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

Jingjing Zhang and Petros Elia

***Abstract*—Building on the recent coded-caching breakthrough by Maddah-Ali and Niesen, the work here considers the K-user cache-aided wireless multi-antenna (MISO) symmetric broadcast channel (BC) with random fading and imperfect feedback, and analyzes the throughput performance as a function of feedback statistics and cache size. In this setting, our work identifies the optimal cache-aided degrees-of-freedom (DoF) within a factor of 4, by identifying near-optimal schemes that exploit the new synergy between coded caching and delayed CSIT, as well as by exploiting the unexplored interplay between caching and feedback-quality.**

The derived limits interestingly reveal that — the combination of imperfect quality current CSIT, delayed CSIT, and coded caching, guarantees that — the DoF gains have an initial offset defined by the current CSIT quality, and then that the additional gains attributed to coded caching are exponential, in the sense that any linear decrease in the required DoF performance, allows for an exponential reduction in the required cache size.

I. Introduction

Recent work by [1] explored — for the single-stream broadcast setting — how careful caching of content at the receivers, and proper encoding across different users' requested data, can allow for higher communication rates. The key idea was to use coding in order to create multicast opportunities, even if the different users requested different data content. This *coded caching* approach involved two phases: the placement phase and the delivery phase. In the placement phase, content that was predicted to be popular (a library of commonly requested files), was coded and placed across user's caches. In the delivery phase — which started when users requested specific files from the predicted library of files — the transmitter encoded across users' requests, taking into consideration the existing cache contents. This approach — which translated to efficient interference removal gains that were termed as 'coded-caching gains' — was shown in [1] to provide substantial performance improvement that far exceeded the 'local' caching gains from only pre-storing content at local caches.

Our interest here is to explore coded caching, not in the original single-stream setting in [1], but rather in the feedback-aided multi-antenna wireless BC. This wireless and multi-antenna element now automatically brings to the fore a largely unexplored and involved relationship between coded caching and CSIT-type feedback quality. This relationship carries particular importance because both CSIT and coded caching are powerful and crucial ingredients in handling interference, because they are both hard to implement individually, and because their utility is affected by one another (often adversely, as we will see). Our work tries to understand how CSIT and caching resources jointly improve performance, as well as tries to shed some light on the interplay between coded caching and feedback.

1) Motivation for the current work: A main motivation in [1] and in subsequent works, was to employ coded caching to remove interference. Naturally, in wireless networks, the ability to remove interference is very much linked to the quality and timeliness of feedback, and thus any attempt to further our understanding of the role of coded caching in these networks, stands to benefit from understanding the interplay between coded caching and (variable quality) feedback. This joint exposition becomes even more meaningful when we consider the connections that exist between feedback-usefulness and cached side-information at receivers, where principally the more side information receivers have, the less feedback information the transmitter might need.

This approach is also motivated by the fact that feedback is hard to get in a timely manner, and hence is typically far from ideal and perfect. Thus, given the underlying links between the two, perhaps the strongest reason to jointly consider coded caching and feedback, comes from the prospect of using coded caching to alleviate the constant need to gather and distribute CSIT, which — given typical coherence durations — is an intensive task that may have to be repeated hundreds of times per second during the transmission of content. This suggests that content prediction of a predetermined library of files during the night (off peak hours), and a subsequent caching of parts of this library content again during the night, may go beyond boosting performance, and may in fact offer the additional benefit of alleviating the need for prediction, estimation, and communication of CSIT during the day, whenever requested files are from the library. Our idea of exploring the interplay between feedback (timeliness and quality) and coded caching, hence draws directly from this attractive promise that content prediction, once a day, can offer repeated and prolonged savings in CSIT.

A. Cache-aided broadcast channel model

1) K-user BC with pre-filled caching: In the symmetric K-user multiple-input single-output (MISO) BC of interest

The authors are with the Communications 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 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.

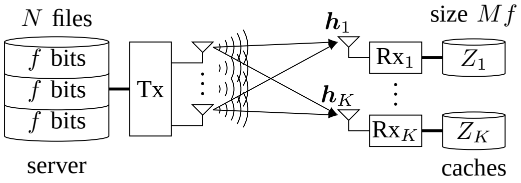

Fig. 1. Cache-aided K-user MISO BC.

here, the K-antenna transmitter, communicates to K single-antenna users. The transmitter has access to a library of $N \geq K$ distinct files $W_1, W_2, \dots, W_N$, each of size $|W_n| = f$ bits. Each user $k \in \{1, 2, \dots, K\}$ has a cache Z_k, of size $|Z_k| = Mf$ bits, where naturally $M \leq N$. Communication consists of the aforementioned *content placement phase* and the *delivery phase*. In the placement phase — which corresponds to communication during off-peak hours — the caches $\{Z_k\}_{k=1}^{K}$ are pre-filled with content from the N files $\{W_n\}_{n=1}^{N}$. Then the delivery phase commences when each user k requests, any *one* file $W_{R_k} \in \{W_n\}_{n=1}^{N}$, out of the N library files. Each file is requested with equal probability. 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. We will consider the normalized

$$\gamma \triangleq \frac{M}{N} \tag{1}$$

as well as the cumulative

$$\Gamma \triangleq \frac{KM}{N} = K\gamma. \tag{2}$$

The latter simply means that the sum of the cache sizes, is Γ times the volume of the N-file library.

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

$$y_k = \boldsymbol{h}_k^T \boldsymbol{x} + z_k, \quad k = 1, \dots, K \tag{3}$$

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 the desired file W_{R_k}.

