_{23}) \end{bmatrix} \tag{8}$$

$$= \begin{bmatrix} A_{34} \oplus C_{14} \\ B_{34} \oplus D_{23} \\ \mathbf{h}_3^T(\mathbf{h}_2^{\perp} A_{34} \oplus C_{14} + \mathbf{h}_1^{\perp} B_{34} \oplus D_{23}) \\ \mathbf{h}_4^T(\mathbf{h}_2^{\perp} A_{34} \oplus C_{14} + \mathbf{h}_1^{\perp} B_{34} \oplus D_{23}) \end{bmatrix} \tag{9}$$

where we have ignored noise for simplicity.

Hence we see that user 1 and user 2 only receive the first and second XOR respectively (due to the design of the precoders), which means that each of these two users can decode their desired subfiles, i.e. A_{34} and B_{34} respectively,

[3]Thus if for example $L > K\gamma$, we would apply memory sharing, splitting each file in two parts so that the first part is cached with redundancy $K\gamma'' = L$ while the other part would not be cached ($K\gamma' = 0$). Note that, as proved in [32], this approach has a very small impact on the DoF performance, corresponding to a multiplicative loss of at most 2.

[4]The same principles can be applied in FDD mode.

[5]Naturally there are better precoders, but ZF keeps the notation and impact clean, and it suffices for the DoF exposition we seek.

[6]We will assume that sets are ordered.

[7]In the examples we will use the standard simplified notation where $W^{(d_1)} = A, W^{(d_2)} = B$, and so on.

[8]Here we need to point out that the number of subfiles that each user in λ needs to have cached in order to decode its desired subfile is much smaller than in the original scheme of [1], a fact that has been exploited in [38] to show how users without caches can have the full cache-aided DoF in a multiple-antenna environment.

[9]Here the reader is warned that there is a 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.

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 must cache-out subfiles A_{34}, B_{34} and D_{23} (which, by design of the placement are already cached at user 3) in order to decode the desired C_{14}, while user 4 must cache-out subfiles A_{34}, B_{34} and C_{14} (which, by design of the placement, are already cached at user 4) to decode the desired subfile D_{23}. In order to achieve this, users 3 and 4 need to employ their cached content and, also, need some CSI knowledge; user 3 needs products $\mathbf{h}_3^T\mathbf{h}_2^{\perp}$ and $\mathbf{h}_3^T\mathbf{h}_1^{\perp}$, while user 4 needs $\mathbf{h}_4^T\mathbf{h}_2^{\perp}$ and $\mathbf{h}_4^T\mathbf{h}_1^{\perp}$.

We will now describe the two parts of the feedback acquisition process; the uplink part which informs the transmitter of the channels of users 1 and 2, and the downlink part which feeds-back the required composite global CSIR to users 3 and 4.

In the uplink part, users 1 and 2 transmit training symbols so that the transmitter can estimate those channels. This amounts to 2 training time slots in total for the CSIT.

On the other hand, in the downlink training part, users 3 and 4 must each acquire global (composite) CSIR[10]*. As is common, the training sequences for each of the two precoders takes the form*

$$\begin{bmatrix} h_1^{\perp}(1)S(1,1) \\ \vdots \\ h_1^{\perp}(L)S(L,1) \end{bmatrix}, \quad \cdots, \quad \begin{bmatrix} h_1^{\perp}(1)S(1,P) \\ \vdots \\ h_1^{\perp}(L)S(L,P) \end{bmatrix}$$

$$\begin{bmatrix} h_2^{\perp}(1)S(1,1) \\ \vdots \\ h_2^{\perp}(L)S(L,1) \end{bmatrix}, \quad \cdots, \quad \begin{bmatrix} h_2^{\perp}(1)S(1,P) \\ \vdots \\ h_2^{\perp}(L)S(L,P) \end{bmatrix}.$$

employing the standard pilot matrix $\mathbf{S}_{P\times L}$, where P corresponds to the number of training vectors that will be used for channel estimation. Now, during channel uses 1 to P, user 3 will receive

$$y_3(t=j) = \sum_{i=1}^{L} h_1^{\perp}(i)h_3(i)S(i,j) + w_3(t), \quad j \in [P]$$

and then during channel uses $P+1$ to $2P$, it will receive

$$y_3(t=P+j) = \sum_{i=1}^{L} h_2^{\perp}(i)h_3(i)S(i,j) + w_3(t), \quad j \in [P]$$

from which it can successfully estimate $\mathbf{h}_3^T\mathbf{h}_1^{\perp}$ and $\mathbf{h}_3^T\mathbf{h}_2^{\perp}$.

