EURECOM
Department of Mobile Communications
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-15-307

Limits of Cache-Aided Wireless BC: Interplay between Coded-Caching and CSIT Feedback

August 25th, 2015

Jingjing Zhang and Petros Elia

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {jingjing.zhang, petros.elia}@eurecom.fr

[1]EURECOM's research is partially supported by its industrial members: BMW Group Research and Technology, IABG, Monaco Telecom, Orange, Monaco Telecom, SAP, ST Microelectronics, Symantec.

Limits of Cache-Aided Wireless BC: Interplay between Coded-Caching and CSIT Feedback

Jingjing Zhang and Petros Elia

Abstract

We consider the K-user cache-aided wireless multi-antenna broadcast channel (BC) with random fading and imperfect feedback, and analyze the throughput performance as a function of feedback statistics and cache size, identifying the optimal cache-aided degrees-of-freedom (DoF) performance within a factor of 2. In our setting where a single transmitter communicates — using non-timely and imperfect-quality channel state information (CSIT) — to K independent users with pre-filled caches, the work identifies near-optimal schemes that combine data caching, folding and precoding, to efficiently utilize caching and feedback resources. Our schemes will reveal interesting connections between MAT-type schemes and caching. Interestingly in the large K setting, where the schemes are often DoF optimal, the derived limits reveal the surprising fact that full (perfect) CSIT can be completely substituted (without performance losses) by combining a vanishingly small portion of delayed CSIT, with a vanishingly small fraction of the files content per user's cache. The key lies in finding the right balance between cache-induced gains of multicasting common information, and CSIT-induced gains of broadcasting private information. It also builds on the retrospective nature shared by both coded caching and (communicating with) non-timely feedback, where in both cases the transmitter — which has timely knowledge of the information content — must act retrospectively to compensate for not knowing the 'destination' (channel and user identity) of this content.

Index Terms

Coded caching, CSIT, multiple-input single-output (MISO), CSIT gain, cache-aided degrees of freedom

Contents

1 Introduction **1**
- 1.1 Caching-aided broadcast channel model . . . 1
  - 1.1.1 K-user BC with pre-filled caching . . . 1
- 1.2 Coded caching and CSIT-type feedback . . . 1
- 1.3 Measures of performance . . . 2
- 1.4 Prior work . . . 3
- 1.5 Notation and assumptions . . . 4

2 Main results **4**
- 2.1 BC with caching - Throughput results . . . 4
- 2.2 Large BC with modest amount of caching - $K \gg 1,\ \gamma_{tot} \ll K$ . . 5
- 2.3 Translating caching gain to CSIT gain . . . 6
- 2.4 Cache-aided CSIT gains in the large K regime . . . 6
- 2.5 How much caching is needed to fully substitute CSIT . . . 7
- 2.6 Vanishing fraction of delayed CSIT . . . 8

3 Combining retrospective transmission and retrospective coded caching **10**
- 3.1 Placement phase . . . 10
- 3.2 Delivery phase: folding . . . 11
- 3.3 Delivery of folded and private information with imperfect CSIT . 12
- 3.4 Decoding . . . 15
- 3.5 Duration Calculation . . . 15

4 Bounding the performance gap **16**
- 4.1 Bounding the performance gap to optimal, for the case of $\alpha = 0$. 16
  - 4.1.1 Case 1: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\gamma \leq \frac{1}{44}$ and $K \geq 2$. 17
  - 4.1.2 Case 2: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\frac{1}{44} \leq \gamma \leq \frac{1}{4}$ and $K \geq 2$ . . . 18
  - 4.1.3 Case 3: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\frac{1}{4} \leq \gamma \leq \frac{1}{2}, K \geq 2$. 20
  - 4.1.4 Case 4: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\frac{1}{2} \leq \gamma \leq \frac{K-1}{K}, K \geq 2$ 21
- 4.2 Bounding the performance gap, for the case of $\alpha > 0$ . . . 21
  - 4.2.1 Case 1: Proving that $\frac{T(\alpha>0)}{T^*(\alpha>0)} \leq 2$ for $\frac{1}{4} \leq \gamma \leq \frac{K-1}{K}, \forall K$. 22
  - 4.2.2 Case 2: Proving that $\frac{T(\alpha>0)}{T^*(\alpha>0)} \leq 2$ for $0 \leq \gamma \leq \frac{1}{4}, \forall K$ . . . 22

5 Appendix - Lower bound on T^* **23**

6 Appendix - Proof of the asymptotic optimality **23**

7 Appendix - Additional proofs **24**
- 7.1 Proof of Lemma 2 . . . 24
- 7.2 Proof of vanishing fraction of delayed CSIT cost due to caching from Section 2.6 . . . 25

1 Introduction

Our interest here is to explore the idea of coded caching (cf. [1]) in the feedback-aided multi-antenna wireless BC.

1.1 Caching-aided broadcast channel model

1.1.1 K-user BC with pre-filled caching

In the K-user multiple-input single-output (MISO) broadcast channel (BC) of interest here, the K-antenna transmitter, communicates to K single-antenna receiving users. At the transmitter, there is a total of $N \geq K$ distinct files $W_1, W_2, \ldots, W_N$, each of size $|W_i| = f$ bits. Each user $k \in \{1, 2, \ldots, K\}$ has a cache Z_k, of size $|Z_k| = Mf$ bits, where naturally $M \leq N$. Communication consists of two distinct phases; the content *placement phase* and the *delivery phase*. During the placement phase — which usually corresponds to communication during off-peak hours — the caches $Z_1, Z_2, \ldots, Z_K$ are pre-filled with content from the N files $\{W_i\}_{i=1}^N$. The delivery phase commences once each user k requests from the transmitter, any *one* file $W_{R_k} \in \{W_i\}_{i=1}^N$, out of the N available 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.

For each transmission, the received signals at each user k, will be modeled as

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

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 user k.

1.2 Coded caching and CSIT-type feedback

CSIT is typically of imperfect-quality as it is hard to obtain in a timely and reliable manner. In the high-SNR (high P) setting, this current-CSIT quality is concisely represented in the form of the normalized quality exponent [2] [3]

$$\alpha := -\lim_{P \to \infty} \frac{\log \mathbb{E}[||\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k||^2]}{\log P}, \; k \in \{1, \ldots, K\} \tag{2}$$

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 $\boldsymbol{h}_k$. The range of interest is $\alpha \in [0, 1]$ (cf. [4]). We also assume availability of delayed CSIT ([5]), where now the delayed estimates of any channel, can be received without error but with arbitrary delay.

In normalizing the caching resources, described by M, we will consider

$$\gamma := \frac{M}{N} \tag{3}$$

as well as the cumulative cache size

$$\gamma_{tot} := \frac{KM}{N} = K\gamma. \tag{4}$$

1.3 Measures of performance

The general objective here is to identify caching and transmission schemes that jointly reduce T, under specific constraints on caching size Mf, on N, and under specific constraints on the CSIT-quality α. Specifically, as in [1], the measure of performance here is the duration T — in time slots, per file served per user — needed to complete the delivery process, for any request. The link capabilities, and the time scale, are normalized such that one time slot corresponds to the optimal amount of time it would take to communicate a single file to a single receiver, had their been no caching and no interference. As a result, in the high P setting of interest — where the capacity of a single-user MISO channel scales as $\log P$ — we proceed to set

$$f = \log P \tag{5}$$

which guarantees that the two measures of performance, here and in [1], are the same and can thus be directly compared[1].

A simple inversion would lead to an equivalent measure of the *cache-aided degrees of freedom* (cache-aided sum DoF)

$$d(\gamma, \alpha) = \frac{K}{T} \tag{6}$$

which is a measure of cumulative throughput.

To insightfully measure this synergistic effect of coded-caching and CSIT in removing interference, we consider

$$\alpha_j(\gamma, \alpha) := \arg\min_{\alpha'} \{\alpha' : (1-\gamma)T^*(\gamma = 0, \alpha') \leq T(\gamma, \alpha)\} \tag{7}$$

to reflect the boosted CSIT quality α_j that would have been needed to reach the achieved $T(\gamma, \alpha)$, had there been no coded-caching gain. As K increases, this measure will increasingly reflect the fraction of interference jointly removed by CSIT and coded-caching.

[1] We note that setting $f = \log P$ is simply a normalization of choice, and does not carry a 'forced' relationship between SNR and file sizes. The essence of the derived results would remain the same for any other non-trivial normalization.

1.4 Prior work

Various works, not considering caching, have sought to understand the effect of feedback on the performance of the networks. This is a massive and active area of research, which incorporates many facets and considers many settings of wireless communications. Focusing here just on the broadcast channel and just on the setting of imperfect and delayed CSIT feedback, a small — and certainly incomplete — samples of interesting works includes the work by Maddah-Ali and Tse [5] who considered a simple crisp setting of fast random fading that completely obsolete CSIT is in fact useful, this is achieved by designing ways to utilize the interference that was created at the transmitter as a result of this CSIT delay. In addition to CSIT delays, modern wireless communications must consider CSIT that is of imperfect quality, i.e., where the transmitter has channel estimates that come with estimation errors. Different works, such as in [2,3,6–9], have dealt with this imperfect-quality issue. Specifically, in the BC with perfect delayed CSIT and current CSIT with quality exponent α, the work by Chen et. al [9] showed that the inner bound of the sum DoF is $\frac{K}{H_K}(1-\alpha)+K\alpha$, which matches the outer bound from the work of de Kerret et. al [8]. Other related work can be found in [4,10–16].

On the other hand, the benefits of caching on reducing the load on the networks, came with the work by Maddah-Ali and Niesen in [1] who considered a caching system where a server is connected to multiple users through a shared, error-free link. A novel coded caching approach applied offers a multicast gain by designing carefully the pre-filled caching content at the receivers to mitigate the load of the link. Then, they extended their work to decentralized caching in [17] which achieved a performance close to that of [1] despite the lack of coordination of the content placement. Another work by Ji et al. in [18] considered a combination caching network in which a source is connected to multiple user nodes through a layer of relay nodes, such that each user node with caching is connected to a distinct subset of the relay nodes, and the fundamental limits of this setting is analyzed. Ghasem et al. mainly focused on the tighter lower bounds on coded caching in [19].