B. Coded caching and CSIT-type feedback

Communication also takes place in the presence of feedback. CSIT-type feedback is crucial in handling interference, and can thus substantially reduce the resulting duration T of the delivery phase. This 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

$$\alpha \triangleq - \lim_{P \to \infty} \frac{\log \mathbb{E}[||\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k||^2]}{\log P}, \; k \in \{1, \dots, K\} \tag{4}$$

where $\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k$ denotes the estimation error between the channel $\boldsymbol{h}_k$ and the current CSIT estimate $\hat{\boldsymbol{h}}_k$. The range of interest[1] is $\alpha \in [0, 1]$. We also assume availability of delayed CSIT (as in for example [3], as well as in a variety of subsequent works [4]–[15]) where now the delayed estimates of any channel, can be received without error but with arbitrary delay, even if this delay renders this CSIT completely obsolete. As it is argued in [4], this mixed CSI model (partial current CSIT, and delayed CSIT) nicely captures different realistic settings that might involve channel correlations and an ability to improve CSI as time progresses. This same CSI model is particularly well suited for our caching-related setting here, because it explicitly reflects two key ingredients that are directly intertwined with coded caching; namely, feedback timeliness and quality.

a) Intuitive links between α and γ: As we will see, α is not only linked to the performance — where a higher α allows for better interference management and higher performance over the delivery link — but is also linked to caching; after all, the bigger the γ, the more side information the receivers have, the less interference one needs to handle (at least in symmetric systems), and the smaller the α that is potentially needed to steer interference. This means that principally, a higher γ implies that more common information needs to be sent, which may (in some cases) diminish the utility of feedback which primarily aims to facilitate the opposite which is the transmission of private information. For example, in the presence of $\Gamma = K - 1$ (we will see later), there is no need for CSIT to achieve the optimal performance.

C. Measures of performance in current work

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 wireless 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 there 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 $\log_2(P)$ — we set

$$f = \log_2(P) \tag{5}$$

which guarantees that the two measures of performance, here and in [1], are the same and can thus be directly compared[2].

A simple inversion leads to the equivalent measure of the per-user DoF

$$d(\gamma, \alpha) = \frac{1 - \gamma}{T} \tag{6}$$

which captures the joint effect of coded caching and feedback[3].

[1] In the high SNR regime of interest here, $\alpha = 0$ corresponds to having essentially no current CSIT (cf. [2]), while having $\alpha = 1$ corresponds (again in the high SNR regime) to perfect and immediately available CSIT.

[2] Setting $f = \log_2(P)$ is simply a normalization choice, and does not carry a 'forced' relationship between SNR and file size. The essence of the derived results would remain the same for any other non-trivial choice.

[3] The DoF measure is designed to exclude the benefits of having some content already available at the receivers (local caching gain), and thus to limit the DoF between 0, and the interference free optimal DoF of 1.

D. Notation and assumptions

1) Notation: We will use $H_n \triangleq \sum_{i=1}^{n} \frac{1}{i}$ to represent the n-th harmonic number, and $\epsilon_n \triangleq H_n - \log(n)$ to represent its logarithmic approximation error, for some integer n. We remind the reader that ϵ_n decreases with n, and that $\epsilon_\infty \triangleq \lim_{n\to\infty} H_n - \log n \approx 0.5772$. $\mathbb{Z}$ 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 $[K] \triangleq \{1, 2, \cdots, K\}$. 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 will write $g(P) \doteq P^B$ to denote $\lim_{P\to\infty} \frac{\log_2 g(P)}{\log_2 P} = B$. Logarithms are of base e, unless we use $\log_2(\cdot)$ which will represent a logarithm of base 2.

2) Main assumptions: Throughout this work we adopt the mixed-CSIT model, and also adhere to the common convention (see for example [6]) 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. We also make the soft assumption that the transmitter *during the delivery phase* is aware of the feedback statistics. Removing this assumption entails, for $\alpha > 0$, a performance penalty which is small.

E. Prior work

Deviating from single-stream error free links, different works have considered the use of coded caching in different wireless networks, without though particular consideration for CSIT feedback quality. For example, work by Huang et al. in [16], considered a cache-aided wireless fading BC where each user experiences a different link quality, and proposed a suboptimal communication scheme. Further work by Timo and Wigger in [17] considered an erasure broadcast channel and explored how the cache-aided system efficiency can improve by employing unequal cache sizes that are functions of the different channel qualities. Another work can be found in [18] where Maddah-Ali and Niesen studied the wireless interference channel where each transmitter has a local cache, and showed distinct benefits of coded caching that stem from the fact that content-overlap at the transmitters allows effective interference cancellation. Further related work on caching can be found in [19]–[24].

Work that combines caching and feedback considerations in wireless networks, has only just recently started. A reference that combines these, can be found in [25] where Deghel et al. considered a MIMO interference channel (IC) with caches at the transmitters. In this setting, whenever the requested data resides within the pre-filled caches, the data-transfer load of the backhaul link is alleviated, thus allowing for these links to be instead used for exchanging CSIT that supports interference alignment. A concurrent work can be found in [26] where Ghorbel et al. studied the capacity of the cache-enabled broadcast packet erasure channel with ACK/NACK feedback. In this setting, Ghorbel et al. cleverly showed — interestingly also using a retrospective type algorithm, this time by Gatzianas et al. in [27] — how feedback can improve performance by informing the transmitter when to resend the packets that are not received by the intended user and which are received by unintended users, thus allowing for multicast opportunities. Another work by Shariatpanahi can be found in [28] where perfect CSIT feedback was consider in one networks: linear networks. More recent works on can be found in [29], [30]. The first work that considers the actual interplay between coded caching and CSIT quality, can be found in [31] which considered the easier problem of how the optimal cache-aided performance (with coded caching), can be achieved with reduced quality CSIT.

F. Outline and contributions

In Section II we present the achievable $T(\gamma, \alpha)$, for $\Gamma \in [K]$, $\alpha \in [0, 1]$, and prove it to be less than four times the optimal, thus identifying the optimal $T^*(\gamma, \alpha)$ within a factor of 4. From the per-user DoF perspective, we see that even a very small γ can offer a substantial DoF boost.

In Section III we discuss practical implications. In Corollary 2b we describe the savings in current CSIT that we can have due to coded caching, while in Corollary 2d we quantify the intuition that, with coded-caching, there is no reason to improve CSIT beyond a certain threshold quality.

In Section IV we present the caching-and-delivery schemes, which build on the interesting connections between MAT-type retrospective transmission schemes (cf. [3]) and coded caching. The caching and transmission algorithms are calibrated so that the caching algorithm — which is modified from [1] to adapt the caching redundancy to α — creates the same multi-destination delivery problem that is efficiently solved by the last stages of the QMAT scheme.

II. Throughput of cache-aided BC as a function of CSIT quality and caching resources

The following results hold for the (K, M, N, α) cache-aided K-user wireless MISO BC with random fading, $\alpha \in [0, 1]$ and $N \geq K$, where $\gamma = \frac{M}{N}$ and $\Gamma = K\gamma$. We begin with an outer bound (lower bound) on the optimal T^*.