Using the same training symbols, user 4 will receive (at the same time as user 3)

$$y_4(t=j) = \sum_{i=1}^{L} h_4(i)h_1^{\perp}(i)S(i,j) + w_4(t), \quad j \in [P]$$

$$y_4(t=P+j) = \sum_{i=1}^{L} h_4(i)h_2^{\perp}(i)S(i,j) + w_4(t), \quad j \in [P]$$

which can be used to estimate $\mathbf{h}_4^T\mathbf{h}_1^{\perp}$ and $\mathbf{h}_4^T\mathbf{h}_2^{\perp}$.

Thus, the above downlink sequence amounts to 2 training time-slots, and it provides the active users with all the CSI they need to decode successfully their subfiles.

IV. Scheme Description

We proceed to present the scheme's cache-placement phase, and the feedback-acquisition and content-delivery phase.

A. Placement Phase

The placement phase happens without knowledge of the number of transmit antennas, it does not assume any CSI knowledge, and it follows the original scheme in [1] where each file $W^{(n)}, n \in [N]$, is initially split into $\binom{K}{K\gamma}$ subfiles $W_{\tau}^{(n)}$, each indexed by a $K\gamma$-length set $\tau \subset [K]$, in which case the cache of user $k \in [K]$ takes the form

$$Z_k = \left\{ W_{\tau}^{(n)} : \forall \tau \ni k, |\tau| = K\gamma, \forall n \in [N] \right\}. \tag{10}$$

B. CSI Acquisition

This part takes place at the beginning of the coherence period and it involves first an uplink training phase and then a downlink training phase. In the uplink, a user set λ is selected, comprized of L users, who will transmit pilot signals so that the transmitter can estimate their channel coefficients $\mathbf{h}_k$, $\forall k \in \lambda$ and thus construct precoders $\mathbf{h}_{\lambda\setminus\{k\}}^{\perp}$, $\forall k \in \lambda$. Then, during the downlink training phase, the transmitter will broadcast, for each of the L precoders, P vectors[11] as follows

$$\text{diag}\left(\mathbf{h}_{\lambda\setminus\{k\}}^{\perp}\right)\mathbf{S}_{L\times P}, \quad \forall k \in \lambda$$

which, when multiplied by a user's individual channel, will allow for estimation at some user $q \in \pi \subset [K]\setminus\lambda$ (recall $|\pi| = K\gamma$), the needed global-CSIR products $\mathbf{h}_q^T \cdot \mathbf{h}_{\lambda\setminus\{k\}}^{\perp}$, $\forall k \in \lambda$ and where the operation diag($\mathbf{h}$) creates a square diagonal matrix whose elements are the entries of vector $\mathbf{h}$.

In summary: there are L slots for CSIT because only the L users in λ need to send CSIT, and there are L slots for global CSIR because, for each fixed precoder, one (training symbol) shot suffices to communicate the composite channel-precoder product to any number of users. The above process in the form of a pseudo-algorithm can be found in Algorithm 1.

C. Content Delivery

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[12,13]

$$W_{\tau}^{(d_k)} \to \{W_{\sigma,\tau}^{(d_k)},\ \sigma \subseteq [K] \setminus (\tau \cup \{k\}), |\sigma| = L-1\} \tag{11}$$

$$W_{\sigma,\tau}^{(d_k)} \to \{W_{\sigma,\tau}^{\phi,(d_k)},\ \phi \in [L+K\gamma]\}. \tag{12}$$

[10] Here, we remind to the reader that due to the ZF properties the products $\mathbf{h}_1^T\mathbf{h}_2^{\perp} = 1$ and $\mathbf{h}_1^T\mathbf{h}_1^{\perp} = 0$, $\mathbf{h}_2^T\mathbf{h}_1^{\perp} = 1$ and $\mathbf{h}_2^T\mathbf{h}_2^{\perp} = 0$.

[11] P is simply the minimum number of vectors that are required for a perfect estimation of the intended CSI.

[12] In a small abuse of notation we will refer to the segments of the original subfiles as subfiles, from this point onwards.

[13] We note that, for clarity of exposition and to avoid many indices, the index ϕ of Equation 12 will henceforth be suppressed, thus any $W_{\sigma,\tau}^{\phi,(d_k)}$ will be denoted as $W_{\sigma,\tau}^{(d_k)}$ unless ϕ is explicitly needed.