In addition, prior works have also employed caching to different wireless networks without utilizing CSIT. The work by Maddah-Ali and Niesen in [20] studied an interference channel in which each transmitter is equipped with a local cache, and it shows that three distinct benefits can be obtained from caching, in the sense that content overlap in the transmitter allows effective interference cancellation. Another work of Timo and Wigger [21] indicated that the cache-aided system efficiency was improved by employing unequal cache sizes at the receivers as they experience different channel qualities over an erasure broadcast channel. Niesen et al. in [22] derived inner and outer bounds on caching capacity region of a wireless caching network, where the nodes randomly located on a square and each node in the network requests a message available at a set of caches. Other related work on caching can be found in [23–25].

1.5 Notation and assumptions

We will 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 $[K] \stackrel{\Delta}{=} \{1, 2, \cdots, K\}$. If ψ is a set, then $|\psi|$ will denote its cardinality. 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 dur($\boldsymbol{x}$) to denote the transmission duration of that vector. For example, saying dur($\boldsymbol{x}$) $= \frac{1}{10}T$ simply means that $\boldsymbol{x}$ uses one tenth of the delivery phase.

2 Main results

We first lower bound T.

Lemma 1 *The optimal T^* for the (K, M, N, α) cache-aided K-user MISO BC, is lower bounded as*

$$T^* \geq \max_{s\in\{1,\ldots,min\{\lfloor \frac{N}{M} \rfloor, K\}\}} \frac{s}{d(\gamma = 0, \alpha)}(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}) \tag{8}$$

$$\geq \max_{s\in\{1,\ldots,min\{\lfloor \frac{N}{M} \rfloor, K\}\}} (\frac{1-\alpha}{H_s} + \alpha)^{-1}(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}) \tag{9}$$

where $d(\gamma = 0, \alpha)$ is the optimal sum-DoF for the corresponding K-user $s \times 1$ MISO BC.

Proof: The proof is presented in Section 5 and it uses that $d^*(\gamma = 0, \alpha = 0) = \frac{s}{H_s}$ from [5], and that $d^*(\gamma = 0, \alpha) = \frac{s}{H_s}(1 - \alpha) + s\alpha$, where the outer bound is from [26] and the matching inner bound from [27]. □

2.1 BC with caching - Throughput results

We first consider the case where $\gamma_{tot} \in \{1, 2, \cdots, K\}$, i.e., where $M \in \frac{N}{K}\{0, 1, \cdots, K\}$. We remind the reader that $\gamma = \frac{M}{N}$ and $\gamma_{tot} = K\gamma$.

Theorem 1 *In the (K, M, N, α) cache-aided MISO BC with N files and $K \leq N$ users each with cache of size $M \in \frac{N}{K}\{0, 1, \cdots, K\}$,*

$$T = \frac{(1-\gamma)(H_K - H_{\gamma_{tot}})}{\alpha(H_K - H_{\gamma_{tot}}) + (1-\alpha)(1-\gamma)} \tag{10}$$

is achievable and has a gap from optimal

$$\frac{T}{T^*} < 2 \tag{11}$$

that is at most 2, for all $\alpha, K, \gamma_{tot} \in \{1, 2, \cdots, K\}$.

Proof: The scheme that achieves the above performance is presented in Section 3, while the corresponding gap to optimal is bounded in Section 4. □
We also have the following, under the logarithmic approximation $H_n \approx \log(n)$, for the case of $\alpha = 0$.

Corollary 1a *In the* $(K, M, N, \alpha = 0)$ *cache-aided MISO BC without current CSIT,*

$$T = H_K - H_{\gamma_{tot}} \tag{12}$$

is achievable and has a gap from optimal that is at most 2. Under the logarithmic approximation, or in the large K *regime, the above* T *takes the form*

$$T = \log(\frac{1}{\gamma}). \tag{13}$$

2.2 Large BC with modest amount of caching - $K \gg 1$, $\gamma_{tot} \ll K$

The following states that the derived T from Corollary 1a is asymptotically optimal for K large.

Theorem 2 *In the* (K, M, N, α) *cache-aided MISO BC, in the limit of asymptotically large* K *and with reduced caching size* $M << N$*, the achievable* T *from Corollary 1a is asymptotically optimal, and satisfies*

$$\lim_{K\to\infty} \frac{T}{T^*} = 1, \ \forall \alpha. \tag{14}$$

Proof: The proof is found in Section 6, which calculates the gap between T (from Corollary 1a) to a properly minimized outer bound that derives from Lemma 1. □

2.3 Translating caching gain to CSIT gain

The following calculates the achievable synergistic $\alpha_j(\gamma, \alpha)$, which nicely takes the form $\alpha_j(\gamma, \alpha) = \alpha + \delta_\alpha(\gamma, \alpha)$ for some $\delta_\alpha(\gamma, \alpha)$ that can be termed as the *CSIT gain due to caching*.

Corollary 2a *In the (K, M, N, α) cache-aided MISO BC, then*

$$\alpha_j(\gamma, \alpha) = \alpha + \frac{(1-\alpha)(H_{K\gamma} - \gamma H_K)}{(H_K - 1)(H_K - H_{K\gamma})} \tag{15}$$

is achievable.

Proof: The proof is direct from Theorem 1, and then from Lemma 1. □

For the special case when $\alpha = 0$, we have the following.

Corollary 2b *In the $(K, M, N, \alpha = 0)$ cache-aided BC with $\alpha = 0$, then*

$$\alpha_j(\gamma, \alpha = 0) = \frac{H_{\gamma_{tot}} - \gamma H_K}{(H_K - H_{\gamma_{tot}})(H_K - 1)} \tag{16}$$

is achievable, and the associated gain $\delta_\alpha(\gamma, \alpha)$ has a gap to the optimal gain, that is at most 2.

Proof: The proof is direct from Corollary 2b. □

2.4 Cache-aided CSIT gains in the large K regime

The following insightfully differentiates between the fraction of interference removed by CSIT, and that which is removed by coded caching.

Corollary 2c *In the large K regime, the fraction of interference synergistically removed by CSIT and coded caching*

$$\alpha_j(\gamma, \alpha) = \alpha + (1-\alpha)\frac{1-\gamma}{\log(\frac{1}{\gamma})} \tag{17}$$

is achievable and at least half the optimal.

Proof: The proof is direct from the definition of $\alpha_j(\gamma, \alpha)$ and from Lemma 1, Theorem 1 and Theorem 2. $\square$

The following holds directly from the above.

Corollary 2d *In the large K regime, for $\alpha = 0$,*

$$\alpha_j(\gamma, \alpha = 0) = \frac{1-\gamma}{\log(\frac{1}{\gamma})} \tag{18}$$

is achievable, and has a gap to optimal that is at most 2.

2.5 How much caching is needed to fully substitute CSIT

Let us now put back into the picture the local caching gains, and explore the amount of caching

$$\gamma_{if} := \arg\min_{\gamma'} \{\gamma' : T(\gamma', \alpha = 0) \leq T^*(\gamma = 0, \alpha = 1)\} \tag{19}$$

needed to fully substitute CSIT, i.e., explore how much caching is needed for a system with $\alpha = 0$, to achieve the interference free performance $T^*(\gamma = 0, \alpha = 1) = 1$. The latter expression derives from the well known optimal sum-DoF $d^*(\gamma = 0, \alpha = 1) = K$ of the K-user BC with perfect CSIT. Towards this, we have the following result.

Corollary 2e *In the $(K, M, N, \alpha = 0)$ cache-aided BC with $\alpha = 0$, to achieve perfect CSIT performance $T^* = 1$, (and thus to fully substitute perfect CSIT with caching), it is sufficient to have*

$$\gamma \geq \gamma_{if} = e^{-(1-\epsilon_K+\epsilon_\infty)}. \tag{20}$$

As K increases, the above converges to

$$\gamma_{if} = e^{-1}.$$

Proof: First recall that $T(\gamma, \alpha = 0) = H_K - H_{K\gamma}$, then set $H_K - H_{K\gamma} = 1$, and then recall that $\log(\frac{1}{\gamma}) < H_K - H_{K\gamma} \leq \log(\frac{1}{\gamma}) + \epsilon_K - \epsilon_\infty$. $\square$ Similarly we can consider

$$\gamma_{if,G} := \arg\min_{\gamma'} \{\gamma' : T(\gamma', \alpha = 0) \leq G \cdot T^*(\gamma = 0, \alpha = 1)\} \tag{21}$$

to describe the γ needed to achieve a certain gap $G \geq 1$ from perfect-CSIT optimal performance, i.e., to achieve a performance $G\dot{T}^*(\gamma = 0, \alpha = 1) = G$, and do so with $\alpha = 0$. The following is a small generalization of Corollary 2e.

Corollary 2f *In the $(K, M, N, \alpha = 0)$ cache-aided BC with $\alpha = 0$, to achieve a gap of G to perfect CSIT performance, i.e., to achieve $T = G$, it is sufficient to have*

$$\gamma \geq \gamma_{if,G} = e^{-(G-\epsilon_K+\epsilon_\infty)}. \tag{22}$$

As K increases, the above converges to

$$\gamma_{if} = e^{-G}.$$

Proof: The proof is direct, after setting $H_K - H_{K\gamma} = G$. □

It is interesting to see that in a system with $\alpha = 0$, a reasonable amount of caching can substantially bridge the gap to optimal, irrespective of K. Had there been no caching, this gap, would have otherwise been increasing with K, approximately as $\log(K)$.

2.6 Vanishing fraction of delayed CSIT

In the following we will briefly describe how caching allows for a large reduction in the cost of supporting communications with delayed CSIT. This will be done here for the case where $\alpha = 0$. The analytical details will be presented in the appendix of Section 7.2.

The MAT-inspired schemes that we use and describe in Section 3, can have up to K phases of decreasing time duration and of decreasing cost of communicating CSIT. What we will see is that caching will allow us to bypass the first γ_{tot} phases, which are the longest and most intensive, leaving us with the remaining $K - \gamma_{tot}$ communication phases that are easier to support with delayed feedback because they involve fewer transmissions, with fewer transmit antennas and to fewer users, and thus involving fewer CSIT scalars that must be communicated.