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

$$T^*(\gamma, \alpha) \geq \max_{s\in\{1,\dots,\lfloor \frac{N}{M}\rfloor\}} \frac{1}{(H_s\alpha + 1 - \alpha)}(H_s - \frac{Ms}{\lfloor \frac{N}{s} \rfloor}). \quad (7)$$

Proof: The proof can be found in the journal version of this work [32]. ■

A. Achievable throughput of the cache-aided BC

The following identifies, up to a factor of 4, the optimal T^*, for all $\Gamma \in \{1, 2, \cdots, K\}$ (i.e., $M \in \frac{N}{K}\{1, \cdots, K\}$). The result uses the expression

$$\alpha_{b,\eta} = \frac{\eta - \Gamma}{\Gamma(H_K - H_\eta - 1) + \eta}, \ \eta = \lceil \Gamma \rceil, \ldots, K-1. \quad (8)$$

Note that the above does not hold for $\Gamma = K$, as this would imply no need for delivery.

Theorem 1: In the (K, M, N, α) cache-aided MISO BC with N files, $K \leq N$ users, $\Gamma \in \{1, 2, \cdots, K\}$, and for $\eta = \arg\max_{\eta' \in [\Gamma, K-1] \cap \mathbb{Z}} \{\eta' \ : \ \alpha_{b,\eta'} \leq \alpha\}$, then

$$T = \max\{1-\gamma, \frac{(K-\Gamma)(H_K - H_\eta)}{(K-\eta) + \alpha(\eta + K(H_K - H_\eta - 1))}\} \quad (9)$$

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

Proof: The caching and delivery scheme that achieves the above performance is presented in Section IV. The proof regarding the gap to optimal can be found in the journal version of this work [32]. ■

The above is achieved with a general scheme whose caching phase is a function of α. We will consider a special case ($\eta = \Gamma$) of this scheme, which provides similar performance (it again has a gap to optimal that is bounded by 4), and has the practical advantage that the caching phase need not depend on the CSIT statistics α of the delivery phase. For this case, we can achieve the following performance.

Theorem 2: In the (K, M, N, α) cache-aided MISO BC with $\Gamma \in \{1, 2, \cdots, K\}$,

$$T = \frac{(1-\gamma)(H_K - H_\Gamma)}{\alpha(H_K - H_\Gamma) + (1-\alpha)(1-\gamma)} \quad (10)$$

is achievable and has a gap from optimal

$$\frac{T}{T^*} < 4 \quad (11)$$

that is less than 4, for all α, K. Thus the corresponding per-user DoF takes the form

$$d(\gamma, \alpha) = \alpha + (1-\alpha)\frac{1-\gamma}{H_K - H_\Gamma}. \quad (12)$$

Proof: The scheme that achieves the above performance will be described as a special (simpler) case of the scheme corresponding to Theorem 1. The proof for the gap to optimal can be found in the journal version of this work [32]. ■

The following corollary describes the above achievable T, under the logarithmic approximation $H_n \approx \log(n)$. The presented expression is exact in the large K where $\frac{H_K - H_\Gamma}{\log(\frac{1}{\gamma})} = 1$, which is tight for any *fixed* γ.

Corollary 2a: Under the logarithmic approximation $H_n \approx \log(n)$, the derived T takes the form

$$T(\gamma, \alpha) = \frac{(1-\gamma)\log(\frac{1}{\gamma})}{\alpha \log(\frac{1}{\gamma}) + (1-\alpha)(1-\gamma)} \quad (13)$$

and the derived DoF takes the form

$$d(\gamma, \alpha) = \alpha + (1-\alpha)\frac{1-\gamma}{\log\frac{1}{\gamma}}. \quad (14)$$

For the large K setting, what the above suggests is that current CSIT offers an initial DoF boost of $d^*(\gamma = 0, \alpha) = \alpha$ (cf. [33]), which is then supplemented by a DoF gain

$$d(\gamma, \alpha) - d^*(\gamma = 0, \alpha) \to (1-\alpha)\frac{1-\gamma}{\log(\frac{1}{\gamma})}$$

attributed to the synergy between delayed CSIT and caching. These synergistic gains (see also [34]) are accentuated for smaller values of γ, where we see an exponential effect of coded caching, in the sense that now a microscopic $\gamma = e^{-G}$ can offer a substantial DoF boost

$$d(\gamma = e^{-G}, \alpha) - d(\gamma = 0, \alpha) \approx (1-\alpha)\frac{1}{G}. \quad (15)$$

Example 1: In a MISO BC system with $\alpha = 0$, K antennas and K users, in the absence of caching, the optimal per-user DoF is $d^*(\gamma = 0, \alpha = 0) = 1/H_K$ (cf. [3]) which vanishes to zero as K increases. A DoF of $1/4$ can be guaranteed with $\gamma \approx \frac{1}{50}$ for all K, a DoF of $1/7$ with $\gamma \approx \frac{1}{1000}$, and a DoF of $1/11.7$ can be achieved with $\gamma \approx 10^{-5}$, again for all K.

a) Interplay between CSIT quality and coded caching in the symmetric MISO BC: The derived form in (12) (and its approximation in (14)) nicely capture the synergistic as well as competing nature of feedback and coded caching. It is easy to see for example that the effect from coded-caching, reduces with α and is proportional to $1-\alpha$. This reflects the fact that in the symmetric MISO BC, feedback supports broadcasting by separating data streams, thus diminishing multi-casting by reducing the number of common streams. In the extreme case when $\alpha = 1$, we see — again for the symmetric MISO BC — that the caching gains are limited to local caching gains[4].

III. Cache-aided CSIT reductions

We proceed to explore how coded caching can alleviate the need for CSIT.

A. Cache-aided CSIT gains

To capture the cache-aided reductions on the CSIT load, let us consider

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

which is derived below in the form

$$\bar{\alpha}(\gamma, \alpha) = \alpha + \delta_\alpha(\gamma, \alpha)$$

for some $\delta_\alpha(\gamma, \alpha)$ that can be seen as the *CSIT reduction due to caching* (from $\bar{\alpha}(\gamma, \alpha)$ to the operational α).

[4] This conclusion is general (and not dependent on the specific schemes), because the used schemes are optimal for $\alpha = 1$. The statement holds because we can simply uniformly cache a fraction γ of each file in each cache, and upon request, use perfect-CSIT to zero-force the remaining requested information, to achieve the optimal $T^*(\gamma, \alpha = 1) = 1 - \gamma$, which leaves us with local (data push) caching gains only.

