Feedback-Aided Coded Caching for the MISO BC with Small Caches

Jingjing Zhang and Petros Elia

***Abstract*—This work explores coded caching in the symmetric K-user cache-aided MISO BC with imperfect CSIT-type feedback, for the specific case where the cache size is much smaller than the library size. Building on the recently explored synergy between caching and delayed-CSIT, and building on the tradeoff between caching and CSIT quality, the work proposes new schemes that boost the impact of small caches, focusing on the case where the cumulative cache size is smaller than the library size. This small-cache setting places an additional challenge due to the fact that some of the library content must remain entirely uncached, which forces us to dynamically change the caching redundancy to compensate for this. Our proposed scheme is near-optimal, and the work identifies the optimal cacheaided degrees-of-freedom (DoF) performance within a factor of 4.**

I. Introduction

In the setting of the single-stream broadcast channel, the seminal work in [1] proposed *coded caching* as a technique which — after carefully caching content at the receivers, and properly coding across different users' requested data — provided increased effective throughput and a reduced network load. By using coding to create multicast opportunities — even when users requested different data content — coded caching allowed per-user DoF gains that were proportional to the cache sizes. The fact though that such caches can be comparably small [2], brings to the fore the need to understand how we must efficiently combine reduced caching resources, with any additional complementary resources — such as feedback and spatial dimensions — that may be available in communication networks.

Our aim here is to explore this concept, in the symmetric K-user cache-aided wireless MISO BC. Following in the footsteps of [3], our aim here is to further our understanding of the *joint* effect of coded caching — now with small caches — and (variable quality) feedback, in removing interference and in eventually improving performance. This joint exposition is natural and important because caching and feedback are both powerful and scarce ingredients in wireless networks, and because these two ingredients are intimately connected. These connections will prove particularly crucial here, in boosting the effect of otherwise insufficiently large caches, or otherwise insufficiently refined feedback. The coding challenge here — for the particular case of small caches — is to find a way to ameliorate the negative effect of having to leave some content entirely uncached, which is a problem which paradoxically can become more pronounced in the presence of CSIT resources, as we will see later on.

The authors are with the Communication Systems Department at EURECOM, Sophia Antipolis, 06410, France (email: jingjing.zhang@eurecom.fr, elia@eurecom.fr). The work of Petros Elia was supported by the ANR Jeunes Chercheurs project ECOLOGICAL-BITS-AND-FLOPS.

An initial version of this paper has been reported as Research Report No. RR-15-307 at EURECOM, August 25, 2015, http://www.eurecom.fr/publication/4723 as well as was uploaded on arxiv (version 1) in November 2015.

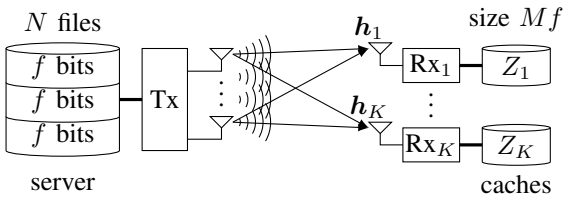

Fig. 1. System setup of cache-aided K-user MISO BC.

A. K-user feedback-aided symmetric MISO BC with small caches

We consider the symmetric K-user wireless MISO BC with a K-antenna transmitter, and K single-antenna receiving users. The transmitter has access to a library of $N \geq K$ distinct files $W_1, W_2, \ldots, W_N$, each of size $|W_n| = f$ bits. Each user $k \in \{1, 2, \ldots, K\}$ has a cache Z_k, of size $|Z_k| = Mf$ bits, where naturally $M \leq N$. A normalized measure of caching resources will take the form

$$\gamma := \frac{M}{N}. \tag{1}$$

Our emphasis here will be on the small cache regime where the cumulative cache size is less than the library size ($K\gamma \leq 1$, i.e., $KM \leq N$), and which will force us to account for the fact that not all content can be cached. We will also touch upon the general small-cache setting where the individual cache size is much less than the library size ($M \ll N$, i.e., $\gamma \ll 1$).

As in [1], communication consists of the aforementioned *content placement phase* (typically taking place during off-peak hours) and the *delivery phase*. During the placement phase, the caches are pre-filled with content from the N files $\{W_n\}_{n=1}^N$ of the library. The delivery phase commences when each user $k = 1, \ldots, K$ requests from the transmitter, any *one* file $W_{R_k} \in \{W_n\}_{n=1}^N$, out of the N library files. Upon notification of the users' requests, the transmitter aims to deliver the (remaining of the) requested files, each to their intended receiver, and the challenge is to do so over a limited (delivery phase) duration T.

a) Channel model: For each transmission, the received signals at each user k, will be modeled as

$$y_k = \boldsymbol{h}_k^T \boldsymbol{x} + z_k, \quad k = 1, \ldots, K \tag{2}$$

where $\boldsymbol{x} \in \mathbb{C}^{K\times 1}$ denotes the transmitted vector satisfying a power constraint $\mathbb{E}(||\boldsymbol{x}||^2) \leq P$, where $\boldsymbol{h}_k \in \mathbb{C}^{K\times 1}$ denotes the channel of user k in the form of the random vector of fading coefficients that can change in time and space, and where z_k represents unit-power AWGN noise at receiver k. At

the end of the delivery phase, each receiving user k combines the received signal observations y_k — accumulated during the delivery phase — with the fixed information in their respective cache Z_k, to reconstruct their desired file W_{R_k}.