In brief — after normalization to account for the condition that each user receives a total of $\log P$ bits — each phase $j = \gamma_{tot} + 1, \gamma_{tot} + 2, \ldots, K$ will have a *normalized* duration $T_j = \frac{1}{j}$. During each phase j, we will need to send CSIT that describes the channel vectors for $K - j$ users, and during this same phase the transmitted vectors will have support $K - j + 1$ because only $K - j + 1$ transmit antennas will need to be active. Thus during phase j, there will be a need to send $T_j(K - j + 1)(K - j) = \frac{1}{j}(K - j + 1)(K - j)$ CSIT scalars, and thus a need to send CSIT on a total of

$$\begin{aligned} L(\gamma_{tot}) &= \sum_{j=\gamma_{tot}+1}^{K} \frac{1}{j}(K - j + 1)(K - j) \\ &= (K^2 + K)(H_K - H_{\gamma_{tot}}) - \frac{K(1-\gamma)(3K - K\gamma - 1)}{2} \end{aligned} \tag{23}$$

channel scalars, while in the absence of caching (corresponding to $\gamma_{tot} = 0$), we will have to send CSIT on

$$L(\gamma_{tot} = 0) = \sum_{j=1}^{K} \frac{1}{j}(K - j + 1)(K - j) \tag{24}$$

$$= (K^2 + K)H_K - \frac{3K^2}{2} + \frac{K}{2} \tag{25}$$

channel scalars.

To reflect the frequency of having to gather CSIT, and to provide a fair comparison between different schemes of different performance that manage to convey different amounts of actual data to the users, we consider the measure $Q(\gamma_{tot})$ that normalizes the above number $L(\gamma_{tot})$ of full CSIT scalars, by the coherence period T_c and by the total number of full data symbols sent (recall that $\alpha = 0$). In our case, under the assumption that each user receives a total of $\log P$ bits, the total number of full data symbols sent is K, and thus we have

$$Q(\gamma_{tot}) = \frac{L(\gamma_{tot})}{T_c K} \tag{26}$$

$$= \frac{(K^2 + K)(H_K - H_{\gamma_{tot}}) - \frac{K(1-\gamma)(3K - K\gamma - 1)}{2}}{T_c K} \tag{27}$$

which means that without caching, we have

$$Q(\gamma_{tot} = 0) = \frac{L(\gamma_{tot})}{T_c K} = \frac{(K + 1)H_K - \frac{3}{2}K + \frac{1}{2}}{T_c}. \tag{28}$$

Consequently we see that in the large K limit,

$$Q(\gamma_{tot}) \to \frac{K\big(\log(\frac{1}{\gamma}) - \frac{3}{2} + 2\gamma - 2\gamma^2\big)}{T_c} \tag{29}$$

while in the absence of caching

$$Q(\gamma_{tot} = 0) \to \frac{1}{T_c} K \log(K) \tag{30}$$

which implies that

$$\lim_{K \to \infty} \frac{Q(\gamma_{tot})}{Q(\gamma_{tot} = 0)} = 0 \tag{31}$$

which in turn tells us that caching allows for a substantial reduction (down to a vanishingly small portion) of the cost of delayed CSIT.

This reduction is important because retrospective delayed-feedback methods suffer from an increased cost of supporting their CSIT requirements (albeit at the benefit of allowing substantial delays in the feedback mechanisms). To quickly

see this, just consider that in the presence of perfect CSIT and zero forcing (no caching), the same cost is

$$Q_{ZF} = \frac{K^2}{T_c K} = \frac{K}{T_c}$$

which means that

$$\lim_{K\to\infty} \frac{Q(\gamma_{tot}=0)}{Q_{ZF}} = \infty \tag{32}$$

which in turn implies that the increase in the cost of supporting the CSIT requirements for retrospective delayed-feedback methods, can be unbounded compared to non-retrospective methods. On the other hand, we see that

$$\lim_{K\to\infty} \frac{Q(\gamma_{tot})}{Q_{ZF}} = \log(\frac{1}{\gamma} - \frac{3}{2} + 2\gamma - 2\gamma^2) \tag{33}$$

which means that

$$\lim_{K\to\infty} \frac{Q(\gamma_{tot})}{Q_{ZF}} < 1,\ \gamma \geq \frac{1}{10}.$$

One interesting conclusion that comes out of this, is that caching can allow for full substitution of current CSIT (as we have seen in Section 2.5), with a very substantial reduction of the cost of delayed CSIT as well.

3 Combining retrospective transmission and retrospective coded caching

Our schemes will reveal interesting connections between MAT-type schemes and caching. The caching algorithm (which generally draws from [1]) can be seen as essentially *'folding'* the different users' data into multi-layered blocks, while the delivery algorithm (which includes a close variant of the last $K - \gamma_{tot}$ levels of MAT) simply unfolds these - or more correctly, it delivers these layers in a manner that allows the receivers to unfold them. Another way to visualize the interesting match of the caching and delivery effort, is to note that the caching algorithm creates the same multi-destination delivery problem, as do the first γ_{tot} levels of the MAT algorithm. Hence, both of these problems are resolved by the delivery algorithm here, which is a close variant of the $K - \gamma_{tot}$ levels of the K-user MAT, and which, fortunately, requires a much reduced delayed-CSIT load than the full K-user MAT, simply because the transmitted vectors have smaller support (only a few antennas transmit), and because — in the presence of caching — they are much more efficient.

3.1 Placement phase

Each of the N files $W_n, n = 1, 2, \ldots, N$ in the library, is split into $\binom{K}{\gamma_{tot}}$ *disjoint* subfiles $W_{n,\tau}, \tau \in \Psi_{\gamma_{tot}}$, where

$$\Psi_{\gamma_{tot}} := \{\tau \subset [K],\ |\tau| = \gamma_{tot}\} \tag{34}$$

and where each cache Z_k is filled as

$$Z_k = \{W_{n,\tau}\}_{n\in[N],\tau\in\Psi^{(k)}_{\gamma_{tot}}} \tag{35}$$

where

$$\Psi^{(k)}_{\gamma_{tot}} := \{\tau \subset [K] \,:\, |\tau| = \gamma_{tot}, \tau \ni k\}. \tag{36}$$

This means that each subfile $W_{n,\tau}$ will be placed in cache Z_k of user k, if and only if $k \in \tau$. This in turn means that each subfile $W_{n,\tau}$ will appear in γ_{tot} different caches. We also recall that since each W_n is of size $|W_n| = f$ bits, and since the subfiles $W_{n,\tau}$ are disjoint, then each subfile $W_{n,\tau}$ is of size $|W_{n,\tau}| = f/\binom{K}{\gamma_{tot}}$ bits.

3.2 Delivery phase: folding

We recall that at this point, the transmitter becomes aware of the file requests $R_k, k = 1,\dots,K$, and thus must deliver each requested file W_{R_k}, by delivering the constituent subfiles $\{W_{R_k,\tau}\}_{\tau\in\Psi_{\gamma_{tot}}}$, to the corresponding receiver k. Let us recall that

1. subfiles $\{W_{R_k,\tau}\}_{\tau\in\Psi^{(k)}_{\gamma_{tot}}}$ are already in Z_k;
2. subfiles $\{W_{R_k,\tau}\}_{\tau\in\Psi_{\gamma_{tot}}\backslash\Psi^{(k)}_{\gamma_{tot}}}$ are directly requested by user k, but are not already available inside Z_k;
3. subfiles $Z_k\backslash\{W_{R_k,\tau}\}_{\tau\in\Psi^{(k)}_{\gamma_{tot}}} = Z_k\backslash W_{R_k}$ are cached inside Z_k, are not directly requested by user k, but will be useful in removing interference.

The folding part corresponds to creating linear combinations (XORs) of different subfiles. Assume that we are trying to deliver a subfile $W_{R_k,\tau} \notin Z_k$ (note that $k \notin \tau$) to user k. Let $P_{k,k'}(\tau)$ be the function that replaces the entry $k' \in \tau$, with the entry k. Then we see that if we deliver

$$W_{R_k,\tau} \oplus (\oplus_{k'\in\tau} \underbrace{W_{R'_k,P_{k,k'}(\tau)}}_{\in Z_k}) \tag{37}$$

the fact that $W_{R'_k,P_{k,k'}(\tau)} \in Z_k$, guarantees that receiver k can recover $W_{R_k,\tau}$, while at the same time similarly guarantees that each other user $k' \in \tau$ can recover its own desired subfile $W_{R'_k,P_{k,k'}(\tau)} \notin Z_{k'}, \forall k' \in \tau$. Hence delivery of $W_{R_k,\tau} \oplus (\oplus_{k'\in\tau} W_{R'_k,P_{k,k'}(\tau)})$ which has size $|W_{R_k,\tau} \oplus (\oplus_{k'\in\tau} W_{R'_k,P_{k,k'}(\tau)})| = f/\binom{K}{\gamma_{tot}}$ bits, automatically guarantees delivery of $W_{R'_k,P_{k,k'}(\tau)}$ to each user $k' \in \tau$, i.e., delivers a total of $\gamma_{tot}+1$ distinct subfiles (each of size $|W_{R'_k,P_{k,k'}(\tau)}| = f/\binom{K}{\gamma_{tot}}$ bits) to $\gamma_{tot}+1$ distinct users. Hence any

$$X_\psi = \oplus_{k\in\psi} W_{R_k,\psi\backslash\{k\}}, \psi \in \Psi_{\gamma_{tot}+1}$$

— referred to here as an *order-($\gamma_{tot}+1$) folded file* — can similarly deliver to user $k \in \psi$, her requested file $W_{R_k,\psi\backslash k}$, i.e., each order-($\gamma_{tot}+1$) folded file X_ψ can deliver — with the assistance of the side information in the caches — a distinct, individually requested subfile, to each of the $\gamma_{tot}+1$ users in ψ. Hence to satisfy all requests $\{W_{R_k}\backslash Z_k\}_{k=1}^K$, one must deliver the following set

$$\mathcal{X}_\Psi = \{X_\psi = \oplus_{k\in\psi} W_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\gamma_{tot}+1}} \tag{38}$$

consisting of $|\mathcal{X}_\Psi| = \binom{K}{\gamma_{tot}+1}$ folded messages of order-($\gamma_{tot}+1$), where each folded message has size

$$|X_\psi| = f/\binom{K}{\gamma_{tot}} \text{ (bits)}. \tag{39}$$

3.3 Delivery of folded and private information with imperfect CSIT

The challenge here is to find an efficient method to deliver all the common (order-$\gamma_{tot}+1$) messages from $\mathcal{X}_\Psi$.

The delivery phase draws from the last $K-\gamma_{tot}$ phases of the MAT algorithm in [5], where each phase $j, j \in [\gamma_{tot}+1, K]\cap\mathbb{Z}$ aims to deliver order-j folded messages (cf. (38)) with the help of delayed CSIT. It is shown that the phase j takes $(K-j+1)\binom{K}{j}$ common symbols of order j, and creates $j\binom{K}{j+1}$ of order $j+1$. We use N_j to denote the number of order-j messages and T_j to denote the duration of phase j, and the entire duration is $T = \sum_{j=1}^{K-\gamma_{tot}} T_j$.