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

$$\bar{\alpha}(\gamma, \alpha) = \alpha + \frac{(1-\alpha)(H_{K\gamma} - \gamma H_K)}{(H_K - 1)(H_K - H_{K\gamma})} \tag{16}$$

is achievable, and implies a cache-aided CSIT reduction

$$\delta_\alpha(\gamma, \alpha) = \frac{(1-\alpha)(H_{K\gamma} - \gamma H_K)}{(H_K - 1)(H_K - H_{K\gamma})}.$$

Proof: The proof is direct from Theorem 2. ■

The above is made more insightful in the large K regime, for which we have the following.

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

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

which implies CSIT reductions of

$$\delta_\alpha(\gamma, \alpha) = (1-\alpha)d(\gamma, \alpha = 0) = (1-\alpha)\frac{1-\gamma}{\log(\frac{1}{\gamma})}.$$

Proof: The proof is direct from the definition of $\bar{\alpha}(\gamma, \alpha)$ and from Theorem 2. ■

Furthermore we have the following which quantifies the intuition that, with coded-caching, there is no reason to improve CSIT beyond a certain threshold quality. The following uses the definition in (8), and it holds for all K.

Corollary 2d: For any $\Gamma \in \{1, \ldots, K\}$, then

$$T^*(\gamma, \alpha) = T^*(\gamma, \alpha = 1) = 1 - \gamma \tag{18}$$

holds for any

$$\alpha \geq \alpha_{b,K-1} = \frac{K(1-\gamma) - 1}{(K-1)(1-\gamma)} \tag{19}$$

which reveals that CSIT quality $\alpha = \alpha_{b,K-1}$ is the maximum needed, as it already offers the same optimal performance $T^*(\gamma, \alpha = 1)$ that would be achieved if CSIT was perfect.

Proof: This is seen directly from Theorem 1 after noting that the achievable T matches $T^*(\gamma, \alpha = 1) = 1 - \gamma$. ■

b) How much caching is needed to partially substitute current CSIT with delayed CSIT (using coded caching to 'buffer' CSI): As we have seen, in addition to offering substantial DoF gains, the synergy between feedback and caching can also be applied to reduce the burden of acquiring current CSIT. What the above results suggest is that a modest γ can allow a BC system with D-CSIT to approach the performance attributed to current CSIT, thus allowing us to partially substitute current with delayed CSIT, which can be interpreted as an ability to buffer CSI. A simple calculation — for the large-K regime — can tell us that

$$\gamma'_\alpha \triangleq \arg\min_{\gamma'}\{\gamma' : d(\gamma', \alpha = 0) \geq d^*(\gamma = 0, \alpha)\} = e^{-1/\alpha}$$

which means that $\gamma'_\alpha = e^{-1/\alpha}$ suffices to achieve — in conjunction with delayed CSIT — the optimal DoF performance $d^*(\gamma = 0, \alpha)$ associated to a system with delayed CSIT and α-quality current CSIT.

Example 2: Let K be very large, and consider a BC system with delayed CSIT and α-quality current CSIT, where $\alpha = 1/5$. Then $\gamma'_{\alpha=1/5} = e^{-5} \approx 1/150$ which means that

$$d^*(\gamma \approx 1/150, \alpha = 0) \geq d^*(\gamma = 0, \alpha = 1/5)$$

which says that the same high-K per-user DoF performance $d^*(\gamma = 0, \alpha = 1/5)$, can be achieved by substituting all current CSIT with coded caching employing $\gamma \approx 1/150$.

IV. Cache-aided retrospective communications

We proceed to describe the achievable scheme, and in particular the process of placement, folding-and-delivery, and decoding. In the end we calculate the achievable duration T.

The caching part is modified from [1] to *'fold'* (linearly combine) the different users' data into multi-layered blocks, in a way such that the subsequent Q-MAT algorithm (cf. [33]) (specifically the last $K - \eta_\alpha$ ($\eta_\alpha \in \{\Gamma, \ldots, K-1\}$) phases of the QMAT algorithm) can efficiently deliver these blocks. Equivalently the algorithms are calibrated so that the caching algorithm creates a multi-destination delivery problem that is the same as that which is efficiently solved by the last stages of the QMAT- type communication scheme. We henceforth remove the subscript in η_α and simply use η, where now the dependence on α is implied.

A. Placement phase

We proceed with the placement phase which modifies on the work of [1] such that when the CSIT quality α increases, the algorithm caches a decreasing portion from each file, but does so with increasing redundancy. The idea is that the higher the α, the more private messages one can deliver directly without the need to multicast, thus allowing for some of the data to remain entirely uncached, which in turn allows for more redundancy across users' caches.

Here each file W_n ($|W_n| = f$ bits), is split into two parts

$$W_n = (W_n^c, W_n^{\bar{c}}), n = 1, 2, \ldots, N \tag{20}$$

where W_n^c (c for 'cached') will be placed into one or more 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 CSIT — in a manner that causes manageable interference and hence does not necessarily benefit from coded caching. The split is such that

$$|W_n^c| = \frac{KMf}{N\eta} \tag{21}$$

where $\eta \in \{\Gamma, \ldots, K-1\}$ is a positive integer, the value of which will be decided later on such that it properly regulates how much to cache from each W_n. Now for any specific η, we equally divide W_n^c into $\binom{K}{\eta}$ subfiles $\{W_{n,\tau}^c\}_{\tau \in \Psi_\eta}$,

$$W_n^c = \{W_{n,\tau}^c\}_{\tau \in \Psi_\eta} \tag{22}$$

where[5] $\Psi_\eta \triangleq \{\tau \subset [K] \ : \ |\tau| = \eta\}$ and each subfile has size

$$|W_{n,\tau}^c| = \frac{KMf}{N\eta\binom{K}{\eta}} = \frac{Mf}{N\binom{K-1}{\eta-1}} \text{ bits.} \tag{23}$$

[5] We recall that in the above, τ and $W_{n,\tau}^c$ are sets, thus $|\tau|, |W_{n,\tau}^c|$ denote cardinalities; $|\tau| = \eta$ means that τ has η different elements from $[K]$, while $|W_{n,\tau}^c|$ describes the size of $W_{n,\tau}^c$ in bits.

