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 symmetric broadcast channel with random fading and imperfect feedback, and analyzes the throughput performance as a function of feedback statistics and cache size. In this setting, this paper identifies the optimal cache-aided degrees-of-freedom (DoF) within a factor of 4, by identifying near-optimal schemes that exploit a new synergy between coded caching and delayed CSIT, as well as by exploiting the unexplored interplay between caching and feedback-quality. The DoF expressions reveal an initial gain due to current CSIT, and an additional gain due to coded caching, which is exponential in the sense that any linear decrease in the required DoF performance, allows for an exponential reduction in the required cache size. In the end, this paper reveals three new aspects of caching: a synergy between memory and delayed feedback, a tradeoff between memory and current CSIT, and a powerful ability to provide cache-aided feedback savings.**

***Index Terms*— Coded Caching, Prefetching, Broadcast channel, Feedback, Channel State Information at the Transmitter (CSIT), degrees-of-freedom (DoF), MIMO, Cache Memory.**

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 — which went beyond storing popular content closer to the user — involved two phases; the placement phase (during off peak hours) and the delivery phase (during peak hours). During the placement phase, content that was predicted to be popular (a library of commonly requested files), was coded and placed across user's caches. During the delivery phase — which started when users requested specific files from the predicted library of files — the transmitter encoded across different users' requested data content, taking into consideration the requests and 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 the aforementioned traditional 'data push' methods that only pre-store 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 and feedback timeliness. 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.

A. 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 the available 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

The work was supported by the European Research Council under the European Union's Horizon 2020 research and innovation program under Grant 725929.

The authors are with the Communication Systems Department, EURECOM, Sophia Antipolis 06410, France (e-mail: jingjing.zhang@eurecom.fr; elia@eurecom.fr).

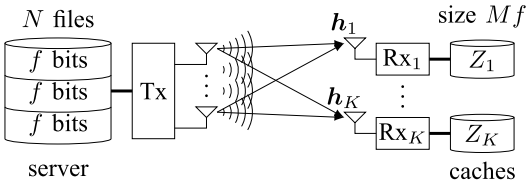

Fig. 1. Cache-aided K-user MISO BC.

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.

1) Cache-Aided Broadcast Channel Model:

a) K-user BC with pre-filled caching: In the symmetric K-user multiple-input single-output (MISO) broadcast channel of interest here, the K-antenna transmitter, communicates to K single-antenna receiving users. The transmitter has access to a library of $N \geq K$ distinct files $W_1, W_2, \ldots, W_N$, each of size $|W_n| = f$ bits. Each user $k \in \{1, 2, \ldots, K\}$ has a cache Z_k, of size $|Z_k| = Mf$ bits, where naturally $M \leq N$. Communication consists of the aforementioned *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 $\{W_n\}_{n=1}^N$. The delivery phase commences when each user k requests from the transmitter, any *one* file $W_{R_k} \in \{W_n\}_{n=1}^N$, out of the N library files. Each file can be 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

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

where the latter simply means that the sum of the sizes of the caches across all users, is Γ times the volume of the N-file library. As in [1], we will first consider the case where $\Gamma = \{1, 2, \cdots K\}$, while for non integer Γ, the result will be that corresponding to $\lfloor \Gamma \rfloor$.

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

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

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 their desired file W_{R_k}.

2) Coded Caching and CSIT-Type Feedback: Communication also takes place in the presence of channel state information at the transmitter. CSIT-type feedback is typically of imperfect-quality as it is hard to obtain in a timely and reliable manner. In the high-SNR (high P) regime of interest, this current-CSIT quality is concisely represented in the form of the normalized quality exponent [2], [3]

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

where $\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k$ denotes the Gaussian estimation error between the current CSIT estimate $\hat{\boldsymbol{h}}_k$ and the estimated channel $\boldsymbol{h}_k$. In this sense, the variance (power) of the error scales as $P^{-\alpha}$. The range of interest[1] is $\alpha \in [0, 1]$. We also assume availability of delayed CSIT (as in for example [6], as well as in a variety of subsequent works [2], [3], [7]–[14], see also [12], [13], [15], [16] as well as [17]–[19]) 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.

a) Motivating the mixed-CSIT model: As it is argued in [2], this mixed CSI model (partial current CSIT, and good-quality delayed CSIT) nicely captures different realistic settings that might involve channel correlations and an ability to improve CSI as time progresses. The idea is that fast feedback (that is received well within the coherence period) might be of reduced refinement, which is though improved later with additional delayed feedback. Another way to motivate this setting is by recalling that in practice, current CSI is obtained from prediction using the delayed CSIT.