During phase j, the transmitter sends

$$\boldsymbol{x}_t = \boldsymbol{x}_{c,t} + \boldsymbol{g}_{1,\eta}a_{1,\eta} + \cdots + \boldsymbol{g}_{k,t}a_{k,t} + \cdots + \boldsymbol{g}_{K,t}a_{K,t} \tag{40}$$

for any time instant $t \in [\sum_{i=1}^{j-1} T_i, \sum_{i=1}^{j} T_i]$, where $\boldsymbol{g}_{k,t}$ are the precoders that are designed to be orthogonal to the channel estimates of all users k' other than k satisfying

$$\hat{\boldsymbol{h}}_{k',t}^T \boldsymbol{g}_{k,t} = 0, \forall k' \notin [K]\backslash\{k\} \tag{41}$$

We use the notation

$$P_t^{(c)} \triangleq \mathbb{E}||\boldsymbol{x}_{c,t}||^2, P_t^{(a_k)} \triangleq \mathbb{E}|a_{k,t}|^2 \tag{42}$$

to denote the power of $\boldsymbol{x}_{c,t}$ and $a_{k,t}$ respectively, and we use $r_t^{(c)}$ and $r_t^{(a_k)}$ to denote the rate of $\boldsymbol{x}_{c,t}$ and $a_{k,t}$ at time t. So the power and rate are formulated as

$$\begin{aligned} &P_t^{(c)} \doteq P,\ P_t^{(a_k)} \doteq P^\alpha, \\ &r_t^{(c)} = (1-\alpha)f,\ r_t^{(a_k)} = \alpha f, k = 1,\ldots,K. \end{aligned} \tag{43}$$

Hence, during phase j, $\{\boldsymbol{x}_{c,t}\}_{\forall t}$ and $\{a_{k,t}\}_{\forall t}$ can deliver $(1-\alpha)fT_j$ bits and $\alpha f T_j$ bits respectively, where $t \in [\sum_{i=1}^{j-1} T_i, \sum_{i=1}^{j} T_i]$.

To convey $\{X_\psi\}_{\psi\in\Psi_{\gamma_{tot}}}$, we split each X_ψ into two parts as follows

$$X_\psi = (X_\psi^{(p)}, X_\psi^{(c)}) = (\oplus_{k\in\psi} W^{(p)}_{R_k,\psi\backslash\{k\}}, \oplus_{k\in\psi} W^{(c)}_{R_k,\psi\backslash\{k\}}).$$

The delivery of the first part is converted into that of private messages, i.e., elements $\{W^{(p)}_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\gamma_{tot}+1}}$ — private information for user k — are delivered by $\{a_{k,t}\}_{t=0}^{T}$ through $K-\gamma_{tot}$ phases, The second part $\{X_\psi^{(c)}\}_{\psi\in\Psi_{\gamma_{tot}}}$ is handled by $\boldsymbol{x}_c$.

For the first part, each $\{a_{k,t}\}_{t=0}^{T}$ carries equal size information of the desired $\binom{K-1}{\gamma_{tot}}$ subfiles $\{W^{(p)}_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\gamma_{tot}+1}}$. In this way, within duration T, each

$$X_\psi^{(p)} = \oplus_{k\in\psi} W^{(p)}_{R_k,\psi\backslash\{k\}}$$

can be conveyed by $\frac{f\alpha T}{\binom{K-1}{\gamma_{tot}}}$ bits.

Now we focus on the second part. $\{\boldsymbol{x}_{c,t}\}_{t=0}^{T}$ takes place over $K-\gamma_{tot}$ phases. Phase j delivers order-$(\gamma_{tot}+j)$ symbols— each one of them useful for a subset of $\gamma_{tot}+j$ users, and generates order-$(\gamma_{tot}+j+1)$ symbols. During phase $K-\gamma_{tot}$, the transmitter sends the fully common information intended for all receivers and no more symbols are generated.

phase 1: The information of $\{X_\psi\}_{\psi\in\Psi_{\gamma_{tot}+1}}$ are delivered by $\{\boldsymbol{x}_{c,t}\}_{t=0}^{T_1}$, describing a sequential time-sharing transmission of $\{\boldsymbol{x}_\psi\}_{\forall\psi}$. Each

$$\boldsymbol{x}_\psi = [x_{\psi,1}, \ldots, x_{\psi,K-\gamma_{tot}}, 0, \ldots, 0]^T \tag{44}$$

maps the content of $X_\psi^{(c)}$ with $\theta = \frac{f}{\binom{K}{\gamma_{tot}}} - \frac{f\alpha T}{\binom{K-1}{\gamma_{tot}}}$ bits, where $\{x_{\psi,i}\}_{i=1}^{K-\gamma_{tot}}$ are independent scalars. Thus each $x_{\psi,i}$ carries equal size of $\frac{\theta}{K-\gamma_{tot}}$ bits. Hence the duration for carrying each $\boldsymbol{x}_\psi$ is $\text{dur}(\boldsymbol{x}_\psi) = \frac{\theta}{(K-\gamma_{tot})(1-\alpha)f}$, and consequently,

$$T_1 = \binom{K}{\gamma_{tot}+1}\text{dur}(\boldsymbol{x}_\psi) = \frac{\binom{K}{\gamma_{tot}+1}\theta}{(K-\gamma_{tot})(1-\alpha)f} \tag{45}$$

After overloading $\{X_\psi\}_{\psi\in\Psi_{\gamma_{tot}+1}}$, each user k receives one observation of $K-\gamma_{tot}$ symbols $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\gamma_{tot}}$, which is denoted by $L_{\psi,k}$ with size $|L_{\psi,k}| = \frac{\theta}{K-\gamma_{tot}}$ bits. For any receiver in ψ, if the transmitter can somehow send $K-\gamma_{tot}-1$ overheard equations received by all the receivers in $[K]\backslash\psi$ with the aid of delayed CSIT, then receiver k can solve the elements $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\gamma_{tot}}$. In other words, each overheard equation $L_{\psi,k}$ is simultaneously useful for all the receivers in ψ and will be somehow sent in phase 2.

phase 2: We proceed to see how the overheard equations from phase 1 are delivered during this phase. In the presence of delayed CSIT, these overheard equations are available to the transmitter after phase 1. Note that

$$\Psi_{\gamma_{tot}+2} = \{\psi \in [K], |\psi| = \gamma_{tot} + 2\} \tag{46}$$

For each ψ, $\boldsymbol{x}_\psi = [x_{\psi,1}, \ldots, x_{\psi,K-\gamma_{tot}-1}, 0, \ldots, 0]^T$ maps the content of

$$f_i(\{L_{\psi\setminus\{k\},k}\}_{k\in\psi}), i = 1, \ldots, \gamma_{tot} + 1$$

, where f_i is a random linear combination of the $\gamma_{tot} + 2$ elements $\{L_{\psi\setminus\{k\},k}\}_{\forall k\in\psi}$ created by the transmitter and the coefficients are shared with the receivers. The transmission of the sequences $\{\boldsymbol{x}_\psi\}_{\forall\psi}$, described by $\{\boldsymbol{x}_{c,t}\}_{t=T_1}^{T_1+T_2}$, takes place in phase 2 using time-sharing. Each $x_{\psi,i}, i = 1, \ldots, K-\gamma_{tot}-1$ carries $\frac{|L_{\psi,k}|(\gamma_{tot}+1)}{K-\gamma_{tot}-1}$ bits and thus $\text{dur}(\boldsymbol{x}_\psi) = \frac{|L_{\psi,k}|(\gamma_{tot}+1)}{(K-\gamma_{tot}-1)(1-\alpha)f}$. We can get duration T_2

$$T_2 = \binom{K}{\gamma_{tot}+2}\text{dur}(\boldsymbol{x}_\psi) = T_1\frac{\gamma_{tot}+1}{\gamma_{tot}+2} \tag{47}$$

Now, we focus on how the transmission allows the solution of the overheard equations from phase 1. Consider each ψ, receiver $k \in \psi$ is able to remove the known $L_{\psi\setminus\{k\},k}$ from $f_i(\{L_{\psi\setminus\{k\},k}\}_{\forall k\in\psi}), i = 1, \ldots, \gamma_{tot} + 1$ since it is a received signal in phase 1 and obtains $\gamma_{tot} + 1$ independent linear combinations of $\{L_{\psi\setminus\{k'\},k'}\}_{\forall k'\in\psi\setminus\{k\}}$, which can be solved easily. Each other user $k' \in \psi$ acts the same and acquires the desired $\gamma_{tot} + 1$ observations. It is easy to see that phase 1 creates $(\gamma_{tot} + 1)\binom{K}{\gamma_{tot}+2}$ symbols of the form $f_i(\{L_{\psi\setminus\{k\},k}\}_{k\in\psi}), \forall i, \forall\psi$, each is an order-$(\gamma_{tot} + 2)$ aimed for $\gamma_{tot} + 2$ receivers in ψ.