Now drawing from [1], the caches are filled as follows

$$Z_k = \{W^c_{n,\tau}\}_{n\in[N],\tau\in\Psi^{(k)}_\eta} \tag{24}$$

where $\Psi^{(k)}_\eta \triangleq \{\tau \in \Psi_\eta \ : \ k \in \tau\}$. Hence, each $W^c_{n,\tau}$ (thus each part of W^c_n) is repeated η times in the caches. As η increases with α, this means that CSIT allows for a higher redundancy in the caches; instead of content appearing in Γ different caches, it appears in $\eta \geq \Gamma$ caches instead, which will translate into multicast messages that are intended for more receivers.

B. Data folding

At this point, the transmitter becomes aware of the file requests $R_k, \forall k$, and must now deliver each requested file W_{R_k}, by delivering the constituent subfiles $\{W^c_{R_k,\tau}\}_{\tau\in\Psi_\eta\backslash\Psi^{(k)}_\eta}$ as well as $W^{\bar{c}}_{R_k}$, all to the corresponding receiver k. Recall that:

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

We assume the communication here has duration T. Thus for each k and a chosen η, we split each subfile $W^c_{R_k,\tau}$, $\tau \in \Psi_\eta\backslash\Psi^{(k)}_\eta$ (with size $|W^c_{R_k,\tau}| = \frac{Mf}{N\binom{K-1}{\eta-1}}$ (cf. (23))) into

$$W^c_{R_k,\tau} = [W^{c,f}_{R_k,\tau} \ \ W^{c,\bar{f}}_{R_k,\tau}] \tag{25}$$

where $W^{c,f}_{R_k,\tau}$ corresponds to information that appears in a cache somewhere and that will be eventually 'folded' (XORed) with other information, whereas $W^{c,\bar{f}}_{R_k,\tau}$ corresponds to information that is cached somewhere but that will not be folded with other information. The split yields

$$|W^{c,\bar{f}}_{R_k,\tau}| = \frac{f\alpha T - f(1-\frac{KM}{N\eta})}{\binom{K-1}{\eta}} \text{ (bits)} \tag{26}$$

where in the above, $f\alpha T$ represents the load for each user without causing interference during the delivery phase, where $f(1-\frac{KM}{N\eta})$ is the amount of uncached information, and where $|W^{c,f}_{R_k,\tau}| = |W^c_{R_k,\tau}| - |W^{c,\bar{f}}_{R_k,\tau}|$.

We proceed to fold cached content by creating linear combinations (XORs) from $\{W^{c,f}_{R_k,\tau}\}_{\tau\in\Psi_\eta\backslash\Psi^{(k)}_\eta}, \forall k$. We will use $P_{k,k'}(\tau)$ to be the function that replaces inside τ, the entry $k' \in \tau$, with the entry k. As in [1], if we deliver

$$W^{c,f}_{R_k,\tau} \oplus (\oplus_{k'\in\tau} \underbrace{W^{c,f}_{R_{k'},P_{k,k'}(\tau)}}_{\in Z_k}) \tag{27}$$

the fact that $W^{c,f}_{R_{k'},P_{k,k'}(\tau)} \in Z_k$, guarantees that receiver k can recover $W^{c,f}_{R_k,\tau}$, while at the same time guarantees that each other user $k' \in \tau$ can recover its own desired subfile $W^{c,f}_{R'_k,P_{k,k'}(\tau)} \notin Z_{k'}, \forall k' \in \tau$.

Hence delivery of each $W^{c,f}_{R_k,\tau} \oplus (\oplus_{k'\in\tau} W^{c,f}_{R_{k'},P_{k,k'}(\tau)})$ of size $|W^{c,f}_{R_k,\tau} \oplus (\oplus_{k'\in\tau} W^{c,f}_{R_{k'},P_{k,k'}(\tau)})| = |W^{c,f}_{R_k,\tau}|$ (cf. (23)), automatically guarantees delivery of $W^{c,f}_{R_{k'},P_{k,k'}(\tau)}$ to each user $k' \in \tau$, i.e., simultaneously delivers a total of $\eta+1$ distinct subfiles (each again of size $|W^{c,f}_{R_{k'},P_{k,k'}(\tau)}| = |W^{c,f}_{R_k,\tau}|$ bits) to $\eta+1$ distinct users. Hence *any*

$$X_\psi \triangleq \oplus_{k\in\psi} W^{c,f}_{R_k,\psi\backslash\{k\}}, \psi \in \Psi_{\eta+1} \tag{28}$$

— which is of the same form as in (27), and which is referred to here as an *order-*$(\eta+1)$ *folded message* — can similarly deliver to user $k \in \psi$, her requested file $W^{c,f}_{R_k,\psi\backslash k}$, which in turn means that each order-$(\eta+1)$ folded message X_ψ can deliver — with the assistance of the side information in the caches — a distinct, individually requested subfile, to each of the $\eta+1$ users $k \in \psi$ ($\psi \in \Psi_{\eta+1}$).

Thus to satisfy all requests $\{W_{R_k}\backslash Z_k\}_{k=1}^K$, the transmitter must deliver

- uncached messages $W^{\bar{c}}_{R_k}, \forall k$
- cached but unfolded messages $\{W^{c,\bar{f}}_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\eta+1}}, \forall k$
- and the entire set

$$\mathcal{X}_\Psi \triangleq \{X_\psi = \oplus_{k\in\psi} W^{c,f}_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\eta+1}} \tag{29}$$

consisting of $|\mathcal{X}_\Psi| = \binom{K}{\eta+1}$ folded messages of order-$(\eta+1)$, each of size (cf. (26),(23))

$$|X_\psi| = |W^{c,f}_{R_k,\tau}| = \frac{f(1-\gamma-\alpha T)}{\binom{K-1}{\eta}} \text{ (bits).} \tag{30}$$

C. Transmission

We now focus on the transmission of the aforementioned messages by adapting the QMAT algorithm from [33],

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

$$\boldsymbol{x}_t = \mathbf{G}_{c,t}\boldsymbol{x}_{c,t} + \sum_{\ell\in\bar{\mathcal{S}}} \boldsymbol{g}_{\ell,t} a^*_{\ell,t} + \sum_{k=1}^{K} \boldsymbol{g}_{k,t} a_{k,t} \tag{31}$$

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 $k, k \in [K]$, is simultaneously orthogonal to the CSI estimate for the channels of all other users ($\boldsymbol{g}_{l,t}$ acts the same), thus guaranteeing $\hat{\boldsymbol{h}}^T_{k',t}\boldsymbol{g}_{k,t} = 0, \ \ \forall k' \in [K]\backslash k$.