b) Feedback model: Communication will also take place in the presence of channel state information at the transmitter. Motivated by the fact that CSIT-type feedback is typically hard to obtain in a timely and reliable manner, we will here consider the mixed CSIT model (cf. [4], see also [5]) where feedback offers a combination of imperfect-quality current (instantaneously available) CSIT, together with additional (perfect-accuracy) delayed CSIT. In this setting, the channel estimation error of the current channel state is assumed to scale in power as $P^{-\alpha}$, for some CSIT quality exponent[1]

$$\alpha := -\lim_{P\to\infty} \frac{\log \mathbb{E}[||\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k||^2]}{\log P},\ k \in \{1,\dots,K\} \qquad (3)$$

where $\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k$ denotes the estimation error between the current CSIT estimate $\hat{\boldsymbol{h}}_k$ and the estimated channel $\boldsymbol{h}_k$.

This mixed CSIT model, in addition to being able to capture different realistic scenarios such as that of using predictions and feedback to get an estimate of the current state of a time-correlated channel, it is also well suited for cache aided networks because, by mixing the effect of delayed and current feedback, it allows us to concisely capture the powerful synergies between caching and delayed CSIT (cf. [7]) as well as the tradeoffs between feedback quality and cache size (cf. [3]).

B. Measures of performance, notation and assumptions

1) Measures of performance in current work: Our aim is to design schemes that, for any $K\gamma < 1$ and any $\alpha \in [0,1]$, reduce the duration $T(\gamma,\alpha)$ — in time slots, per file served per user — needed to complete the delivery process, *for any request*.

2) Notation: We will use

$$\Gamma := \frac{KM}{N} = K\gamma \qquad (4)$$

to represent the cumulative (normalized) cache size, in the sense that the sum of the sizes of the caches across all users, is a fraction Γ of the volume of the N-file library. We will also use the notation $H_n := \sum_{i=1}^{n} \frac{1}{i}$, to represent the n-th harmonic number, and we will use $\epsilon_n := H_n - \log(n)$ to represent its logarithmic approximation error, for some integer n. We remind the reader that ϵ_n decreases with n, and that $\epsilon_\infty := \lim_{n\to\infty} H_n - \log(n)$ is approximately 0.5772. $\mathbb{Z}$ will represent the integers, $\mathbb{Z}^+$ the positive integers, $\mathbb{R}$ the real numbers, $\binom{n}{k}$ the n-choose-k operator, and $\oplus$ the bitwise XOR operation. We will use $[K] := \{1,2,\cdots,K\}$. If ψ is a set, then $|\psi|$ will denote its cardinality. For sets A and B, then $A\backslash B$ denotes the difference set. Complex vectors will be denoted by lower-case bold font. We will use $||\boldsymbol{x}||^2$ to denote the magnitude of a vector $\boldsymbol{x}$ of complex numbers. For a transmitted vector $\boldsymbol{x}$, we will use $\text{dur}(\boldsymbol{x})$ to denote the transmission duration of that vector. For example, having $\text{dur}(\boldsymbol{x}) = \frac{1}{10}T$ would simply mean that the transmission of vector $\boldsymbol{x}$ lasts one tenth of the delivery phase. We will also use $\doteq$ to denote *exponential equality*, i.e., we write $g(P) \doteq P^B$ to denote $\lim_{P\to\infty} \frac{\log g(P)}{\log P} = B$. Similarly $\dot{\geq}$ and $\dot{\leq}$ will denote exponential inequalities. Logarithms are of base e, unless we use $\log_2(\cdot)$ which will represent a logarithm of base 2.

3) Main assumptions: In addition to the aforementioned mixed CSIT assumptions, we will adhere to the common convention (see for example [8]) of assuming perfect and global knowledge of delayed channel state information at the receivers (delayed global CSIR), where each receiver must know (with delay) the CSIR of (some of the) other receivers. We will assume that the entries of *each specific* estimation error vector are i.i.d. Gaussian. For the outer (lower) bound to hold, we will make the common assumption that the current channel state must be independent of the previous channel-estimates and estimation errors, *conditioned on the current estimate* (there is no need for the channel to be i.i.d. in time). Furthermore, as with most works on coded caching, we will assume uniform file popularity, as well as that $N \geq K$.

C. Prior work

In terms of feedback, our work builds on many different works including [8], as well as other subsequent works [4], [5], [9]–[16] that incorporate different CSIT-quality considerations. In terms of caching, our work is part of a sequence of works (cf. [17]–[22]) that is inspired by the work in [1] and which try to understand the limits of coded caching in different scenarios. Additional interesting works include [19], [23]–[30], as well as the work in [31] which considered coded caching — in the single stream case with $K \geq N$ — in the presence of very small caches with $KM \leq 1$, corresponding to the case where pooling all the caches together, can at most match the size of a single file.

In spirit, our work is closer to different works that deviate from the setting of having single-stream error free links, such as the works by Timo and Wigger in [32] and by Ghorbel et al. [33] on the erasure broadcast channel, the work by Maddah-Ali and Niesen in [34] on the wireless interference channel with transmitter-side caching, and our work in [35].