After phase 2, we use $L_{\psi,k}, \psi \in \Psi_{\gamma_{tot}+2}$ to denote the received signal at receiver k. Like before, each receiver $k, k \in \psi$ needs $K - \gamma_{tot} - 2$ extra observations of $x_{\psi,1}, \ldots, x_{\psi,K-\gamma_{tot}-1}$ which will be seen from $L_{\psi,k'}, \forall k' \notin \psi$, which will come from order-$(\gamma_{tot} + 3)$ messages that are created by the transmitter and will be sent in the third phase.

phase j $(3 \le j \le K - \gamma_{tot})$*:* Note that

$$\Psi_{\gamma_{tot}+j} = \{\psi \in [K], |\psi| = \gamma_{tot} + j\}. \tag{48}$$

Similarly, during phase j, $f_i(\{L_{\psi\setminus\{k\},k}\}_{\forall k\in\psi}), i = 1, \ldots, \gamma_{tot}+j-1$ are delivered by $\boldsymbol{x}_\psi = [x_{\psi,1}, \ldots, x_{\psi,K-(\gamma_{tot}+j)+1}, 0, \ldots, 0]^T$ for each ψ. Like before, to solve $x_{\psi,1}, \ldots, x_{\psi,K-(\gamma_{tot}+j)+1}$, which can be seen from $L_{\psi,k'}, \forall k' \notin \psi$, order-$(\gamma_{tot} + j+1)$ messages are generated and will be sent in the next phase. For the last phase, fully common messages are sent from the transmitter and no more messages are created. Finally, the whole transmission is finished. We have

$$T_j = T_1\frac{\gamma_{tot}+1}{\gamma_{tot}+j}, j = 3, 2, \cdots, K - \gamma_{tot} \tag{49}$$

3.4 Decoding

After transmission, the received signals $y_k,\ k = 1, 2, \ldots, K$ for each s phase take the form

$$y_{k,t} = \underbrace{\boldsymbol{h}_{k,t}^T \boldsymbol{x}_{c,t}}_{P} + \underbrace{\boldsymbol{h}_{k,t}^T \boldsymbol{g}_{k,t} a_{k,t}}_{P^\alpha} + \underbrace{\sum_{i=1, i\neq k}^{K} \boldsymbol{h}_{k,t}^T \boldsymbol{g}_{i,t} a_{i,t}}_{p^0} + \underbrace{z_{k,t}}_{P^0} \tag{50}$$

where we see that, due to the power allocation and CSIT quality, symbols $a_{k,t}$ do not cause interference to unintended users; at least not above the noise level. At this point, each user k can decode $\boldsymbol{h}_{k,t}^T \boldsymbol{x}_{c,t}$ by treating all other signals as noise. Consequently, user k removes $\boldsymbol{h}_{k,t}^T \boldsymbol{x}_{c,t}$, and decodes its private symbol $a_{k,t}$. Then it can recover each $X_\psi^{(p)}$ with XOR operation.

After decoding $\{\boldsymbol{h}_{k,t}^T \boldsymbol{x}_{c,t}\}_{t=0}^T$, each receiver k reconstructs the overheard equations from backwards until phase 2. Then k obtains enough observations to solve $K - \gamma_{tot}$ symbols $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\gamma_{tot}}$. As a result, $X_\psi^{(c)}$ can be recovered.

Finally, user k can reconstruct X_ψ, and then it can recover $\{W_{R_k,\tau}\}_{\tau\in\Psi_{\gamma_{tot}}, k\notin\tau}$, from which W_{R_k} is obtained. All the other users act the same.

3.5 Duration Calculation

Now we focus on the performance of the duration. Based on the above, we get the entire duration

$$T = \sum_{j=1}^{K-\gamma_{tot}} T_j = \sum_{i=1}^{K-\gamma_{tot}} T_1 \frac{\gamma_{tot}+1}{\gamma_{tot}+i} \tag{51}$$

For each user k, the total amount of subfiles $\bigcup_{\tau\in\Psi_{\gamma_{tot}}, k\notin\tau} W_{R_k,\tau}^{(p)}$ is $\binom{K-1}{\gamma_{tot}}$, thus each a_k carries $(T\alpha f)/\binom{K-1}{\gamma_{tot}}$ bits of each subfile within T. Therefore, we can get the equivalent size of each $X_\psi^{(p)}$. Hence, the total amount of $\bigcup_{\psi\in\Psi_{\gamma_{tot}+1}} X_\psi^{(p)}$ that is delivered by a_k is $T\alpha f\binom{K}{\gamma_{tot}+1} \Big/ \binom{K-1}{\gamma_{tot}}$. It is shown that the second part $\bigcup_{\psi\in\Psi_{\gamma_{tot}+1}} X_\psi^{(c)}$ with $\theta\binom{K}{\gamma_{tot}+1}$ bits can be conveyed by $\boldsymbol{x}_c$ within T using delayed CSIT. Consequently, to deliver $\bigcup_{\psi\in\Psi_{\gamma_{tot}+1}} X_\psi$, we have

$$\frac{f\binom{K}{\gamma_{tot}+1}}{\binom{K}{\gamma_{tot}}} - \frac{T\alpha f\binom{K}{\gamma_{tot}+1}}{\binom{K-1}{\gamma_{tot}}} = \theta\binom{K}{\gamma_{tot}+1} \tag{52}$$

From (45), (51) and (52), we obtain

$$,T = \frac{(1-\gamma)(H_K - H_{\gamma_{tot}})}{\alpha(H_K - H_{\gamma_{tot}}) + (1-\alpha)(1-\gamma)} \tag{53}$$

It is obvious that when $\gamma_{tot} = K\gamma = K - 1$, we have $T = \frac{1}{K}, \forall\alpha$, which also achieves the optimal $1-\gamma$ under perfect current CSIT by delivering the fully common information. At this point, coded caching reduces the need of CSIT.

Specially, for the case of $\alpha = 0, \beta = 1$, we have $T = H_K - H_{\gamma_{tot}}$ and achieve DoF for sending $\{X_\psi\}_{\psi\in\Psi_{\gamma_{tot}+1}}$ as follows

$$d_{\gamma_{tot}+1} = \frac{f\binom{K}{\gamma_{tot}+1}}{\binom{K}{\gamma_{tot}}Tf} = \frac{K-\gamma_{tot}}{\gamma_{tot}+1}\frac{1}{H_K - H_{\gamma_{tot}}} \tag{54}$$

We describe our MAT-caching algorithm as follows. First assume that each common message generated from caching, e.g. $X_\psi, \psi \in \Psi_{\gamma_{tot}+1}$, is an order-$j(j \leq K)$ message—intended for j users simultaneously, then the scheme comprises $K - j + 1$ phase.

- In phase 1, the order-$j(j \leq K)$ messages are delivered by the transmitter from $K - j + 1$ of the transmit antennas.
- In phase $i \in \{2, \ldots, K - j\}$, order-$(j + i - 1)$ messages generated from the previous phase are sent from $K - (j + i) + 2$ of the transmit antennas. Note that, in the last phase, fully common messages are sent and no more messages are created.

Consequently, the corresponding DoF is

$$d_j = \frac{K-j+1}{j}\frac{1}{H_K - H_{j-1}} \tag{55}$$

In the scheme we describe above, for $\boldsymbol{x}_c$, $j = \gamma_{tot} + 1$.

Our method MAT-caching explores the role of delayed CSIT in reducing the duration of caching-aided MISO BC for any user demand. It is shown that after phase 1, the scheme jumps to phase $j + 1$ of MAT scheme directly. Comparing with MAT scheme, the scheme skips the first $j - 1$ phases with the assistance of caching.

4 Bounding the performance gap

4.1 Bounding the performance gap to optimal, for the case of $\alpha = 0$

We want to prove that

$$\frac{T(\alpha = 0)}{T^*(\alpha = 0)} \leq 2,$$

i.e., that the achievable $T(\alpha = 0) = H_K - H_{K\gamma}$ in (12), is within a factor of 2 from the optimal $T^*(\alpha = 0)$, which was bounded in Lemma 1 as

$$T^* \geq \max_{s\in\{1,\ldots,\min\{\lfloor \frac{N}{M}\rfloor,K\}\}} \frac{s}{d_{\sum}}(1 - \frac{M}{\lfloor \frac{N}{s}\rfloor}). \tag{56}$$

Using the result in [5] that says that for $\alpha = 0$, the sum-DoF is $d_{\sum} = \frac{s}{H_s}$, we see that the gap is bounded as

$$\frac{T}{T^*} \leq \frac{H_K - H_{K\gamma}}{\max\limits_{s\in\{1,\ldots,\lfloor \frac{1}{\gamma}\rfloor\}} H_s(1 - \frac{M}{\lfloor \frac{N}{s}\rfloor})}. \tag{57}$$

We will show that this gap is less than 2.

The proof is split into five parts: the first part will focus on the region $\gamma \leq \frac{1}{44}$, the second part on the region $\frac{1}{44} \leq \gamma \leq \frac{1}{4}$, the third part on the region $\frac{1}{4} \leq \gamma \leq \frac{1}{2}$ and the fourth part on the region $\frac{1}{2} \leq \gamma \leq \frac{K-1}{K}$.

4.1.1 Case 1: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\gamma \leq \frac{1}{44}$ and $K \geq 2$

First consider $\gamma \leq \frac{1}{44}$, and let us focus our attention to the case where $K \geq 5$ since when $K \leq 2$, there is no value of $\gamma \leq \frac{1}{44}$. Then

$$\frac{T}{T^*} \leq \max_{\gamma\in[\frac{1}{K},\frac{1}{44}]\cap(\mathbb{Z}/K)} \frac{H_K - H_{K\gamma}}{\max\limits_{s\in[1,\frac{K}{4}]\cap\mathbb{Z}} H_s(1 - \frac{M}{\lfloor \frac{N}{s}\rfloor})} \tag{58}$$

$$\leq \max_{\gamma\in[\frac{1}{K},\frac{1}{44}]} \frac{H_K - H_{K\gamma}}{\max\limits_{s\in[11,\frac{K}{4}]\cap\mathbb{Z}} H_s(1 - \frac{M}{\lfloor \frac{N}{s}\rfloor})} \tag{59}$$

$$\leq \max_{\gamma\in[\frac{1}{K},\frac{1}{44}]} \frac{\log(\frac{1}{\gamma}) + \epsilon_5 - \epsilon_\infty}{\max\limits_{s\in[11,\frac{K}{4}]\cap\mathbb{Z}} (\log s + \epsilon_\infty)(1 - \gamma s\frac{5}{4})} \tag{60}$$

where (58) holds because $H_s(1 - \frac{M}{\lfloor \frac{N}{s}\rfloor}) < 0$ when $s > \lfloor \frac{1}{\gamma}\rfloor$, where (59) holds to reflect the change of the maximizing regions for γ and s, and where (60) holds because ϵ_K decreases with K and because $H_K - \log(K) \leq \epsilon_5$, $H_{K\gamma} - \log(K\gamma) > \epsilon_\infty$, $H_s > \log(s) + \epsilon_\infty$, and because $(\lfloor \frac{N}{s}\rfloor)/\frac{N}{s} \geq \frac{4}{5}$, $s \leq \frac{N}{4}$. Continuing from (60), we have that

$$\frac{T}{T^*} \leq \max_{s_c\in[11,\frac{K}{4}]\cap\mathbb{Z}} \max_{\gamma\in[\frac{1}{4s_c},\frac{1}{4(s_c-1)}]} \frac{\log(\frac{1}{\gamma}) + \epsilon_5 - \epsilon_\infty}{(\log s_c + \epsilon_\infty)(1 - \gamma s_c\frac{5}{4})} \tag{61}$$

because $\max\limits_{s\in[11,\frac{K}{4}]\cap\mathbb{Z}} (\log s + \epsilon_\infty)(1 - \gamma s\frac{5}{4}) \geq (\log s_c + \epsilon_\infty)(1 - \gamma s_c\frac{5}{4})$ for any γ and for any $s_c \in [11, \frac{K}{4}]\cap\mathbb{Z}$, and where the split of the maximization $\max_{\gamma\in[\frac{1}{K},\frac{1}{44}]}$

into the double maximization $\max_{s_c\in[11,\frac{K}{4}]\cap\mathbb{Z}}\max_{\gamma\in[\frac{1}{4s_c},\frac{1}{4(s_c-1)}]}$ reflects the fact that we heuristically choose $s=s_c$ when $\gamma\in[\frac{1}{4s_c},\frac{1}{4(s_c-1)}]$. Now we perform a simple change of variables, introducing s' such that $\gamma=\frac{1}{4s'}$. Hence, a γ range of $\gamma\in[\frac{1}{4s_c},\frac{1}{4(s_c-1)}]$, corresponds to an s' range of $s'\in[s_c-1,s_c]$.