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 [6].

[6] We will henceforth avoid going into the details of the QMAT scheme. Some aspects of this scheme are similar to MAT, 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. An element that is hidden from the presentation here is that, while the QMAT scheme has many rounds, and while decoding spans more than one round, we will — in a slight abuse of notation — focus on describing just one round, which we believe is sufficient for the purposes of this paper here.

We will allocate the rates 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 we will allocate power such that

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

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

Remark 1: Recall that instead of employing matrix notation, after normalization, we use the concept of signal duration $\text{dur}(\boldsymbol{x})$ required for the transmission of some vector $\boldsymbol{x}$. We also note that due to time normalization, the time index $t \in [0, T]$, need not be an integer.

For any α, our scheme will be defined by an integer $\eta \in [\Gamma, K-1] \cap \mathbb{Z}$, which will be chosen as

$$\eta = \arg \max_{\eta' \in [\Gamma, K-1]\cap\mathbb{Z}} \{\eta' \ : \ \alpha_{b,\eta'} \leq \alpha\} \tag{32}$$

for

$$\alpha_{b,\eta} = \frac{\eta - \Gamma}{\Gamma(H_K - H_\eta - 1) + \eta}. \tag{33}$$

η will define the amount of cached information that will be folded ($\{W^{c,f}_{R_k,\tau}\}_{\tau\in\Psi_\eta\backslash\Psi_\eta^{(k)}}$), and thus also the amount of cached information that will not be folded ($\{W^{c,\overline{f}}_{R_k,\tau}\}_{\tau\in\Psi_\eta\backslash\Psi_\eta^{(k)}}$) and which will be exclusively carried by the different $a_{k,t}$. In all cases,

- all of $\{X_\psi\}_{\psi\in\Psi_{\eta+1}}$ (which are functions of the cached-and-to-be-folded $\{W^{c,f}_{R_k,\tau}\}_{\tau\in\Psi_\eta\backslash\Psi_\eta^{(k)}}$) will be exclusively carried by $\boldsymbol{x}_{c,t},\ t \in [0, T]$, while
- all of the uncached $W^{\overline{c}}_{R_k}$ (for each $k = 1, \ldots, K$) and all of the cached but unfolded $\{W^{c,\overline{f}}_{R_k,\tau}\}_{\tau\in\Psi_\eta\backslash\Psi_\eta^{(k)}}$ will be exclusively carried by $a_{k,t}, t \in [0, T]$.

Transmission of $\{X_\psi\}_{\psi\in\Psi_{\eta+1}}$: From [33], we know that the transmission relating to $\boldsymbol{x}_{c,t}$ can be treated independently from that of $a_{k,t}$ with the assistance of $a^*_{\ell,t}$, simply because $a_{k,t}$ do not actually interfere with decoding of $\boldsymbol{x}_{c,t}$, as a result of the scheme, and as a result of the chosen power and rate allocations which jointly adapt to the CSIT quality α. For this reason, we can treat the transmission of $\boldsymbol{x}_{c,t}$ separately.

Hence we first focus on the transmission of $\{X_\psi\}_{\psi\in\Psi_{\eta+1}}$, which will be sent using $\boldsymbol{x}_{c,t},\ t \in [0, T]$ using the last $K-\eta$ phases of the QMAT algorithm in [33] corresponding to having the ZF symbols $a_{k,t}$ set to zero. For simplicity, we will label these phases starting from phase $\eta+1$ and terminating in phase K. Each phase $j = \eta+1, \ldots, K$ aims to deliver order-j folded messages (cf. (29)), and will do so gradually: phase j will deliver (in addition to other information) $N_j \triangleq (K-j+1)\binom{K}{j}$ order-j messages which carry information that has been requested by j users, and in doing so, it will generate $N_{j+1} \triangleq j\binom{K}{j+1}$ signals that are combinations of received signals from $j+1$ different users, and where these N_{j+1} signals will be conveyed in the next phase $j+1$. In the last phase $j = K$, fully common symbols that are useful and decoded by all users will be sent, thus allowing each user to go back and retroactively decode the data of phase $j = K-1$, which will then be used to decode the data in phase $j = K-2$ and so on, until they reach phase $j = \eta+1$ (first phase) which will end the task. We proceed to describe these phases and T_j denotes the duration of phase j.

Phase $\eta+1$: In this first phase of duration $T_{\eta+1}$, the information in $\{X_\psi\}_{\psi\in\Psi_{\eta+1}}$ is delivered by $\boldsymbol{x}_{c,t},\ t \in [0, T_{\eta+1}]$, which can also be rewritten in the form of a sequential transmission of shorter-duration K-length vectors

$$\boldsymbol{x}_\psi = [x_{\psi,1}, \ldots, x_{\psi,K-\eta}, 0, \ldots, 0]^T \tag{34}$$

for different ψ, where each vector $\boldsymbol{x}_\psi$ carries exclusively the information from each X_ψ, and where this information is uniformly split among the $K-\eta$ independent scalar entries $x_{\psi,i},\ i = 1, \ldots, K-\eta$, each carrying

$$\frac{|X_\psi|}{(K-\eta)} = \frac{f(1-\gamma-\alpha T)}{\binom{K-1}{\eta}(K-\eta)} \tag{35}$$

bits (cf. (30)). Hence, given that the allocated rate for $\boldsymbol{x}_{c,t}$ (and thus the allocated rate for each $\boldsymbol{x}_\psi$) is $(1-\alpha)f$, we have that the duration of each $\boldsymbol{x}_\psi$ is

$$\text{dur}(\boldsymbol{x}_\psi) = \frac{|X_\psi|}{(K-\eta)(1-\alpha)f}. \tag{36}$$

Given that $|\mathcal{X}_\Psi| = \binom{K}{\eta+1}$, then

$$T_{\eta+1} = \binom{K}{\eta+1}\text{dur}(\boldsymbol{x}_\psi) = \frac{\binom{K}{\eta+1}|X_\psi|}{(K-\eta)(1-\alpha)f}. \tag{37}$$

After each transmission of $\boldsymbol{x}_\psi$, the received signal $y_{k,t},\ t \in [0, T_{\eta+1}]$ of desired user k ($k \in \psi$) takes the form