Algorithm 1: Training Phase

```
Uplink Training
Transmitter selects set λ ⊂ [K] of L users.
Users in λ sequentially transmit from set s_up of
  predetermined pilot symbols.
Transmitter receives pilots and forms h_k, ∀k ∈ λ.
Transmitter constructs precoders h_k^⊥, ∀k ∈ λ.
Downlink Training
for k ∈ λ (Select Precoder) do
    Select S_{P×L} (Set of P downlink pilot vectors) for
     p ∈ P do
        Transmit:
```

$$\mathbf{x}_k(p) = \begin{bmatrix} \mathbf{h}_k^{\perp}(1)\mathbf{S}(1,p) \\ \vdots \\ \mathbf{h}_k^{\perp}(L)\mathbf{S}(L,p) \end{bmatrix}.$$

In the following we describe how, for every transmission, the transmitter will first create a vector of L XORs, and will then precode these with the appropriate set of precoders.

1) Individual XOR design: For any two disjoint sets

$$\mu \subset [K], \nu \subset [K], \ \mu \cap \nu = \emptyset$$

where $|\mu| = \frac{K\gamma}{L} + 1, \ \ |\nu| = K\gamma\frac{L-1}{L}$, and for an arbitrary $\sigma \subseteq \Big([K] \setminus (\mu \cup \nu)\Big), \ |\sigma| = L - 1$, we construct the XOR

$$X_{\mu}^{\nu,\sigma} = \bigoplus_{k\in\mu} W_{\sigma,(\nu\cup\mu)\setminus\{k\}}^{(d_k)} \tag{13}$$

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

- each subfile is requested by one user that belongs to user set μ, and where
- all subfiles are known by all users in the user set 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 in their cache, while set σ is a selected subset of $L - 1$ users from the set λ.

***Example* 3.** *Let us consider the $L = 2$ MISO BC with $K\gamma = 4$. Let $\mu = \{1, 2, 3\}, \nu = \{4, 5\}$ and consider some arbitrary $\sigma \in [K] \setminus \{1, 2, 3, 4, 5\}$, $|\sigma| = 1$. Then, the designed XOR takes the form*

$$X_{\{123\}}^{45,\sigma} = W_{\{\sigma,\underbrace{2345}_{\tau}\}}^{(d_1)} \oplus W_{\{\sigma,1345\}}^{(d_2)} \oplus W_{\{\sigma,1245\}}^{(d_3)} \tag{14}$$

and it delivers the subfiles requested by the users in set μ, while each element of the XOR is cached at all users of set ν. Users $1, 2$ and 3 work in the traditional way to cache out each others' files in order to get their own (e.g. user 1 caches out $W_{\{\sigma,1345\}}^{(d_2)} \oplus W_{\{\sigma,1245\}}^{(d_3)}$ to get its own $W_{\{\sigma,2345\}}^{(d_1)}$), while users 4 and 5 are fully protected (since they have cached all 3 subfiles) against this entire undesired XOR. As a quick verification, we see that each index τ (which indicates the set of users that have cached the specific subfile) has size $|\tau| = K\gamma = 4$ which adheres to the available cache-size constraint, which tells us that each file can be stored with redundancy $K\gamma = 4$.

Algorithm 2: Delivery Phase

```
1 for λ ⊂ [K], |λ| = L (precode users in λ) do
2     Create H_λ^{-1}
3     for π ⊂ ([K] \ λ), |π| = Kγ do
4         Break π into some F_i i ∈ [L] : |F_i| = Kγ/L,
            ⋃_{i∈[L]} F_i = π, F_i ∩ F_j = ∅, ∀i, j ∈ [L]
5         for s ∈ {0, 1, ..., L − 1} do
6             r_i = ((s + i − 1) mod L) + 1, i ∈ [L]
7             Transmit
```

$$\mathbf{x}_{\lambda,\pi}^{s} = H_{\lambda}^{-1} \cdot \begin{bmatrix} X_{\lambda(1)\cup\mathcal{F}_{r_1}}^{\pi\setminus\mathcal{F}_{r_1},\lambda\setminus\lambda(1)} \\ X_{\lambda(2)\cup\mathcal{F}_{r_2}}^{\pi\setminus\mathcal{F}_{r_2},\lambda\setminus\lambda(2)} \\ \vdots \\ X_{\lambda(L)\cup\mathcal{F}_{r_L}}^{\pi\setminus\mathcal{F}_{r_L},\lambda\setminus\lambda(L)} \end{bmatrix}. \tag{15}$$

2) Design of vector of XORs: At this point we describe how the L XORs of a transmitted vector are chosen. As explained above, the aim of every transmission is to serve different subfiles to $L + K\gamma$ users (there is no data repetition), while requiring CSIT from only L users. This is described by the following sequence of steps in Algorithm 2. Up to now, we have seen that