II. Main results

The following identifies, up to a factor of 4, the optimal T^*, for all $\Gamma \in [0,1]$. We use the expression

$$\alpha_{b,\eta} = \frac{\eta - \Gamma}{\Gamma(H_K - H_\eta - 1) + \eta},\ \eta = 1,\dots,K-1. \qquad (5)$$

Theorem 1: In the (K,M,N,α) cache-aided MISO BC with N files, $K \leq N$ users, and $KM \leq N$ ($\Gamma \leq 1$), then for $\eta = 1,\dots,K-2$,

$$T = \begin{cases} \frac{H_K - \Gamma}{1-\alpha+\alpha H_K}, & 0 \leq \alpha < \alpha_{b,1} \\ \frac{(K-\Gamma)(H_K - H_\eta)}{(K-\eta)+\alpha(\eta+K(H_K-H_\eta-1))}, & \alpha_{b,\eta} \leq \alpha < \alpha_{b,\eta+1} \\ 1-\gamma, & \frac{K-1-\Gamma}{(K-1)(1-\gamma)} \leq \alpha \leq 1 \end{cases} \qquad (6)$$

is achievable, and has a gap from optimal that is less than 4 ($\frac{T}{T^*} < 4$), for all α, K. For $\alpha \geq \frac{K-1-\Gamma}{(K-1)(1-\gamma)}$, T is optimal.

[1]The range of interest is $\alpha \in [0,1]$; in the high SNR regime of interest here, $\alpha = 0$ corresponds to having essentially no current CSIT, while having $\alpha = 1$ corresponds (again in the high SNR regime) to perfect and immediately available CSIT (cf. [6]).

Proof: The scheme that achieves the above performance is presented in Section III, while the corresponding gap to optimal is bounded in Section IV. ∎

Furthermore directly from the above, for $\alpha = 0$, we have the following.

Corollary 1a: In the MISO BC with $\Gamma \leq 1, \alpha = 0$, then

$$T = H_K - \Gamma \tag{7}$$

is achievable and has a gap from optimal that is less than 4.

Directly from Theorem 1, we have the following corollary which quantifies the CSIT savings

$$\delta(\gamma,\alpha) := \arg\min_{\alpha'} \{\alpha' : (1-\gamma)T^*(\gamma=0,\alpha') \leq T(\gamma,\alpha)\} - \alpha$$

that we can have as a result of properly exploiting small caches. This reflects the CSIT reductions (from $\alpha + \delta(\gamma,\alpha)$ to the operational α) that can be achieved due to coded caching, without loss in performance.

Corollary 1b: In the (K, M, N, α) cache-aided BC with $\Gamma \leq 1$, then

$$\delta(\gamma,\alpha) = \begin{cases} \frac{\gamma(K-H_K)}{H_K - K\gamma}(\alpha + \frac{1}{H_K - 1}), & 0 \leq \alpha < \alpha_{b,1} \\ \frac{(1-\alpha)(KH_\eta - \eta H_K)}{KH_{\eta+1}(H_K-1)}, & \alpha_{b,\eta} \leq \alpha < \alpha_{b,\eta+1} \\ 1-\alpha, & \alpha \geq \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)}. \end{cases} \tag{8}$$

The last case in the above equation shows how, in the presence of caching, we need not acquire CSIT quality that exceeds $\alpha = \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)}$.

III. Cache-aided QMAT with very small caches

We now describe the communication scheme. Part of the challenge, and a notable difference from the case of larger caches, is that due to the fact that now $\Gamma < 1$, some of the library content must remain entirely uncached. This uncached part is delivered by employing a combination of multicasting and ZF which uses current CSIT. The problem though remains when α is small because then current CSIT can only support a weak ZF component, which in turn forces us to send some of this uncached private information using multicasting, which itself must be calibrated not to intervene with the multicasting that utilizes side information from the caches. For this range of smaller α, our scheme here will differ from that when α is big (as well as from the scheme for $\Gamma \geq 1$). When α is bigger than a certain threshold value $\alpha_{b,1}$, we will choose to cache even less data from the library, which though we will cache with higher redundancy[2]. Calibrating this redundancy as a function of α, allows us to strike the proper balance between ZF and delayed-CSIT aided coded caching. For this latter part, we will use our scheme from [3] which we do not describe here.

We consider the range[3] $\alpha \in [0, \alpha_{b,1}]$, and proceed to set $\eta = 1$ (cf. (5) from Theorem 1), such that there is no overlapping content in the caches ($Z_k \cap Z_i = \varnothing$).

[2]Higher redundancy here implies that parts of files will be replicated in more caches.

[3]The remaining range of α will be briefly addressed at the end of this section.

A. Placement phase

During the placement phase, each of the N files $W_n, n = 1, 2, \ldots, N$ ($|W_n| = f$ bits) in the library, is split into two parts

$$W_n = (W_n^c, W_n^{\bar{c}}) \tag{9}$$

where W_n^c (c for 'cached') will be placed in different caches, while the content of $W_n^{\bar{c}}$ ($\bar{c}$ for 'non-cached') will never be cached anywhere, but will instead be communicated — using current and delayed CSIT — in a manner that avoids interference without depending on caches. The split is such that

$$|W_n^c| = \frac{KMf}{N} = K\gamma f \text{ bits.} \tag{10}$$

Then, we equally divide W_n^c into K subfiles $\{W_{n,k}^c\}_{k\in[K]}$, where each subfile has size

$$|W_{n,k}^c| = \frac{Mf}{N} = \gamma f \text{ bits} \tag{11}$$

and the caches are filled as follows

$$Z_k = \{W_{n,k}^c\}_{n\in[N]} \tag{12}$$

such that each subfile $W_{n,k}^c$ is stored in Z_k.