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\mathbf{G}_{c,t}\boldsymbol{x}_{c,t}}_{L_{\psi,k},\text{power} \ \doteq \ P} + \underbrace{\boldsymbol{h}^T_{k,t}\sum_{\ell\in\psi}\boldsymbol{g}_{\ell,t}a^*_{\ell,t}}_{\doteq \ P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{P^{\alpha}} \tag{38}$$

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

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a^*_{k,t}}_{\text{power} \ \doteq \ P} + \underbrace{\boldsymbol{h}^T_{k,t}\sum_{\substack{\ell\in\psi\\ \ell\neq k}}\boldsymbol{g}_{\ell,t}a^*_{\ell,t} + \underbrace{\boldsymbol{h}^T_{k,t}\mathbf{G}_{c,t}\boldsymbol{x}_{c,t}}_{L_{\psi,k}, \doteq \ P^{1-\alpha}}}_{i_{\psi,k}, \doteq \ P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{P^{\alpha}} \tag{39}$$

where in both cases, we ignored the Gaussian noise and the ZF noise up to P^0. Each user $k \in [K]$ receives a linear combination $L_{\psi,k}$ of the transmitted $K-\eta$ symbols $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\eta}$. Next the transmitter will somehow send an additional $K-\eta-1$ signals $L_{\psi,k},\ k \in [K]\backslash\psi$ (linear combinations of $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\eta}$ as received — up to noise level — at each user $k' \in [K]\backslash\psi$) which will help each user $k \in \psi$ resolve the already sent $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\eta}$. This will be done in the next phase $j = \eta+2$.

Phase $\eta+2$: The challenge now is for signals $\boldsymbol{x}_{c,t},\ t \in (T_{\eta+1}, T_{\eta+1} + T_{\eta+2}]$ to convey all the messages of the form

$$i_{\psi,k},\ \forall k \in [K]\backslash\psi,\ \forall\psi \in \Psi_{\eta+1}$$

to each receiver $k \in \psi$. Note that $\boldsymbol{h}_{k,t}^T \sum_{\substack{\ell \in \psi \\ \ell \neq k}} \boldsymbol{g}_{\ell,t} a_{\ell,t}^*$ is the residual interference of the previous round which can be removed easily and each of the above linear combinations, is now — during this phase — available (up to noise level) at the transmitter. Let

$$\Psi_{\eta+2} = \{\psi \in [K] \ : \ |\psi| = \eta + 2\}$$

and consider for each $\psi \in \Psi_{\eta+2}$, a transmitted vector

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

which carries the contents of $\eta{+}1$ $(l = 1, \cdots, \eta{+}1)$ different elements

$$f_l = (\bar{i}_{\psi\backslash\{k\},k} \oplus \bar{i}_{\psi\backslash\{k'\},k'}), k \neq k', k, k' \in \psi$$

where $\bar{i}_{\psi\backslash\{P_k\},P_k}$ is the quantization of $i_{\psi\backslash\{P_k\},P_k}$ from phase 1. f_l are predetermined and known at each receiver. The transmission of $\{\boldsymbol{x}_\psi\}_{\forall\psi\in\Psi_{\eta+2}}$ is sequential.

It is easy to see that there is a total of $(\eta+1)\binom{K}{\eta+2}$ XORs in the form of f_l, and each can be taken as an order-$(\eta+2)$ signal intended for $\eta + 2$ receivers in ψ. Using this, and following the same steps used in phase $\eta + 1$, we have that

$$T_{\eta+2} = \binom{K}{\eta+2} \text{dur}(\boldsymbol{x}_\psi) = T_{\eta+1} \frac{\eta+1}{\eta+2}. \tag{40}$$

We now see that for each ψ, each receiver $k \in \psi$ recalls their own observation $i_{\psi\backslash\{k\},k}$ from the previous phase, and removes it from f_l, thus now being able to acquire the $\eta + 1$ independent linear combinations $\{L_{\psi\backslash\{k\},k}\}_{\forall k\in\psi\backslash\{k\}}$ by easily removing the auxiliary symobls. The same holds for each other user $k \in \psi$.

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

Phase j $(\eta + 3 \leq j \leq K)$: Generalizing the described approach to any phase $j = \eta{+}3, \ldots, K$, we will use $\boldsymbol{x}_{c,t},\ t \in [\sum_{i=\eta+1}^{j-1} T_i, \sum_{i=\eta+1}^{j} T_i]$ to convey all the messages of the form

$$i_{\psi,k},\ \forall k \in [K]\backslash\psi,\ \forall\psi \in \Psi_{j-1}$$

to each receiver $k \in \psi$. For each $\psi \in \Psi_j \triangleq \{\psi \in [K] \ : \ |\psi| = j\}$, the transmitted vector

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

will carry the contents of $j - 1$ different XORs $f_l, l = 1, \ldots, j-1$ of the j elements $\{\bar{i}_{\psi\backslash\{k\},k}\}_{\forall k\in\psi}$ created by the transmitter. After the sequential transmission of $\{\boldsymbol{x}_\psi\}_{\forall\psi\in\Psi_j}$, each receiver k can obtain the $j{-}1$ independent linear combinations $\{L_{\psi\backslash\{k\},k}\}_{\forall k\in\psi\backslash\{k\}}$ again by removing the auxiliary symbols. The same holds for each other user $k' \in \psi$. As with the previous phases, we can see that

$$T_j = T_{\eta+1}\frac{\eta+1}{j},\ j = \eta+3, \ldots, K. \tag{41}$$

This process finishes at phase $j = K$, during which each

$$\boldsymbol{x}_\psi = [x_{\psi,1}, 0, 0, \ldots, 0]^T$$

carries a single scalar that is decoded easily by all. Based on this, backwards decoding will allow for users to retrieve $\{X_\psi\}_{\psi\in\Psi_{\eta+1}}$. This is described afterwards. In treating the decoding part, we briefly recall that each $a_{k,t},\ k = 1, \ldots, K$ carries (during period $t \in [0, T]$), all of the uncached $W_{R_k}^{\bar{c}}$ and all of the unfolded $\{W_{R_k,\tau}^{c,\bar{f}}\}_{\tau\in\Psi_\eta\backslash\Psi_\eta^{(k)}}$.