- In Step 1, a set λ of L users is chosen.
- In Step 2, a (ZF-type) precoder H_{λ}^{-1} is designed to separate the L users in λ.
- In Step 3, another set $\pi \subseteq [K] \setminus \lambda$ of $K\gamma$ 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, the set π of $K\gamma$ users is partitioned into L non-overlapping sets $\mathcal{F}_i, \ i \in [L]$, each having $\frac{K\gamma}{L}$ users.
- In Step 5 all users from λ are associated to a distinct set $\mathcal{F}_i$, as a function of parameter s that takes values from $\{0, 1, ..., L - 1\}$. For example, when $s = 0$, the first XOR of the vector will be intended for users in set $\lambda(1) \cup \mathcal{F}_1$ (while completely known by all users in $\pi \setminus \mathcal{F}_1$), the second XOR will be intended for the users in the set $\lambda(2) \cup \mathcal{F}_2$ (while completely known by all users in $\pi \setminus \mathcal{F}_2$) and so on. On the other hand, when $s = 1$ the first XOR will be intended for users in $\lambda(1) \cup \mathcal{F}_2$ (while completely known by all users in $\pi \setminus \mathcal{F}_2$), the second XOR will be for users in $\lambda(2) \cup \mathcal{F}_3$ (while completely known by all users in $\pi \setminus \mathcal{F}_3$) and so on. In particular, step 5 (and the operation in step 6, as shown in Algorithm 2), allows us

to iterate over all sets $\mathcal{F}_i$, associating every time a distinct set $\mathcal{F}_i$ to a distinct user from group λ, until all users from set λ have been associated with all sets $\mathcal{F}_i$.
- Then in the last step (Step 7), the vector of the L XORs is transmitted after being precoded by matrix H_λ^{-1}.

3) Decoding at the users: By design of the XORs (cf. (13)), the constructed vector guarantees (together with the precoder) that the users in λ can decode the single XOR (due to ZF) that they receive, and from there (due to caching) proceed to decode their own file, while also guaranteeing that each user in π has cached all subfiles that are found in the entire vector, apart from its desired subfile. Further, the training phase of Section IV-B has provided the users of set π with all the necessary CSI estimates (specifically, it has provided the receivers with all the necessary precoder-channel composite scalars) to perform 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 decoding process at some of the users.

For some user k belonging in set λ, the decoding process is simple. The received message takes the form

$$\begin{aligned} y_k &= \mathbf{h}_k^T H_\lambda^{-1} \begin{bmatrix} X_{\lambda(1)\cup\mathcal{F}_{r_1}}^{\pi\backslash\mathcal{F}_{r_1},\lambda\backslash\lambda(1)} \\ X_{\lambda(2)\cup\mathcal{F}_{r_2}}^{\pi\backslash\mathcal{F}_{r_2},\lambda\backslash\lambda(2)} \\ \vdots \\ X_{\lambda(L)\cup\mathcal{F}_{r_L}}^{\pi\backslash\mathcal{F}_{r_L},\lambda\backslash\lambda(L)} \end{bmatrix} \\ &= X_{\{k\}\cup\mathcal{F}_{r_k}}^{\pi\backslash\mathcal{F}_{r_k},\lambda\backslash\{k\}}. \end{aligned}$$

Due to the design of the remaining XOR (see eq. (13)), all but one subfiles have been cached by user k, 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 $m \in \pi$ takes the form

$$y_m = \mathbf{h}_m^T H_\lambda^{-1} \begin{bmatrix} X_{\lambda(1)\cup\mathcal{F}_{r_1}}^{\pi\backslash\mathcal{F}_{r_1},\lambda\backslash\lambda(1)} \\ X_{\lambda(2)\cup\mathcal{F}_{r_2}}^{\pi\backslash\mathcal{F}_{r_2},\lambda\backslash\lambda(2)} \\ \vdots \\ X_{\lambda(L)\cup\mathcal{F}_{r_L}}^{\pi\backslash\mathcal{F}_{r_L},\lambda\backslash\lambda(L)} \end{bmatrix} \tag{16}$$

$$= \sum_{j=1}^{L} \mathbf{h}_m^T \mathbf{h}_{\lambda\backslash\lambda(j)}^{-1} X_{\lambda(j)\cup\mathcal{F}_{r_j}}^{\pi\backslash\mathcal{F}_{r_j},\lambda\backslash\lambda(j)}. \tag{17}$$

First, we can observe that due to the process described in algorithm 1, user m has estimated all products $\mathbf{h}_m^T \mathbf{h}_{\lambda\backslash\lambda(j)}^{-1}$, $\forall j \in [L]$ that appear in eq. (17). Then, by taking account of the fact that $\mathcal{F}_{r_i} \cap \mathcal{F}_{r_j} = \emptyset$, $i \neq j$ we can see that user m belongs in one of the $\mathcal{F}_{r_j}$ subsets of π, which means that user m has stored the content of all but one XORs (see eq. (13)), thus can remove them from eq. (17). Further, the remaining XOR, due to its structure (cf. Eq. (13)), is decodable by user m.