$$\frac{T}{T^*}\leq \max_{s_c\in[11,\frac{K}{4}]\cap\mathbb{Z}}\max_{s'\in[s_c-1,s_c]}\frac{\log(4s')+\epsilon_5-\epsilon_\infty}{(\log s_c+\epsilon_\infty)(1-\frac{5}{4}\frac{1}{4}\frac{s_c}{s'})} \tag{62}$$

$$\leq \max_{s_c\in[11,\frac{K}{4}]\cap\mathbb{Z}}\frac{\log(4s_c)+\epsilon_5-\epsilon_\infty}{(\log s_c+\epsilon_\infty)(1-\frac{5}{16}\frac{s_c}{s_c-1})} \tag{63}$$

$$\leq \frac{32}{21}\max_{s_c\in[11,\frac{K}{4}]\cap\mathbb{Z}}\frac{\log(4s_c)+\epsilon_5-\epsilon_\infty}{(\log s_c+\epsilon_\infty)} \tag{64}$$

$$\leq \frac{32}{21}+\frac{32}{21}\max_{s_c\in[11,\frac{K}{4}]\cap\mathbb{Z}}\frac{\log(4)+\epsilon_5-2\epsilon_\infty}{(\log(s_c)+\epsilon_\infty)} \tag{65}$$

$$=\frac{32}{21}+\frac{32}{21}\frac{\log(4)+\epsilon_5-2\epsilon_\infty}{(\log(11)+\epsilon_\infty)}<2. \tag{66}$$

Specifically, if $\frac{K}{4}$ is not an integer, $\gamma\in[\frac{1}{K},\frac{1}{4\lfloor\frac{K}{4}\rfloor}]$ is not considered in the above and only $\gamma=\frac{1}{K}$ is evolved due to the fact that $\frac{i}{K}\geq\frac{1}{4\lfloor\frac{K}{4}\rfloor},\forall i\geq 2$ when $K\geq 45$. For $\gamma=\frac{1}{K}$, we set $s_c=\lfloor\frac{1}{4\gamma}\rfloor=\lfloor\frac{K}{4}\rfloor$ where $s_c\geq 11$. Based on the above, we have

$$\frac{T}{T^*}\leq\frac{\log K+\epsilon_{44}-\epsilon_\infty}{(\log s_c+\epsilon_\infty)(1-\gamma s_c\frac{5}{4})} \tag{67}$$

$$\leq\frac{16}{11}\frac{\log(4s_c+3)+\epsilon_{44}-\epsilon_\infty}{(\log s_c+\epsilon_\infty)} \tag{68}$$

$$\leq\frac{16}{11}\frac{\log(4s_c)+\log(\frac{47}{44})+\epsilon_{44}-\epsilon_\infty}{(\log s_c+\epsilon_\infty)} \tag{69}$$

$$\leq\frac{16}{11}+\frac{16}{11}\frac{\log 4+\log(\frac{47}{44})+\epsilon_{44}-2\epsilon_\infty}{(\log s_c+\epsilon_\infty)} \tag{70}$$

$$<2,\forall s_c\geq 11 \tag{71}$$

where (67) holds from (61), where (68) holds because $K\leq 4s_c+3$ when $s_c=\lfloor\frac{K}{4}\rfloor$ and because $\gamma s_c\leq\frac{1}{4}$, where (69) holds because $\log(4s_c+3)-\log(4s_c)\leq\log 47-\log 44,\forall s_c\geq 11$.

4.1.2 Case 2: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)}\leq 2$ for $\frac{1}{44}\leq\gamma\leq\frac{1}{4}$ and $K\geq 2$

There are two sections.

1) For $K \geq 35$, we have

$$\frac{T}{T^*} \leq \frac{\log \frac{1}{\gamma} + \epsilon_{35} - \epsilon_\infty}{\max\limits_{s\in\{1,\ldots,\lfloor \frac{1}{\gamma}\rfloor\}} H_s(1-\frac{\gamma s}{1-\frac{s-1}{N}})} \tag{72}$$

$$\leq \frac{\log \frac{1}{\gamma} + \epsilon_{35} - \epsilon_\infty}{\max\limits_{s\in\{1,\ldots,\lfloor \frac{1}{\gamma}\rfloor\}} H_s(1-\frac{\gamma s}{1-\frac{s-1}{35}})} \tag{73}$$

$$\leq \max_{s_c\in[2,11]\cap\mathbb{Z}} \max_{\gamma\in[\frac{1}{4s_c},\frac{1}{4(s_c-1)}]} \frac{\log \frac{1}{\gamma} + \epsilon_{35} - \epsilon_\infty}{H_{s_c}(1-\frac{\gamma s_c}{1-\frac{s_c-1}{35}})} \tag{74}$$

$$=: \max_{s_c\in[2,11]\cap\mathbb{Z}} \max_{\gamma\in[\frac{1}{4s_c},\frac{1}{4(s_c-1)}]} g(s_c,\gamma) \tag{75}$$

where (72) holds because $H_K - \log(K) \leq \epsilon_{35}$, $H_{K\gamma} - \log(K\gamma) > \epsilon_\infty, \forall K \geq 35$ and because $\lfloor \frac{N}{s} \rfloor \geq \frac{N-(s-1)}{s}$, where (73) holds because $N \geq K \geq 35$, where (74) holds because of the fact that $s = s_c$ is chosen when $\gamma \in [\frac{1}{4s_c}, \frac{1}{4(s_c-1)}]$, which covers the maximization $\max_{\gamma\in[\frac{1}{44},\frac{1}{4}]}$.

No we focus on $g(s_c,\gamma)$ for each s_c. We note that

$$\begin{aligned}\frac{dg(s_c,\gamma)}{d\gamma} &= \big(\frac{1}{\gamma}(\frac{35\gamma s_c}{36-s_c} - 1) + (\log\frac{1}{\gamma} + \epsilon_{35})\frac{35 s_c}{36-s_c}\big)/A \\ &= g'_N/A \end{aligned} \tag{76}$$

for some $A > 0$, where g'_N denote the above numerator. To maximize $g(s_c,\gamma)$, first our task is to find the behavior of $\frac{dg(s_c,\gamma)}{d\gamma}$. We can see that $\frac{dg'_N}{d\gamma} = \frac{1}{\gamma}(\frac{1}{\gamma} - \frac{35 s_c}{36-s_c}) \geq 0$, $\forall\gamma \leq \frac{36-s_c}{35 s_c}$. For each $s_c \in [2,11]\cap\mathbb{Z}$, with the γ range $[\frac{1}{4s_c}, \frac{1}{4(s_c-1)}]$, to guarantee $\frac{dg'_N}{d\gamma} \geq 0$, we need to have $\frac{36-s_c}{35 s_c} \geq \frac{1}{4(s_c-1)}$, which can be guaranteed by direct calculation. Hence, we can get that $\frac{dg'_N}{d\gamma} \geq 0, \forall\gamma \in [\frac{1}{44}, \frac{1}{4}]$.

Lemma 2 *For the function $g(\gamma,s)$, we use $g'_N(\gamma,s)$ and $g'_D(\gamma,s)$ to denote the numerator and the denominator of $\frac{dg(\gamma,s)}{d\gamma}$ respectively. Having the increasing $g'_N(\gamma,s)$ in γ and positive $g'_D(\gamma,s)$, within range $\gamma \in [\gamma_1,\gamma_2], \gamma_1 \leq \gamma_2$, we obtain*

$$\max_{\gamma\in[\gamma_1,\gamma_2]} g(\gamma,s) = \max\{g(\gamma=\gamma_1,s), g(\gamma=\gamma_2,s)\} \tag{77}$$

Proof: See Section 7.1. □

From Lemma 2, for $K \geq 35$, with $\frac{dg'_N}{d\gamma} \geq 0$, we have

$$\begin{aligned}\frac{T}{T^*} &\leq \max_{s_c \in [2,11] \cap \mathbb{Z}} \max_{\gamma \in [\frac{1}{4s_c}, \frac{1}{4(s_c-1)}]} g(s_c, \gamma) \\ &\leq \max_{s_c \in [2,11] \cap \mathbb{Z}} \max\{g(s_c, \gamma = \frac{1}{4s_c}), g(s_c, \gamma = \frac{1}{4(s_c-1)})\} \\ &\leq 2 \end{aligned} \tag{78}$$

by direct calculation.

2) For $2 \leq K \leq 34$ and $\frac{1}{44} \leq \gamma \leq \frac{1}{4}$, we have

$$\frac{T}{T^*} \leq \frac{H_K - H_{K\gamma}}{\max\limits_{s \in \{1,\ldots,\lfloor \frac{1}{\gamma} \rfloor\}} H_s(1 - \frac{\gamma s}{1 - \frac{s-1}{N}})} \tag{79}$$

$$\leq \frac{H_K - H_{K\gamma}}{H_{s_c}(1 - \frac{\gamma s_c}{1 - \frac{s_c - 1}{K}})} \tag{80}$$

where (79) holds because $\lfloor \frac{N}{s} \rfloor \geq \frac{N-(s-1)}{s}$, where (80) holds because $N \geq K$ and we set $s_c = \left\lceil \frac{1}{4\gamma} \right\rceil$. By directly calculation, it is shown that $\frac{T}{T^*} \leq 2, \forall \gamma \in [\frac{1}{44}, \frac{1}{4}] \cap (\mathbb{Z}/K)$.