D. Decoding

The whole transmission lasts $K - \eta$ phases. For each phase $j = \eta + 1, \cdots, K$ and the corresponding ψ, the received signal $y_{k,t},\ t \in [\sum_{i=\eta+1}^{j-1} T_i, \sum_{i=\eta+1}^{j} T_i]$ of desired user k $(k \in \psi)$ takes the same form as in (38), while the received signal for the other users $k \in [K]\backslash\psi$ takes the same form as in (39). As we see in [33], after each phase, $i_{\psi,k}$ is first quantized with $(1 - 2\alpha)^+ \log P$ bits, which results in a residual quantizaton noise $n_{\psi,k}$ with power scaling as P^α.

Then, the transmitter quantizes the quantization noise $n_{\psi,k}$ with an additional $\alpha \log P$ bits, which will be carried by the auxiliary data symbols in the corresponding phase in the next round (here we 'load' this round with additional requests from the users). In this way, we can see that the 'common' signal $\boldsymbol{x}_{c,t}$ can also be decoded at user $k \in [K]\backslash\psi$ with the assistance of an auxiliary data symbol from the next round. After this, each user k removes $\boldsymbol{h}_{k,t}^T \mathbf{G}_{c,t} \boldsymbol{x}_{c,t}$ from their received signals, and readily decode their private symbols $a_{k,t},\ t \in [0, T]$, thus allowing for retrieval of their own unfolded $\{W_{R_k,\psi\backslash\{k\}}^{c,\bar{f}}\}_{\psi\in\Psi_{\eta+1}}$ and uncached $W_{R_k}^{\bar{c}}$.

In terms of decoding the common information, as discussed above, each receiver k will perform a backwards reconstruction of the sets of overheard equations

$$\begin{gathered}\{L_{\psi,k},\ \forall k \in [K]\backslash\psi\}_{\forall\psi\in\Psi_K} \\ \vdots \\ \downarrow \\ \{L_{\psi,k},\ \forall k \in [K]\backslash\psi\}_{\forall\psi\in\Psi_{\eta+2}}\end{gathered}$$

until phase $\eta + 2$. At this point, each user k has enough observations to recover the original $K - \eta$ symbols $x_{\psi,1}, x_{\psi,2}, \ldots, x_{\psi,K-\eta}$ that fully convey X_ψ, hence each user k can reconstruct their own set $\{W_{R_k,\psi\backslash\{k\}}^{c,f}\}_{\psi\in\Psi_{\eta+1}}$ which, combined with the information from the $a_{k,t},\ t = [0, T]$ allow for each user k to reconstruct $\{W_{R_k,\psi\backslash\{k\}}^{c}\}_{\psi\in\Psi_{\eta+1}}$ which is then combined with Z_k to allow for reconstruction of the requested file W_{R_k}.

E. Calculation of T

To calculate T, we recall from (41) that

$$T = \sum_{j=\eta+1}^{K} T_j = T_{\eta+1} \sum_{j=\eta+1}^{K} \frac{\eta+1}{j} = (\eta+1)(H_K - H_\eta)T_{\eta+1} \tag{42}$$

which combines with (35) and (37) to give

$$T = \frac{(K-\Gamma)(H_K - H_\eta)}{(K-\eta) + \alpha(\eta + K(H_K - H_\eta - 1))} \tag{43}$$

as stated in Theorem 1. The bound by $T = 1 - \gamma$ seen in the theorem, corresponds to the fact that the above expression (43) applies, as is, only when $\alpha \leq \alpha_{b,K-1} = \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)}$ which corresponds to $\eta = K-1$ (where X_ψ are fully common messages, directly desired by all), for which we already get the best possible $T = 1-\gamma$, and hence there is no need to go beyond $\alpha = \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)}$.

V. Conclusions

This work studied the previously unexplored interplay between coded-caching and feedback quality and timeliness. This is motivated by the fact that CSIT and coded caching are two powerful ingredients that are hard to obtain, and by the fact that these ingredients are intertwined in a synergistic and competing manner. In addition to the substantial cache-aided DoF gains revealed here, the results suggest the interesting practical ramification that distributing predicted content 'during the night', can offer continuous amelioration of the load of predicting and disseminating CSIT 'during the day'.

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] A. G. Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling a conjecture by Lapidoth, Shamai and Wigger on the collapse of degrees of freedom under finite precision CSIT," *CoRR*, vol. abs/1403.1541, 2014. [Online]. Available: http://arxiv.org/abs/1403.1541

[3] 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.

[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] 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," Jan. 2013, available on arXiv: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," Jul. 2013, in *ISIT13*, available on arXiv:1302.0806.

[10] 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.

[11] 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.

[12] 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.

[13] M. Torrellas, A. Agustin, and J. Vidal, "Retrospective interference alignment for the MIMO interference broadcast channel," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015.

[14] A. Bracher and M. Wigger, "Feedback and partial message side-information on the semideterministic broadcast channel," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015.

[15] S. Lashgari, R. Tandon, and S. Avestimehr, "Three-user MISO broadcast channel: How much can CSIT heterogeneity help?" in *Proc. IEEE International Conference on Communications (ICC)*, June 2015, pp. 4187–4192.

[16] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *CoRR*, vol. abs/1504.01452, 2015. [Online]. Available: http://arxiv.org/abs/1504.01452

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

[18] 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.

[19] 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

[20] 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

[21] 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

[22] 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.

[23] 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*.

[24] 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

[25] 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), 2015 IEEE 16th International Workshop on*, June 2015, pp. 655–659.

[26] 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

[27] M. Gatzianas, L. Georgiadis, and L. Tassiulas, "Multiuser broadcast erasure channel with feedback — capacity and algorithms," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5779–5804, Sept 2013.

[28] S. P. Shariatpanahi, A. S. Motahari, and B. H. Khalaj, "Multi-server coded caching," *CoRR*, vol. abs/1503.00265, 2015. [Online]. Available: http://arxiv.org/abs/1503.00265

[29] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *CoRR*, vol. abs/1605.00203, 2016. [Online]. Available: http://arxiv.org/abs/1605.00203

[30] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *CoRR*, vol. abs/1606.03175, 2016. [Online]. Available: http://arxiv.org/abs/1606.03175

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

[32] 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

[33] 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

[34] J. Zhang and P. Elia, "The synergistic gains of coded caching and delayed feedback," *CoRR*, vol. abs/1604.06531, Apr. 2016. [Online]. Available: http://arxiv.org/abs/1604.06531