D. Calculating the DoF performance

a) Showing that each desired subfile is transmitted exactly once: The first task here is to show that, for a given fixed subfile $W_{\sigma,\tau}^{(d_k)}$, each of the $K\gamma + L$ sub-subfiles (defined by the same fixed (σ, τ, k) and are differentiated using the $K\gamma + L$ different $\phi \in [K\gamma + L]$ – the notation of which, as you may recall, we suppress), will appear in $K\gamma + L$ different transmissions $\mathbf{x}_{\lambda,\pi}^s$, for some λ, π, s.

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 $|\sigma| = L - 1, |\lambda| = L, |\pi| = |\tau| = K\gamma$. 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), since $|\sigma \cup \tau \cup k| = K\gamma + L$ and because $|\sigma \cup \tau| = K\gamma + L - 1$ (which means that $k \notin \sigma \cup \tau$), we can conclude that $\lambda = \sigma \cup \{k\}$. Given also that

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

means that fixing (σ, τ, k), points to a single λ and a single π. For any fixed (λ, π), in our algorithm, step 5 identifies L specific sub-subfiles which are defined by the same (σ, τ, k), thus can be differentiated by L different $\phi \in [K\gamma + L]$; these L sub-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 fixed (σ, τ, k) (which also defines the set of active users) — there can be $K\gamma$ different sets λ which take the following form

$$\lambda = \sigma \cup \tau(i), \quad i \in [K\gamma].$$

This means that fixing (σ, τ, k) corresponds to $K\gamma$ 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 $K\gamma$ 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 2 that user k can belong in exactly 1 set $\mathcal{F}_{r_i}, i \in [L]$, let that be $\mathcal{F}_{r_j}$, which means that from all L transmissions of Step 5, a sub-subfile belonging to category $W_{\sigma,\tau}^{(d_k)}$ will be transmitted in exactly one transmission, i.e. the transmission which will have XOR

$$X_{\tau(i)\cup\mathcal{F}_{r_j}}^{\pi\backslash\mathcal{F}_{r_j},\sigma}$$

In total, for all the different (λ, π) sets, subfile $W_{\sigma,\tau}^{(d_k)}$ will be transmitted $K\gamma + L$ times.

Finally, since we showed that an arbitrary subfile, $W_{\sigma,\tau}^{(d_k)}$, will be transmitted exactly $K\gamma + L$ times, this implies that all subfiles of interest will be transmitted by spanning through all possible λ, π sets.

b) Calculating the DoF performance: The resulting DoF can now be easily seen to be $d_\Sigma = L + K\gamma$ by recalling that each transmission includes $K\gamma + L$ different subfiles, and by recalling that no subfile is ever repeated. A quick verification, accounting for the subpacketization

$$S = \binom{K}{K\gamma}\binom{K-K\gamma-1}{L-1}(K\gamma+L)$$

and accounting for the number of iterations in each step, gives that

$$\mathcal{T} = \frac{\overbrace{\binom{K}{L}}^{Step\ 1}\overbrace{\binom{K-L}{K\gamma}}^{Step\ 3}\cdot\overbrace{L}^{Step\ 5}}{\binom{K}{K\gamma}\binom{K-K\gamma-1}{L-1}(K\gamma+L)} = \frac{K(1-\gamma)\cdot L}{L(K\gamma+L)} = \tag{18}$$

$$= \frac{K(1-\gamma)}{K\gamma+L} \tag{19}$$

which implies a DoF of

$$d_\Sigma = \frac{K(1-\gamma)}{\mathcal{T}} = L + K\gamma.$$

□

The following example employs the complete notation $W^{\phi,(d_k)}_{\sigma,\tau}$ in order to demonstrate the iteration over all subfiles. Similar to before, we use $A^{(\phi)}_{\sigma,\tau}$ to refer to $W^{\phi,(d_1)}_{\sigma,\tau}$, $B^{(\phi)}_{\sigma,\tau}$ to refer to $W^{\phi,(d_2)}_{\sigma,\tau}$, and so on.