4.1.3 Case 3: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\frac{1}{4} \leq \gamma \leq \frac{1}{2}, K \geq 2$

We choose $s = 1$ and there are two sections.

1) For $K \geq 5$, we have

$$\frac{T}{T^*} \leq \frac{\log \frac{1}{\gamma} + \epsilon_5 - \epsilon_\infty}{1 - \gamma} =: f(\gamma) \tag{81}$$

where $H_K - \log(K) \leq \epsilon_5$, $H_{K\gamma} - \log(K\gamma) > \epsilon_\infty, \forall K \geq 5$.

Now we focus on $f(\gamma)$ and we have

$$\frac{df(\gamma)}{d\gamma} = \frac{1 - \gamma^{-1} - \log\gamma + \epsilon_5 - \epsilon_\infty}{(1-\gamma)^2} = \frac{f'_N(\gamma, s)}{f'_D(\gamma, s)}$$

Note that $f'_D(\gamma, s) > 0, \forall \gamma < 1$. Then

$$\frac{df'_N(\gamma)}{d\gamma} = \gamma^{-2} - \gamma^{-1} \geq 0, \forall \gamma \in [\frac{1}{4}, \frac{1}{2}]$$

From Lemma 2, we see that

$$\max_{\gamma \in [\frac{1}{4}, \frac{1}{2}]} f(\gamma) = \max\{f(\frac{1}{2}), f(\frac{1}{4})\} < 2 \Rightarrow \frac{T}{T^*} < 2 \tag{82}$$

2) For $K = 2, 3, 4$. With $s = 1$, $\frac{T}{T^*} \leq \frac{H_K - H_{K\gamma}}{1-\gamma} < 2$, $\frac{1}{4} \leq \gamma \leq \frac{1}{2}$ can be seen from direct calculation.

4.1.4 Case 4: Proving that $\frac{T(\alpha=0)}{T^*(\alpha=0)} \leq 2$ for $\frac{1}{2} \leq \gamma \leq \frac{K-1}{K}, K \geq 2$

We set $s = 1$. Let $\gamma = \frac{K-j}{K}$, where $j = 1, 2, \ldots, \lfloor \frac{K}{2} \rfloor$ since $\gamma \in [\frac{1}{2}, \frac{K-1}{K}]$. Therefore,

$$\begin{aligned}
\frac{T}{T^*} &\leq \frac{H_K - H_{K\gamma}}{1-\gamma} = \frac{H_K - H_{(K-j)}}{j/K} \\
&= \frac{1}{j}\left(\frac{K}{K-j+1} + \frac{K}{K-j+2} + \cdots + 1\right) \\
&= \frac{1}{j}\left(1 + \frac{j-1}{K-(j-1)} + 1 + \frac{j-2}{K-(j-2)} + \cdots + 1\right) \\
&= 1 + \frac{1}{j}\left(\frac{j-1}{K-(j-1)} + \frac{j-2}{K-(j-2)} + \cdots + \frac{1}{K-1}\right) \\
&\leq 2
\end{aligned} \tag{83}$$

since $j \leq \frac{K}{2}$.

As a result, we have that for $\gamma_{tot} = K\gamma \geq 1$, the performance $T = H_K - H_{K\gamma}$ of the proposed MAT-caching scheme, is within a multiplication factor of 2 from the optimal, i.e.,

$$\frac{T}{T^*} \leq 2 \tag{84}$$

$\forall K, \forall N \geq K, \forall \gamma \in \{\frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}\}$. The values of $K\gamma$ imply that all the information from N files is repeated $K\gamma$ times in the caches.

4.2 Bounding the performance gap, for the case of $\alpha > 0$

In this section, we will show that

$$\frac{T(\alpha > 0)}{T^*(\alpha > 0)} \leq 2,$$

Using that $d_{\sum} = \frac{s}{H_s}(1-\alpha) + s\alpha$, we see that the gap is bounded as

$$\frac{T}{T^*} \leq \frac{\frac{(1-\gamma)(H_K - H_{K\gamma})}{\alpha(H_K - H_{K\gamma}) + (1-\alpha)(1-\gamma)}}{\max\limits_{s \in \{1,\ldots,\lfloor \frac{N}{M} \rfloor\}} \frac{H_s}{(1-\alpha)+\alpha H_s}\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)} =: g(s, \gamma) \tag{85}$$

We will prove that this gap is less than 2.

Before starting the proof, from the previous section, we have already shown that

$$\frac{H_K - H_{K\gamma}}{\max\limits_{s \in \{1,\ldots,\lfloor \frac{N}{M} \rfloor\}} H_s\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)} \leq \frac{H_K - H_{K\gamma}}{H_{s_c}\left(1 - \frac{M}{\lfloor \frac{N}{s_c} \rfloor}\right)} \leq 2 \tag{86}$$

where s_c is chosen according to different γ ranges, which will be useful for the proof.

The proof consists two parts: the first part will focus on the region $\frac{1}{4} \leq \gamma \leq \frac{K-1}{K}$, and the second part on the region $0 \leq \gamma \leq \frac{1}{4}$.

4.2.1 Case 1: Proving that $\frac{T(\alpha>0)}{T^*(\alpha>0)} \leq 2$ for $\frac{1}{4} \leq \gamma \leq \frac{K-1}{K}, \forall K$

We set $s = 1$. From (86), it is shown that $\frac{H_K - H_{K\gamma}}{1-\gamma} \leq 2, \forall \gamma \in [\frac{1}{4}, \frac{K-1}{K}], \forall K$. Thus we have

$$\frac{T}{T^*} \leq \frac{H_K - H_{K\gamma}}{1-\gamma} \leq 2 \tag{87}$$

since $T(\alpha > 0) \leq T(\alpha = 0) = H_K - H_{K\gamma}$.

4.2.2 Case 2: Proving that $\frac{T(\alpha>0)}{T^*(\alpha>0)} \leq 2$ for $0 \leq \gamma \leq \frac{1}{4}, \forall K$

We can see that

$$g(s_c, \gamma) \leq \frac{H_K - H_{K\gamma}}{H_{s_c}(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor})} \Rightarrow H_{s_c} \leq \frac{H_K - H_{K\gamma}}{1-\gamma} \tag{88}$$

is a sufficient condition to guarantee $\frac{T}{T^*} \leq 2, \forall \gamma \in [0, \frac{1}{4}]$. If (88) works, we can get that

$$\log(s_c) \leq \frac{\log(\gamma)}{\gamma - 1} - \epsilon_2, \epsilon_2 = H_2 - \log 2. \tag{89}$$

is a sufficient condition since $H_{s_c} \leq \log(s_c) + \epsilon_2, \forall s_c \geq 2, \forall \gamma \in [0, \frac{1}{4}], \forall K$. Since $\gamma \in [\frac{1}{4s_c}, \frac{1}{4(s_c-1)}]$, so $s_c \leq \frac{1}{4\gamma} + 1$, then we only need to guarantee

$$\log(\frac{1}{4\gamma} + 1) \leq \frac{\log(\gamma)}{\gamma - 1} - \epsilon_2. \tag{90}$$

If (90) holds, then

$$\log(\frac{1}{4\gamma}) + \log 2 \leq \frac{\log(\gamma)}{\gamma - 1} - \epsilon_2 \tag{91}$$

is a sufficient condition. We set $f = \log(\frac{1}{4\gamma}) + \log 2 - \frac{\log(\gamma)}{\gamma - 1} + \epsilon_2$ and we use

$$\frac{df(\gamma)}{d\gamma} = \frac{1 + \log(\gamma) - \gamma}{(1-\gamma)^2} \leq 0$$

to get

$$\max_{\gamma \in [0, \frac{1}{4}]} f = f(\gamma = \frac{1}{4}) = \log 2 - \frac{4}{3} \log 4 + \epsilon_2 \leq 0$$

which proves (91) and then (88), (89) and (90) hold, i,e., $\frac{T}{T^*} \leq 2$.

5 Appendix - Lower bound on T^*

Let us first create the outer (lower) bound on T, using basic cut-set bound arguments. Consider a simplified setting, where there are s users ($s \in \{1, \ldots, \lfloor \frac{N}{M} \rfloor\}$). During the placement phase, the users' corresponding caches $Z_1, \ldots, Z_s$ are filled, while during the delivery phase, each of the s users makes $\lfloor \frac{N}{s} \rfloor$ sequential requests, corresponding to a total of $s\lfloor \frac{N}{s} \rfloor$ requested files $W_1, \ldots, W_{s\lfloor \frac{N}{s} \rfloor}$ by all the users together. Note that for integer $\frac{N}{s}$, these requests span all N files. We now consider a total of $\lfloor \frac{N}{s} \rfloor$ sequential transmissions $X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}$, such that X_1 and $Z_1, \ldots, Z_s$ can reconstruct $W_1, \ldots, W_s$, such that similarly X_2 and $Z_1, \ldots, Z_s$ can reconstruct $W_{s+1}, \ldots, W_{2s}$, and so on, until we have that $X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}$ and $Z_1, \ldots, Z_s$ can reconstruct all the requested files $W_1, \ldots, W_{s\lfloor \frac{N}{s} \rfloor}$.

To apply the cut-set bound, we place the $\lfloor \frac{N}{s} \rfloor$ broadcasting signals $X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}$, each of duration T, on one side of the cut, together with all the caches $Z_1, \ldots, Z_s$, and then on the other side of the cut, we place all the requests of s users for a total of $s\lfloor \frac{N}{s} \rfloor$ files, each of size f. Hence it follows that

$$\begin{aligned}
&\lfloor \frac{N}{s} \rfloor d_{\sum} T + sM \geq H(Z_1, \ldots, Z_s, X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}) \\
&\geq H(Z_1, \ldots, Z_s, X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor} | W_1, \ldots, W_{s\lfloor \frac{N}{s} \rfloor}) + s\lfloor \frac{N}{s} \rfloor (1 - \epsilon_f) \\
&\geq s\lfloor \frac{N}{s} \rfloor (1 - \epsilon_f) \qquad (92)
\end{aligned}$$

where we have used that the $K \times s$ interference-free MIMO channel provides $d_{\sum}$ degrees of freedom as a result of a certain of CSIT quality (this is in the limit of $f \to \infty$), and where we have used Fano's inequality. In the same limit of $f \to \infty$, we have that $\epsilon_f \to 0$. Thus solving for T, and optimizing over all possible choices of s, we obtain

$$T \geq \max_{s \in \{1, \ldots, \lfloor \frac{N}{M} \rfloor\}} \frac{s}{d_{\sum}} (1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}) \qquad (93)$$

proving the theorem.