B. Delivery phase

Upon notification of the requests $W_{R_k}, k = 1, \ldots, K$, we first further split $W_{R_k,k}^{\bar{c}}$ into two parts, $W_{R_k,k}^{\bar{c},p}$ and $W_{R_k,k}^{\bar{c},\bar{p}}$ that will be delivered in two different ways that we describe later, and whose sizes are such that

$$|W_{R_k,k}^{\bar{c},p}| = \alpha f T, \quad |W_{R_k,k}^{\bar{c},\bar{p}}| = f(1 - K\gamma - \alpha T). \tag{13}$$

Then we fold all $W_{R_k,\psi\setminus\{k\}}^c$ to get a set

$$X_\psi := \oplus_{k\in\psi} W_{R_k,\psi\setminus\{k\}}^c, \psi \in \Psi_2 \tag{14}$$

of so-called *order-2 XORs* (each XOR is meant for two users), and where $\Psi_2 := \{\psi \in [K] : |\psi| = 2\}$. Each of these XORs has size

$$|X_\psi| = \gamma f \text{ bits} \tag{15}$$

and they jointly form the XOR set

$$\mathcal{X}_\Psi := \{X_\psi = \oplus_{k\in\psi} W_{R_k,\psi\setminus\{k\}}^c\}_{\psi\in\Psi_2} \tag{16}$$

of cardinality $|\mathcal{X}_\Psi| = \binom{K}{2}$.

In the end, we must deliver

- $W_{R_k}^{\bar{c},p}$, $k = 1, \cdots, K$, privately to user k, using mainly current CSIT
- $W_{R_k}^{\bar{c},\bar{p}}$, $k = 1, \cdots, K$, using mainly delayed CSIT
- $\{W_{R_k,\psi\setminus\{k\}}^c\}_{\psi\in\Psi_2}$, $k = 1, \cdots, K$ by delivering the XORs from $\mathcal{X}_\Psi$, each to their intended pair of receivers.

This delivery is described in the following.

a) Transmission: We describe how we adapt the QMAT algorithm from [36] to deliver the aforementioned messages, with delay T.

While we will not go into all the details of the QMAT scheme, we note that some aspects of this scheme are similar to MAT schemes (cf. [8]), and a main new element is that QMAT applies digital transmission of interference, and a double-quantization method that collects and distributes residual interference across different rounds, in a manner that allows for ZF and MAT to coexist at maximal rates. The main ingredients include MAT-type symbols of different degrees of multicasting, ZF-type symbols for each user, and auxiliary symbols that diffuse interference across different phases and rounds. Many of the details of this scheme are 'hidden' behind the choice of $\mathbf{G}_{c,t}$ and behind the loading of the MAT-type symbols and additional auxiliary symbols that are all represented by $\boldsymbol{x}_{c,t}$ below. Another important element involves the use of caches to 'skip' MAT phases, as well as a careful rate- and power-allocation policy.

The QMAT algorithm has K transmission phases. For each phase $i = 1, \cdots, K$, the QMAT data symbols are intended for a subset $\mathcal{S} \subset [K]$ of users, where $|\mathcal{S}| = i$. Here by adapting the algorithm, at each instance $t \in [0, T]$ through the transmission, the transmitted vector takes the form

$$\boldsymbol{x}_t = \mathbf{G}_{c,t}\boldsymbol{x}_{c,t} + \sum_{\ell \in \bar{\mathcal{S}}} \boldsymbol{g}_{\ell,t} a^*_{\ell,t} + \sum_{k=1}^{K} \boldsymbol{g}_{k,t} a_{k,t} \tag{17}$$

with $\boldsymbol{x}_{c,t}$ being a K-length vector for QMAT data symbols, with $a^*_{\ell,t}$ being an auxiliary symbol that carries residual interference, where $\bar{\mathcal{S}}$ is a set of 'undesired' users that changes every phase, and where each unit-norm precoder $\boldsymbol{g}_{k,t}$ for user $k = 1, 2, \ldots, K$, is simultaneously orthogonal to the CSI estimate for the channels of all other users ($\boldsymbol{g}_{l,t}$ acts the same), thus guaranteeing

$$\hat{\boldsymbol{h}}^T_{k',t}\boldsymbol{g}_{k,t} = 0, \quad \forall k' \in [K]\backslash k. \tag{18}$$

Each precoder $\mathbf{G}_{c,t}$ is defined as $\mathbf{G}_{c,t} = [\boldsymbol{g}_{c,t}, \mathbf{U}_{c,t}]$, where $\boldsymbol{g}_{c,t}$ is simultaneously orthogonal to the channel estimates of the undesired receivers, and $\mathbf{U}_{c,t} \in \mathbb{C}^{K\times(K-1)}$ is a randomly chosen, isotropically distributed unitary matrix.