***Example* 4** (Example of scheme)**.** *Consider a transmitter with $L = 2$ antennas, serving $K = 4$ users whose caches allow for caching redundancy $K\gamma = 2$. Each file is split into*

$$S = \overbrace{(K\gamma+L)}^{\phi}\overbrace{\binom{K-K\gamma-1}{L-1}}^{\sigma}\overbrace{\binom{K}{K\gamma}}^{\tau} = 24$$

sub-subfiles and the following are the $\binom{K}{L}\binom{K-L}{K\gamma}L = 12$ transmissions that will satisfy all the users' requests.

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

As we see, the delay is $\mathcal{T} = \frac{12}{24} = \frac{1}{2}$ and the sum-DoF is $d_\Sigma = \frac{K(1-\gamma)}{\mathcal{T}} = 4$.

V. Conclusion

In this work we provided a new multi-antenna coded caching algorithm that achieves the maximum known, near-optimal DoF of $L+K\gamma$ served users at a time, with a feedback cost that is a function only of the number of transmit antennas.

Various benefits of reducing feedback: higher effective DoF, separability of design, and feedback reuse

This feedback reduction has multiple beneficial effects. Firstly, the reduced feedback requirements increase the effective DoF, simply because they allow for more time (within any given coherence block) to transmit actual data rather than wasting resources on feedback acquisition. Secondly, the algorithm allows us to increase the number of users and/or their cache size, without having to change the amount of CSIT feedback and without additional training overheads.

Thirdly, the structure of the algorithm allows for feedback acquisition to happen less often, since by choosing one group of users for precoding (i.e., by choosing set λ in Step 1 of Algorithm 2), the resulting H^{-1}_λ (Step 2) can stay fixed for a large number of time slots (can stay fixed for all possible $\pi \subset [K]\setminus\lambda,\ |\pi| = K\gamma$ of Step 3), thus allowing — within a coherence period — for a substantial reuse of the acquired feedback. This is further demonstrated in the following example.

***Example* 5.** *Let us assume a transmitter with $L=5$ antennas, serving K users, with each user being equipped with a cache of normalized size $\gamma = \frac{1}{10}$. The goal is to compare the ability of different algorithms to reuse existing feedback, and we will do so by comparing the fraction of the overall delivery that can be completed, given a certain amount of acquired CSI. In particular, let us assume that the allowable feedback cost*[14] *is C, where for example in TDD this can imply receiving feedback for only C users, corresponding to C uplink and C downlink feedback acquisition slots. The comparison will be between the here proposed algorithm and the state-of-art multi-antenna coded caching algorithms [26], [27]. We first note that for the proposed algorithm, C needs to satisfy $C \geq L = 5$, while for the state-of-art algorithms, C needs to satisfy $C \geq L + K\gamma = 5 + \frac{K}{10}$.*

To understand the ability of the state of art algorithms to reuse feedback, let us recall from [26], [27] that once feedback is acquired for a set of $C \geq L+K\gamma$ users, then there are

$$(K\gamma+L)\binom{C}{L+K\gamma}\binom{L+K\gamma-1}{K\gamma} \tag{20}$$

transmissions[15] *that can take place without the need for additional feedback.*

[14]For simplicity we assume that the coherence block is long enough to 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.

[15]Where we added the first term in order to match the subpacketization of the proposed algorithm. As a result, in either of the compared algorithms, one transmission carries the same amount of information.

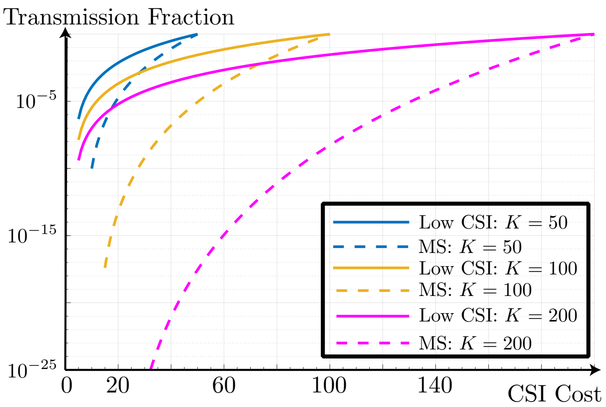

Fig. 2. The CSI cost that is required to complete a fraction of the delivery phase. The cost represents the number of users that need to send feedback. Parameter γ is fixed at value $\gamma = \frac{1}{10}$.

On the other hand, by observing the proposed algorithm, we can see that having feedback for some subset of $C \geq L$ users will allow for

$$L \cdot \binom{C}{L}\binom{K-L}{K\gamma} \tag{21}$$

transmissions. Now comparing (20) *with* (21)*, and given that each transmission in both cases carries the same amount of information, we can conclude that the new algorithm can serve a much larger portion of the delivery. Specifically the new algorithm can serve, for a given CSI cost C,*

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

times more content than the existing algorithms, 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 2, where we show the CSI cost needed to complete any fraction of the entire delivery.