In addition, the mixed CSI model is well suited for cache-aided communications, as it explicitly reflects feedback timeliness and feedback quality which are both (as we will see in this paper) directly intertwined with coded caching. Considering mixed-CSIT is important because the delayed-and-current CSI combination captures aspects that relate feedback to caching; delayed CSIT will introduce a synergy (with coded caching), while the current-CSIT quality (corresponding to the parameter α) will introduce a tradeoff.

b) Intuitive links between feedback-quality and caching (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 wireless 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 transmitted, which may (in some cases) diminish the utility of feedback which primarily aims to facilitate the

[1] In the high SNR regime of interest here, $\alpha = 0$ corresponds to having essentially no current CSIT (cf. [4]), while having $\alpha = 1$ corresponds (again in the high SNR regime) to perfect and immediately available CSIT (cf. [5]).

opposite which is the transmission of private information. It is for example easy to see (we will see this later) that in the presence of $\Gamma = K - 1$, there is no need for CSIT in order to achieve the optimal performance.

3) 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 proceed to 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]

4) Notation and Assumptions: We will use the notation $H_n \triangleq \sum_{i=1}^{n} \frac{1}{i}$, to represent the n-th harmonic number, and we will use $\epsilon_n \triangleq H_n - \log\ (n)$ to represent its logarithmic approximation error, for some integer n. We remind the reader that ϵ_n decreases with n, and that $\epsilon_\infty \triangleq \lim_{n\to\infty} H_n - \log\ (n)$ 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 $[K] \triangleq \{1, 2, \cdots, K\}$. 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. In our high-P setting of interest, we will also use $\doteq$ to denote *exponential equality*, i.e., we will write $g(P) \doteq P^B$ to denote $\lim_{P\to\infty} \frac{\log_2\ g(P)}{\log_2\ P} = B$. Similarly $\dot{\geq}$ and $\dot{\leq}$ will denote exponential inequalities. Logarithms are of base e, unless we use $\log_2(\cdot)$ which will represent a logarithm of base 2.

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. Additional basic assumptions regarding the outer bound, can be found in Appendix A.

5) Prior Work: The benefits of coded caching on reducing interference and improving performance, were revealed in the seminal work by Maddah-Ali and Niesen [1] who considered a caching system where a server is connected to multiple users through a shared link, and designed a novel caching and delivery method that jointly offers a multicast gain that helps mitigate the link load, and which was proven to have a gap from optimal that is at most 12. This work was subsequently generalized in different settings, which included the setting of different cache sizes for which Wang *et al.* in [24] developed a variant of the algorithm in [1] which achieves a gap of at most 12 from the information theoretic optimal. Other extensions included the work in [25] by Maddah-Ali and Niesen who considered the setting of decentralized caching where the achieved performance was shown to be comparable to that of the centralized case [1], despite the lack of coordination in content placement. For the same original single-stream setting of [1], the work of Ji *et al.* in [26] considered a scenario where users make multiple requests each, and proposed a scheme that has a gap to optimal that is less than 18. Again for the setting in [1], the work of Ghasemi and Ramamoorthy in [27], derived tighter outer (lower) bounds that improve upon existing bounds, and did so by recasting the bound problem as one of optimally labeling the leaves of a directed tree. Further work can be found in [28] where Wang *et al.* explored the interesting link between caching and distributed source coding with side information. Interesting conclusions are also drawn in the work of Ajaykrishnan *et al.* in [29], which revealed that the effectiveness of caching in the single stream case, is diminished when N approaches and exceeds K^2.