The rates and the power are set by the QMAT algorithm, such that:

- each $\boldsymbol{x}_{c,t}$ carries $f(1-\alpha)\text{dur}(\boldsymbol{x}_{c,t})$ bits,
- each $a^*_{\ell,t}$ carries $\min\{f(1-\alpha), f\alpha\}\text{dur}(\boldsymbol{g}_{\ell,t}a^*_{\ell,t})$ bits,
- each $a_{k,t}$ carries $f\alpha\text{dur}(\boldsymbol{g}_{k,t}a_{k,t})$ bits,
- and

$$\mathbb{E}\{|\boldsymbol{x}_{c,t}|^2_1\} = \mathbb{E}\{|a^*_{\ell,t}|^2\} \doteq P$$
$$\mathbb{E}\{|\boldsymbol{x}_{c,t}|^2_i\} \doteq P^{1-\alpha}, \mathbb{E}\{|a_{k,t}|^2\} \doteq P^{\alpha}$$

  where $|\boldsymbol{x}_{c,t}|_i, i = 1, 2, \cdots, K$, denotes the magnitude of the i^{th} entry of vector $\boldsymbol{x}_{c,t}$.

The scheme here employs a total of $2K-1$ phases (rather than the K phases in the original Q-MAT), where during the first $K-1$ phases (labeled here as phases $j = 1, \ldots, K-1$), the vector $\boldsymbol{x}_{c,t}$ carries the folded messages $X_\psi \in \mathcal{X}_\Psi$ using the last $K-1$ phases of the MAT algorithm from [8], while for phases $j = K, \ldots, 2K-1$, $\boldsymbol{x}_{c,t}$ now carries $\{W^{\overline{c},\overline{p}}_{R_k}\}_{k\in[K]}$ using the entirety of MAT. In addition, for all $2K-1$ phases, the different $a_{k,t}$ will carry (via ZF) all of the uncached $W^{\overline{c},p}_{R_k}, k = 1, \ldots, K$. The power and rate allocation guarantee that these MAT and ZF components can be carried out in parallel with the assistance of the auxiliary symbols from the next round[4].

In the following, we use T_j to denote the duration of phase j, and $T^{(1)} := \sum_{j=1}^{K-1} T_j$ to denote the duration of the first $K-1$ phases.

b) Summary of transmission scheme for delivery of $\{X_\psi\}_{\psi\in\Psi_2}$*:* Here, $\boldsymbol{x}_{c,t},\ t \in [0, T^{(1)}]$ will have the structure defined by the last $K-1$ phases of (one round of) the QMAT algorithm.

During the first phase ($t \in [0, T_1]$, corresponding to phase 2 of QMAT, where $|\mathcal{S}| = 2$), $\boldsymbol{x}_{c,t}$ will convey all the order-2 messages in $\{X_\psi\}_{\psi\in\Psi_2}$ (each ψ corresponds to each $\mathcal{S}$). Then, at the end of this phase, for each $\psi \in \Psi_2$, and for each $k \in \psi$, the received signal at user k, takes the form

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\mathbf{G}_{c,t}\boldsymbol{x}_{c,t}}_{\text{power } \doteq P} + \underbrace{\boldsymbol{h}^T_{k,t}\sum_{\ell\in\psi}\boldsymbol{g}_{\ell,t}a_{\ell,t}}_{\doteq P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{P^{\alpha}} \tag{19}$$

while the received signal for the other users $k \in [K]\backslash\psi$ takes the form

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a^*_{k,t}}_{\text{power } \doteq P} + \underbrace{\boldsymbol{h}^T_{k,t}\Big(\sum_{\substack{\ell\in\psi\\ \ell\neq k}}\boldsymbol{g}_{\ell,t}a^*_{\ell,t} + \mathbf{G}_{c,t}\boldsymbol{x}_{c,t}\Big)}_{L_{\psi,k'},\ \doteq P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{P^{\alpha}} \tag{20}$$

where in both cases, we ignored the Gaussian noise and the ZF noise up to P^0. Following basic MAT techniques, the interference $L_{\psi,k'}, \forall k'$ is translated into order-3 messages and will be sent in phase $j = 2$. In addition, to separate $\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}$ from the MAT component, as in [36], we use auxiliary data symbols $a^*_{\ell,t}$. Specifically, $L_{\psi,k'}$ is first quantized with $(1-2\alpha)^+ \log P$ bits, leaving the quantizaiton noise $n_{\psi,k'}$ with power scaling in P^α. Then, the transmitter quantizes this quantization noise $n_{\psi,k'}$ with $\alpha \log P$ bits up to the noise level, which will be carried by the auxiliary data symbols in the corresponding phase in the next round. In this way, $\boldsymbol{x}_{c,t}$ *can be decoded using the auxiliary data symbols of the next round, and using order-3 messages from the next phase*.

Given that the allocated 'rate' for $\boldsymbol{x}_{c,t}$ is $(1-\alpha)f$, and given that there is a total of $|\mathcal{X}_\Psi| = \binom{K}{2}$ different order-2 folded messages X_ψ ($|X_\psi| = \gamma f$ bits), the duration T_1 of the first phase, takes the form $T_1 = \frac{\binom{K}{2}|X_\psi|}{(K-1)(1-\alpha)f} = \frac{\gamma\binom{K}{2}}{(K-1)(1-\alpha)}$.