Looking, for example, at the fraction of 10^{-5} and the case of $K = 50$, we see that the proposed algorithm delivers this fraction with feedback cost $C = 7$, while the state-of-art requires a cost of approximately 18*. This effect is further amplified when we visit cases with higher number of users. Specifically, in the case of $K = 100$ the respective costs are* 12 *and* 52*, while in the case of $K = 200$ these costs rise to* 22 *and* 131*, respectively.*

APPENDIX

In this appendix we provide a more involved example that will allow a more in-depth understanding of the mechanics of our algorithm.

***Example* 6.** *We consider the $L = 2$ MISO BC with $K = 6$ users and $\gamma = 2/3$ ($K\gamma = 4$). Then, the required* 30 *transmissions are*

$$\mathbf{x}^1_{12,3456} = H_{12}^{-1}\begin{bmatrix} A^{(1)}_{2,3456} \oplus C^{(1)}_{2,1456} \oplus D^{(1)}_{2,1356} \\ B^{(1)}_{1,3456} \oplus E^{(1)}_{1,2346} \oplus F^{(1)}_{1,2345} \end{bmatrix}$$

$$\mathbf{x}^2_{12,3456} = H_{12}^{-1}\begin{bmatrix} A^{(2)}_{2,3456} \oplus E^{(1)}_{2,1346} \oplus F^{(1)}_{2,1345} \\ B^{(2)}_{1,3456} \oplus C^{(1)}_{1,2456} \oplus D^{(1)}_{1,2356} \end{bmatrix}$$

$$\mathbf{x}^{(1)}_{13,2456} = H_{13}^{-1}\begin{bmatrix} A^{(1)}_{3,2456} \oplus B^{(1)}_{3,1456} \oplus D^{(1)}_{3,1256} \\ C^{(2)}_{1,2456} \oplus E^{(2)}_{1,2346} \oplus F^{(2)}_{1,2345} \end{bmatrix}$$

$$\mathbf{x}^2_{13,2456} = H_{13}^{-1}\begin{bmatrix} A^{(2)}_{3,2456} \oplus E^{(1)}_{3,1246} \oplus F^{(1)}_{3,1245} \\ C^{(3)}_{1,2456} \oplus B^{(3)}_{1,3456} \oplus D^{(2)}_{1,2356} \end{bmatrix}$$

$$\mathbf{x}^1_{14,2356} = H_{14}^{-1}\begin{bmatrix} A^{(1)}_{4,2356} \oplus B^{(1)}_{4,1356} \oplus C^{(1)}_{4,1256} \\ D^{(3)}_{1,2356} \oplus E^{(3)}_{1,2346} \oplus F^{(3)}_{1,2345} \end{bmatrix}$$

$$\mathbf{x}^2_{14,2356} = H_{14}^{-1}\begin{bmatrix} A^{(2)}_{4,2356} \oplus E^{(1)}_{4,1236} \oplus F^{(1)}_{4,1235} \\ D^{(4)}_{1,2356} \oplus B^{(4)}_{1,3456} \oplus C^{(4)}_{1,2456} \end{bmatrix}$$

$$\mathbf{x}^1_{15,2346} = H_{15}^{-1}\begin{bmatrix} A^{(1)}_{5,2346} \oplus B^{(1)}_{5,1346} \oplus C^{(1)}_{5,1246} \\ E^{(4)}_{1,2346} \oplus D^{(5)}_{1,2356} \oplus F^{(4)}_{1,2345} \end{bmatrix}$$

$$\mathbf{x}^2_{15,2346} = H_{15}^{-1}\begin{bmatrix} A^{(2)}_{5,2346} \oplus D^{(1)}_{5,1236} \oplus F^{(1)}_{5,1234} \\ E^{(5)}_{1,2346} \oplus B^{(5)}_{1,3456} \oplus C^{(5)}_{1,2456} \end{bmatrix}$$

$$\mathbf{x}^1_{16,2345} = H_{16}^{-1}\begin{bmatrix} A^{(1)}_{6,2345} \oplus B^{(1)}_{6,1345} \oplus C^{(1)}_{6,1245} \\ F^{(5)}_{1,2345} \oplus D^{(6)}_{1,2356} \oplus E^{(6)}_{1,2346} \end{bmatrix}$$

$$\mathbf{x}^2_{16,2345} = H_{16}^{-1}\begin{bmatrix} A^{(2)}_{6,2345} \oplus D^{(1)}_{6,1235} \oplus E^{(1)}_{6,1234} \\ F^{(6)}_{1,2345} \oplus B^{(6)}_{1,3456} \oplus C^{(6)}_{1,2456} \end{bmatrix}$$