6 Appendix - Proof of the asymptotic optimality

We first focus on the optimality of the schemes having $\alpha = 0$. First set $\gamma_{tot} = K^{\varphi}$, where $\varphi < 1$. We have

$$H_K - H_{K\gamma} \leq \log K + \epsilon_2 - \log(K\gamma) \leq \log \frac{1}{\gamma} + \epsilon_2$$

Hence, for both cases, $T \leq \log \frac{1}{\gamma} + \epsilon_2, \forall \gamma_{tot} \geq 0$. Then,

$$\begin{aligned}\frac{T}{T^*} &\leq \lim_{K\to\infty} \frac{\log \frac{1}{\gamma} + \epsilon_2}{\max\limits_{s\in\{1,\ldots,\min\{K,\lfloor \frac{N}{M}\rfloor\}\}} H_s(1-\frac{M}{\lfloor \frac{N}{s}\rfloor})} \\ &\leq \lim_{K\to\infty} \frac{\log \frac{1}{\gamma} + \epsilon_2}{H_{s_c}(1-\frac{M}{\lfloor \frac{N}{s_c}\rfloor})} \end{aligned} \tag{94}$$

We choose $s_c = \frac{1}{\gamma L}$, where L is a large but finite positive real chosen such that $s_c \in \mathbb{Z}^+$. Then we have

$$s_c = \frac{K}{\gamma_{tot} L} = \frac{K^{1-\varphi}}{L} \gg 1 \tag{95}$$

therefore, $N \geq K = \gamma_{tot} L s_c \gg s_c$. As a result, $\frac{\lfloor \frac{N}{s}\rfloor}{\frac{N}{s}} \to 1$. Consequently, from (94), for both cases, we have

$$\frac{T}{T^*} \leq \lim_{K\to\infty} \frac{\log \frac{1}{\gamma} + \epsilon_2}{\log s_c(1-\gamma s_c)} = \lim_{K\to\infty} \frac{\log L + \log s_c}{\log s_c(1-\frac{1}{L})} \to 1 \tag{96}$$

implying that the schemes are asymptotically optimal.

7 Appendix - Additional proofs

7.1 Proof of Lemma 2

$\frac{dg'_N(\gamma,s)}{d\gamma} \geq 0$ means that $g'_N(\gamma, s)$ is increasing in γ. Additionally, $g'_D(\gamma, s)$ is non-negative, within range $\gamma \in [\gamma_1, \gamma_2], \gamma_1 \leq \gamma_2$. At this point, there are three cases.

Consider case 1. If $g'_N(\gamma_1, s) \geq 0 \Rightarrow g'_N(\gamma, s) \geq 0, \forall \gamma \in [\gamma_1, \gamma_2]$. Hence, we have $\frac{dg(\gamma,s)}{d\gamma} = \frac{g'_N(\gamma,s)}{g'_D(\gamma,s)} \geq 0, \forall \gamma \in [\gamma_1, \gamma_2]$. Consequently,

$$\max_{\gamma\in[\gamma_1,\gamma_2]} g(\gamma, s) = g(\gamma = \gamma_2, s)$$

Consider case 2. If $g'_N(\gamma_1, s) < 0$ & $g'_N(\gamma_2, s) \leq 0 \Rightarrow g'_N(\gamma, s) \leq 0, \forall \gamma \in [\gamma_1, \gamma_2]$. Hence, we have $\frac{dg(\gamma,s)}{d\gamma} \leq 0, \forall \gamma \in [\gamma_1, \gamma_2]$. Consequently,

$$\max_{\gamma\in[\gamma_1,\gamma_2]} g(\gamma, s) = g(\gamma = \gamma_1, s)$$

Consider case 3. If $g'_N(\gamma_1, s) < 0$ & $g'_N(\gamma_2, s) > 0 \Rightarrow \exists$ a unique $\gamma' \in [\gamma_1, \gamma_2]$ $s.t.$, $g'_N(\gamma', s) = 0$. Hence, we have $\frac{dg(\gamma,s)}{d\gamma} \leq 0, \forall \gamma \in [\gamma_1, \gamma']$ & $\frac{dg(\gamma,s)}{d\gamma} \geq 0, \forall \gamma \in [\gamma', \gamma_2]$. Consequently,

$$\max_{\gamma\in[\gamma_1,\gamma_2]} g(\gamma, s) = \max\{g(\gamma = \gamma_1, s), g(\gamma = \gamma_2, s)\}$$

As a result, the above three cases yield the identity

$$\max_{\gamma\in[\gamma_1,\gamma_2]} g(\gamma,s) = \max\{g(\gamma=\gamma_1,s), g(\gamma=\gamma_2,s)\} \tag{97}$$

7.2 Proof of vanishing fraction of delayed CSIT cost due to caching from Section 2.6

The MAT-caching scheme has $K-\gamma_{tot}$ phase, where each phase $j, j \in [\gamma_{tot}+1, K] \cap \mathbb{Z}$ transmits order-j messages in the presence of delayed CSIT. It is shown that the phase j takes $(K-j+1)\binom{K}{j}$ common symbols of order j, and creates $j\binom{K}{j+1}$ of order $j+1$. We use N_j to denote the number of order-j messages and T_j to denote the duration of phase j. Towards this, first we have

$$N_{j+1} = N_j \frac{j\binom{K}{j+1}}{(K-j+1)\binom{K}{j}} = N_j \frac{j(K-j)}{(j+1)(K-j+1)} \tag{98}$$

which implies that

$$N_j = N_{j-1}\frac{(j-1)(K-j+1)}{j(K-j+2)}, \ldots, N_2 = N_1 \frac{K-1}{2K} \tag{99}$$

Hence, we can get

$$N_j = N_1 \frac{K-j+1}{Kj}, j = 1, 2, \ldots, K \tag{100}$$

The order-j symbols generated from phase $j-1$ will be sent in phase j from $K-j+1$ transmit antennas. At this point, we can get the duration of each phase,

$$T_j = \frac{N_j}{K-j+1} = \frac{N_1}{Kj} \tag{101}$$

Towards this, we can see that for phase j, to let the transmitter construct $K-j$ received signals that are not desired by the receiving users, $(K-j+1)(K-j)$ supports every coherence time should be sent back after phase j.

To serve each user a single file with $\log P$ bits, which means $N_1 = K$, so we get $T_j = \frac{1}{j}$ implying that the total duration is $\sum_{j=1}^{K} \frac{1}{j} = H_K$, and $(K-j+1)(K-j)$ supports every coherence time are needed. Hence, The total number of scalars S of $\log P$ bits for each user every coherence time that are needed at the transmitter are

$$\begin{aligned} S &= \sum_{j=\gamma_{tot}+1}^{K} T_j(K-j+1)(K-j) \\ &= (K^2+K)(H_K - H_{\gamma_{tot}}) - \frac{K(1-\gamma)(3K-K\gamma-1)}{2} \end{aligned} \tag{102}$$

Similarly, for MAT scheme, $\gamma_{tot} = 0$, we have

$$S' = \sum_{j=1}^{K} T_j(K-j+1)(K-j) = (K^2+K)H_K - \frac{3K^2}{2} + \frac{K}{2} \tag{103}$$

When K goes to large, we have $\lim_{K\to\infty} S = K^2(\log\frac{1}{\gamma} - \frac{3}{2})$, and $\lim_{K\to\infty} S' = K^2 \log K$, then we have

$$\frac{S}{S'} \approx \frac{\log\frac{1}{\gamma} - \frac{3}{2}}{\log K} \tag{104}$$

For a finite and fixed γ, the ratio goes to 0. Hence, we can see that caching algorithm requires a much reduced delayed-CSIT load than the full K-user MAT algorithm.

We can see that, for the optimal performance, the transmitter needs K^2 supports of the channel information using zero-forcing. The area under the line represents the total supports that each scheme requires. Our MAT-caching needs a fraction of retrospective knowledge of the channel. With γ increasing, the need of channel information is even less than the zero-forcing scheme.

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. 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.

[3] 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.

[4] 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.

[5] 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.

[6] T. Gou and S. 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, Jul. 2012.

[7] 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.

[8] P. de Kerret, X. Yi, and D. Gesbert, "On the degrees of freedom of the k-user time correlated broadcast channel with delayed CSIT," *CoRR*, vol. abs/1301.2138, 2013. [Online]. Available: http://arxiv.org/abs/1301.2138

[9] J. Chen, S. Yang, and P. Elia, "On the fundamental feedback-vs-performance tradeoff over the MISO-BC with imperfect and delayed CSIT," *CoRR*, vol. abs/1302.0806, 2013. [Online]. Available: http://arxiv.org/abs/1302.0806

[10] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045 – 5060, Nov. 2006.

[11] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691 – 1706, Jul. 2003.

[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," Aug. 2012, to appear in *IEEE Trans. Inform. Theory*, available on arXiv:1208.5071.

[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," Feb. 2013, to appear in *ICC13*, available on arXiv:1302.6521.

[16] G. Caire, N. Jindal, and S. Shamai, "On the required accuracy of transmitter channel state information in multiple antenna broadcast channels," in *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on*, Nov 2007, pp. 287–291.

[17] 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

[18] M. Ji, M. F. Wong, A. Tulino, J. Llorca, G. Caire, M. Effros, and M. Langberg, "On the fundamental limits of caching in combination networks," in *Signal Processing Advances in Wireless Communications (SPAWC), 2015 IEEE 16th International Workshop on*, June 2015, pp. 695–699.

[19] 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

[20] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT'2015)*, Hong-Kong, China, 2015.

[21] 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

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

[23] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, "Caching eliminates the wireless bottleneck in video-aware wireless networks," *CoRR*, vol. abs/1405.5864, 2014. [Online]. Available: http://arxiv.org/abs/1405.5864

[24] 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

[25] J. Hachem, N. Karamchandani, and S. Diggavi, "Effect of number of users in multi-level coded caching," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT'2015)*, Hong-Kong, China, 2015.

[26] 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.

[27] J. Chen, S. Yang, and P. Elia, "On the fundamental feedback-vs-performance tradeoff over the MISO-BC with imperfect and delayed CSIT," Feb. 2013, to appear in *ISIT13*, available on arXiv:1302.0806.