For phases $j = 2, \cdots, K-1$ here (which draw from the last $K-2$ phases in [36]), we can similarly calculate their duration T_j to be $T_j = \frac{2}{j+1}T_1$, which in turn implies that

$$T^{(1)} = \sum_{j=1}^{K-1} T_j = T_1 \sum_{j=1}^{K-1} \frac{2}{1+j} = \frac{\Gamma(H_K - 1)}{1-\alpha}. \tag{21}$$

c) Transmission of $\{W^{\overline{c},\overline{p}}_{R_k}\}_{k\in[K]}$*:* Now the remaining information from $\{W^{\overline{c},\overline{p}}_{R_k}\}_{k\in[K]}$, will be conveyed by $\boldsymbol{x}_{c,t}, t \in [T^{(1)}, T]$ (phases $K, \ldots, 2K-1$), where now though we will

[4]We here focus, for ease of description, on describing only one round. For more details on the multi-round structure of the QMAT, please see [36].

use all the phases of the Q-MAT algorithm because now there is no corresponding side information in the caches to help us 'skip' phases. During the first phase of this second part (i.e., during phase $j = K$), we place all of $\{W_{R_k}^{\overline{c},\overline{p}}\}_{k\in[K]}$ in $\boldsymbol{x}_{c,t}, t \in [T^{(1)}, T^{(1)} + T_K]$. Given the allocated rate $(1-\alpha)f$ for $\boldsymbol{x}_{c,t}$, and given that $|\{W_{R_k}^{\overline{c},\overline{p}}\}_{k\in[K]}| = Kf(1-\frac{KM}{N}-\alpha T)$, we see that

$$T_K = \frac{Kf(1-\frac{KM}{N}-\alpha T)}{K(1-\alpha)f} = \frac{(1-\Gamma-\alpha T)}{(1-\alpha)}. \tag{22}$$

Similarly we see that $T_j = \frac{1}{j-K+1}T_K,\ j = K,\dots,2K-1$, which means that

$$\begin{aligned} T - T^{(1)} &= \sum_{j=K}^{2K-1} T_j = T_K \sum_{j=K}^{2K-1} \frac{1}{j-K+1} \\ &= \frac{H_K(1-\Gamma-\alpha T)}{(1-\alpha)}. \end{aligned} \tag{23}$$

Combining (21) and (23), gives the desired

$$T = \frac{H_K - \Gamma}{1-\alpha+\alpha H_K}. \tag{24}$$

d) Communication scheme for $\alpha \in [\alpha_{b,1}, \alpha_{b,K-1}]$: Here, when $\alpha \geq \alpha_{b,1}$, the scheme already exists; we use

$$\eta = \arg \max_{\eta' \in [\Gamma, K-1]\cap\mathbb{Z}} \{\eta' \ :\ \alpha_{b,\eta'} \leq \alpha\} \tag{25}$$

where

$$\alpha_{b,\eta} = \frac{\eta - \Gamma}{\Gamma(H_K - H_\eta - 1) + \eta} \tag{26}$$

and directly from the algorithm designed for the case of $\Gamma \geq 1$ in [3], we get

$$T = \max\{1-\gamma,\ \frac{(K-\Gamma)(H_K - H_\eta)}{(K-\eta)+\alpha(\eta + K(H_K - H_\eta - 1))}\}. \tag{27}$$

IV. Bounding the gap to optimal

This section presents the proof that the gap $\frac{T(\gamma,\alpha)}{T^*(\gamma,\alpha)}$, between the achievable $T(\gamma,\alpha)$ and the optimal $T^*(\gamma,\alpha)$, is always upper bounded by 4, which also serves as the proof of identifying the optimal $T^*(\gamma,\alpha)$ within a factor of 4. The outer bound (lower bound) on the optimal T^*, is taken from [3], and it takes the form

$$T^*(\gamma,\alpha) \geq \max_{s\in\{1,\dots,\lfloor\frac{N}{M}\rfloor\}} \frac{1}{(H_s\alpha + 1 - \alpha)}(H_s - \frac{Ms}{\lfloor\frac{N}{s}\rfloor}). \tag{28}$$

We will here only prove the case of $\alpha = 0, \Gamma \leq 1$, further focusing on the harder case of $K \geq 25$. Due to lack of space, the proof for the rest of the cases can be found in the longer version in [37].

Our aim here is to show that $\frac{T}{T^*} < 4$, where we use the above lower bound, and where we recall that the achievable T took the form

$$T = H_K - K\gamma.$$

We first see that

$$\frac{T}{T^*} \leq \frac{H_K - K\gamma}{\max\limits_{s\in\{1,\dots,\lfloor\frac{N}{M}\rfloor\}} H_s - \frac{Ms}{\lfloor\frac{N}{s}\rfloor}} \leq \frac{H_K - K\gamma}{\max\limits_{s\in\{1,\dots,\lfloor\frac{N}{M}\rfloor\}} H_s - \frac{\gamma s^2}{1-\frac{s-1}{N}}} \tag{29}$$

$$\leq \frac{H_K - K\gamma}{\max\limits_{s\in\{1,\dots,\lfloor\frac{N}{M}\rfloor\}} H_s - \frac{\gamma s^2}{1-\frac{s-1}{K}}} \tag{30}$$

$$\leq \frac{H_K - K\gamma}{H_{s_c} - \frac{K\gamma s_c^2}{K-s_c+1}} =: f_o(\gamma, s_c) \tag{31}$$

where (29) holds because $\lfloor\frac{N}{s}\rfloor \leq \frac{N-(s-1)}{s}$, where (30) holds because $N \geq K$, and where the last step holds because $\gamma \leq \frac{1}{K}$ and because we choose $s_c = \lfloor\sqrt{K}\rfloor$. We proceed to split the proof in two parts: one for $K \geq 25$, and one for $2 \leq K \leq 25$.