$$\mathbf{x}^1_{23,1456} = H_{23}^{-1}\begin{bmatrix} B^{(2)}_{3,1456} \oplus A^{(3)}_{3,2456} \oplus D^{(2)}_{3,1256} \\ C^{(2)}_{2,1456} \oplus E^{(2)}_{2,1346} \oplus F^{(2)}_{2,1345} \end{bmatrix}$$

$$\mathbf{x}^2_{23,1456} = H_{23}^{-1}\begin{bmatrix} B^{(3)}_{3,1456} \oplus E^{(2)}_{3,1246} \oplus F^{(2)}_{3,1245} \\ C^{(3)}_{2,1456} \oplus A^{(3)}_{2,3456} \oplus D^{(2)}_{2,1356} \end{bmatrix}$$

$$\mathbf{x}^1_{24,1356} = H_{24}^{-1}\begin{bmatrix} B^{(2)}_{4,1356} \oplus A^{(3)}_{4,2356} \oplus C^{(2)}_{4,1256} \\ D^{(3)}_{2,1356} \oplus E^{(3)}_{2,1346} \oplus F^{(3)}_{2,1345} \end{bmatrix}$$

$$\mathbf{x}^2_{24,1356} = H_{24}^{-1}\begin{bmatrix} B^{(3)}_{4,1356} \oplus E^{(2)}_{4,1236} \oplus F^{(2)}_{4,1235} \\ D^{(4)}_{2,1356} \oplus A^{(4)}_{2,3456} \oplus C^{(4)}_{2,1456} \end{bmatrix}$$

$$\mathbf{x}^1_{25,1346} = H_{25}^{-1}\begin{bmatrix} B^{(2)}_{5,1346} \oplus A^{(3)}_{5,2346} \oplus C^{(2)}_{5,1246} \\ E^{(4)}_{2,1346} \oplus D^{(5)}_{2,1356} \oplus F^{(4)}_{2,1345} \end{bmatrix}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

By examining any of the above transmitted vectors we can deduce that each transmission serves a total of 6 *users, while the feedback cost is* 2 *training slots for CSIT and* 2 *training slots for global CSIR.*

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," *to appear in International Symposium of Information Theory (ISIT)*, 2018.

[3] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *IEEE Information Theory Workshop (ITW)*, 2016.

[4] 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.

[5] 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.

[6] A. Lozano, R. W. Heath Jr, and J. G. Andrews, "Fundamental limits of cooperation," *arXiv preprint arXiv:1204.0011*, 2012.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] H. Weingarten, S. Shamai, and G. Kramer, "On the compound MIMO Broadcast Channel," in *Proceedings of Annual Information Theory and Applications Workshop UCSD*, Citeseer, 2007.

[13] T. Gou, S. A. Jafar, C. Wang, *et al.*, "On the degrees of freedom of finite state compound wireless networks.," *IEEE Trans. Information Theory*, vol. 57, no. 6, pp. 3286–3308, 2011.

[14] M. A. Maddah-Ali, "On the degrees of freedom of the compound MISO Broadcast Channels with finite states," in *International Symposium on Information Theory Proceedings (ISIT)*, pp. 2273–2277, IEEE, 2010.

[15] 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.

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

[17] 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.

[18] 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.

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

[20] J. Chen and P. Elia, "Toward the performance vs. feedback tradeoff for the two-user miso broadcast channel," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8336–8356, 2013.

[21] 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.

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

[23] A. Destounis, M. Kobayashi, G. Paschos, and A. Ghorbel, "Alpha fair Coded Caching," in *15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2017.

[24] L. Zheng, Q. Yan, Q. Chen, and X. Tang, "On the coded caching delivery design over wireless networks," *arXiv preprint arXiv:1611.04853*, 2016.

[25] 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.

[26] 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.

[27] 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.

[28] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna Coded Caching," in *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[29] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multi-antenna interference management for coded caching," *arXiv preprint arXiv:1711.03364*, 2017.

[30] J. S. P. Roig, D. Gündüz, and F. Tosato, "Interference networks with caches at both ends," in *2017 IEEE International Conference on Communications (ICC)*, May 2017.

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

[32] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communications*, 2018.

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

[34] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *Proc. Allerton Conference*, 2015.

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

[36] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized degrees of freedom of the symmetric cache-aided MISO broadcast channel with partial CSIT," *arXiv preprint arXiv:1712.05244*, 2017.
[37] 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.
[38] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *arXiv preprint arXiv:1806.07800*, 2018.