1) Case 1 ($\alpha = 0, K \geq 25$): Here we see that the derivative of $f_o(\gamma, s_c)$ takes the form

$$\frac{df_o(\gamma,s_c)}{d\gamma} = \frac{K}{A}(\frac{H_K s_c^2}{K-s_c+1} - H_{s_c}) \tag{32}$$

$$\geq \frac{K\log K}{A}(\frac{(\sqrt{K}-1)^2}{K-\sqrt{K}+2} - \frac{1}{2}) \tag{33}$$

$$= \frac{K\log(K)}{A}(\frac{1}{2} - \frac{\sqrt{K}+1}{K-\sqrt{K}+2}) \tag{34}$$

$$\geq 0 \tag{35}$$

where A is easily seen to be positive, where the second step is because $\sqrt{K}-1 \leq s_c \leq \sqrt{K}$ and $H_K \geq \log(K)$, and where the last step is because $0 \leq \frac{\sqrt{K}+1}{K-\sqrt{K}+2} \leq \frac{1}{2}$. Hence

$$\max_{\gamma\in[0,\frac{1}{K}]} f_o(\gamma, s_c) = f_o(\gamma = \frac{1}{K}, s_c) = \frac{H_K - 1}{H_{s_c} - \frac{s_c^2}{K-s_c+1}}. \tag{36}$$

Now it is easy to see that $\frac{s_c^2}{K-s_c-1} \leq \frac{K}{K-\sqrt{K}+1}$ since $s_c = \lfloor\sqrt{K}\rfloor \leq \sqrt{K}$. Now consider the function

$$f(K) := \frac{K}{K-\sqrt{K}+1} - \frac{\log K}{4}$$

and let us calculate its derivative

$$\frac{df(K)}{dK} = \frac{1-\frac{\sqrt{K}}{2}}{(K-\sqrt{K}+1)^2} - \frac{1}{4K} < 0$$

which we see to be negative for any $K \geq 36$. This allows us to conclude that $\max\limits_{K\in[25,\infty]} f(K) = f(25) = 0.3858$, and also that $\frac{s_c^2}{K-s_c-1} \leq \frac{\log K}{4} + 0.3858$.

Now let us go back to (36), where using the above maximization, we can get

$$f_o(\gamma = \frac{1}{K}, s_c) \leq \frac{H_K - 1}{H_{s_c} - (\frac{\log K}{4} + 0.3858)}$$

$$\leq \frac{H_K - 1}{\frac{1}{2}\log K + \epsilon_\infty + \log\frac{4}{5} - (\frac{\log K}{4} + 0.3858)}$$

$$= \frac{\log K + \epsilon_{25} - 1}{\frac{1}{4}\log K + \epsilon_\infty + \log\frac{4}{5} - 0.3858} \tag{37}$$

$$\leq 4 \tag{38}$$

where the second step is because $H_{s_c} \geq \log s_c + \epsilon_\infty \geq \log(\sqrt{K}-1) + \epsilon_\infty \geq \log(\frac{5}{6}\sqrt{K}) + \epsilon_\infty = \frac{1}{2}\log K + \epsilon_\infty + \log\frac{5}{6}$,

and where (37) holds because $H_K \leq \log K + \epsilon_{25}$ since $K \geq 25$.

As stated, the rest of the proof can be found in [37].

V. Conclusions

Our work considered the wireless MISO BC in the presence of two powerful but scarce resources: feedback to the transmitters, and caches at the receivers. Motivated by realistic expectations that cache size — at wireless receivers/end-users — will be small ([2]), and motivated by the well known limitations in getting good-quality and timely feedback, the work combines these scarce and highly connected resources, to conclude that we can attribute non-trivial performance gains even if the caches are with vanishingly small $\gamma \to 0$. This synergy between feedback and caching, allows for a serious consideration of scenarios where even microscopic fractions of the library can be placed at different caches across the network, better facilitating the coexistence of modestly-sized caches and large libraries.

The additional case for $\alpha > 0, \Gamma < 1$ is handled in the extended version in [37], while the case of $\alpha > 0, \Gamma \geq 1$ (but still with $\gamma \ll 1$) can be found in [3].

References

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

[2] S.-E. Elayoubi and J. Roberts, "Performance and cost effectiveness of caching in mobile access networks," in *Proc. of the 2nd International Conference on Information-Centric Networking*, 2015.

[3] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback," *CoRR*, vol. abs/1511.03961, 2015. [Online]. Available: http://arxiv.org/abs/1511.03961

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

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

[6] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845 – 2866, Jun. 2010.

[7] J. Zhang and P. Elia, "The synergistic gains of coded caching and delayed feedback," *CoRR*, vol. abs/1511.03961, 2016. [Online]. Available: http://arxiv.org/abs/1511.03961

[8] M. A. Maddah-Ali and D. N. C. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418 – 4431, Jul. 2012.

[9] J. Chen and P. Elia, "Degrees-of-freedom region of the MISO broadcast channel with general mixed-CSIT," in *Proc. Information Theory and Applications Workshop (ITA)*, Feb. 2013.

[10] P. de Kerret, X. Yi, and D. Gesbert, "On the degrees of freedom of the K-user time correlated broadcast channel with delayed CSIT," Jan. 2013, available on arXiv:1301.2138.

[11] J. Chen, S. Yang, and P. Elia, "On the fundamental feedback-vs-performance tradeoff over the MISO-BC with imperfect and delayed CSIT," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2013.

[12] C. Vaze and M. Varanasi, "The degree-of-freedom regions of MIMO broadcast, interference, and cognitive radio channels with no CSIT," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5254 – 5374, Aug. 2012.

[13] R. Tandon, S. A. Jafar, S. Shamai, and H. V. Poor, "On the synergistic benefits of alternating CSIT for the MISO BC," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4106 – 4128, Jul. 2013.

[14] N. Lee and R. W. Heath Jr., "Not too delayed CSIT achieves the optimal degrees of freedom," in *Proc. Allerton Conf. Communication, Control and Computing*, Oct. 2012.

[15] C. Hao and B. Clerckx, "Imperfect and unmatched CSIT is still useful for the frequency correlated MISO broadcast channel," in *Proc. IEEE Int. Conf. Communications (ICC)*, Budapest, Hungary, Jun. 2013.

[16] G. Caire, N. Jindal, and S. Shamai, "On the required accuracy of transmitter channel state information in multiple antenna broadcast channels," in *Asilomar Conference on Signals, Systems and Computers*, Asilomar, CA, Nov. 2007.

[17] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," *CoRR*, vol. abs/1504.01123, 2015. [Online]. Available: http://arxiv.org/abs/1504.01123

[18] M. A. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," *CoRR*, vol. abs/1301.5848, 2013. [Online]. Available: http://arxiv.org/abs/1301.5848

[19] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order optimal coded delivery and caching: Multiple groupcast index coding," *CoRR*, vol. abs/1402.4572, 2014. [Online]. Available: http://arxiv.org/abs/1402.4572

[20] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *CoRR*, vol. abs/1501.06003, 2015. [Online]. Available: http://arxiv.org/abs/1501.06003

[21] C. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *CoRR*, vol. abs/1504.00553, 2015. [Online]. Available: http://arxiv.org/abs/1504.00553

[22] A. N., N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," *CoRR*, vol. abs/1501.02549, 2015. [Online]. Available: http://arxiv.org/abs/1501.02549

[23] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012.

[24] B. Perabathini, E. Bastug, M. Kountouris, M. Debbah, and A. Conte, "Caching at the edge: a green perspective for 5G networks," *CoRR*, vol. abs/1503.05365, 2015. [Online]. Available: http://arxiv.org/abs/1503.05365

[25] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct 2012.

[26] E. Bastug, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," *CoRR*, vol. abs/1503.05448, 2015. [Online]. Available: http://arxiv.org/abs/1503.05448

[27] J. Hachem, N. Karamchandani, and S. N. Diggavi, "Coded caching for heterogeneous wireless networks with multi-level access," *CoRR*, vol. abs/1404.6560, 2014. [Online]. Available: http://arxiv.org/abs/1404.6560

[28] J. Hachem, N. Karamchandani, and S. Diggavi, "Effect of number of users in multi-level coded caching," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Hong-Kong, China, 2015.

[29] K. Shanmugam, M. Ji, A. Tulino, J. Llorca, and A. Dimakis, "Finite length analysis of caching-aided coded multicasting," 2015, submitted to *IEEE Trans. Inform. Theory - July 2015*.

[30] M. Deghel, E. Bastug, M. Assaad, and M. Debbah, "On the benefits of edge caching for MIMO interference alignment," in *Signal Processing Advances in Wireless Communications (SPAWC)*, Jun. 2015.

[31] Z. C. Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for small buffer users," *CoRR*, vol. abs/1407.1935, 2014. [Online]. Available: http://arxiv.org/abs/1407.1935

[32] R. Timo and M. A. Wigger, "Joint cache-channel coding over erasure broadcast channels," *CoRR*, vol. abs/1505.01016, 2015. [Online]. Available: http://arxiv.org/abs/1505.01016

[33] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," *CoRR*, vol. abs/1509.02074, 2015. [Online]. Available: http://arxiv.org/abs/1509.02074

[34] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jun. 2015.

[35] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *Proc. Allerton Conf. Communication, Control and Computing*, Sep. 2015.

[36] P. de Kerret, D. Gesbert, J. Zhang, and P. Elia, "Optimal sum-DoF of the K-user MISO BC with current and delayed feedback," 2016. [Online]. Available: https://arxiv.org/abs/1604.01653

[37] J. Zhang and P. Elia, "Feedback-aided coded caching for the MISO BC with small caches," *CoRR*, vol. abs/1606.05396, 2016. [Online]. Available: http://arxiv.org/abs/1606.05396