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 [30], considered a cache-aided wireless fading BC where each user experiences a different link quality, and proposed a suboptimal communication scheme that is based on time- and frequency-division and power- and bandwidth-allocation, and which was evaluated using numerical simulations to eventually show that the produced throughput decreases as the number of users increases. Further work by Timo and Wigger in [31] 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 [32] 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.

Different work has also considered the effects of caching in different non-classical channel paradigms. One of the earlier such works that focused on practical wireless network settings, includes the work by Golrezaei *et al.* in [33], which considered

[2]We note that setting $f = \log_2(P)$ 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.

[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.

a downlink cellular setting where the base station is assisted by helper nodes that jointly form a wireless distributed caching network (no coded caching) where popular files are cached, resulting in a substantial increase to the allowable number of users by as much as $400-500\%$. In a somewhat related setting, the work by Perabathini *et al.* [34] accentuated the energy efficiency gains from caching. Further work by Ji *et al.* [35] derived the limits of so-called combination caching networks 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. Additional work can also be found in [36] where Niesen *et al.* considered a cache-aided network where each node is randomly located inside a square, and it requests a message that is available in different caches distributed around the square. Further related work on caching can be found in [37] and [26], [38]–[41].

Work that combines caching and feedback considerations in wireless networks, has only just recently started. A reference that combines these, can be found in [42] 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. An even more recent concurrent work can be found in [43] 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 [44] — 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. The first work that considers the actual interplay between coded caching and CSIT quality, can be found in [45] which considered the easier problem of how the optimal cache-aided performance (with coded caching), can be achieved with reduced quality CSIT.

6) Outline and Contributions: In Section II, Lemma 1, we offer a lower bound for the optimal $T^*(\gamma, \alpha)$. Then in Theorem 1 we calculate the achievable $T(\gamma, \alpha)$, for $\Gamma \in \{1, 2, \cdots, K\}$, $\alpha \in [0, 1]$, and prove it to be less than four times the optimal, thus identifying the optimal $T^*(\gamma, \alpha)$ within a factor of 4. A simpler expression for T (again within a factor of 4 from optimal), and its corresponding per-user DoF, are derived in Theorem 2, while a simple approximation of these is derived in Corollary 2a, where we see that the per-user DoF takes the form $d(\gamma, \alpha) = \alpha + (1-\alpha)\frac{1-\gamma}{\log \frac{1}{\gamma}}$, revealing that even a very small $\gamma = e^{-G}$ can offer a substantial DoF boost which, as K increases, tends to $d(\gamma = e^{-G}, \alpha) - d(\gamma = 0, \alpha) \approx (1-\alpha)\frac{1}{G}$.

In Section III we discuss practical implications. In Corollary 2d we describe the savings in current CSIT that we can have due to coded caching, while in Corollary 2e we quantify the intuition that, in the presence of coded-caching, there is no reason to improve CSIT beyond a certain threshold quality. We also show that caching with approximately $\gamma^{-\frac{1}{\alpha}}$ allows us to entirely remove current CSIT without degrading performance, basically describing the memory cost for buffering CSI.

In Section IV we present four simple examples that offer some intuition on the scheme design, while in Section V we present the caching-and-delivery scheme in its general form, building on the interesting connections between MAT-type retrospective transmission schemes (cf. [6]) and coded caching. Appendix A presents the outer bound proof, and Appendix B the proof for the gap to optimal.

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 and $\alpha \in [0, 1]$, where $\gamma = \frac{M}{N}$ and $\Gamma = K\gamma$. The results hold for $N \geq K$, except the following outer bound (lower bound) on the optimal T^*, which in fact holds for all N, K.

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

$$T^*(\gamma, \alpha) \geq \max_{s \in \{1, \ldots, \min(N,K)\}} \frac{1}{(H_s \alpha + 1 - \alpha)} \left(H_s - \frac{Ms}{\lfloor \frac{N}{s} \rfloor}\right). \tag{7}$$

Proof: The proof is presented in Section VI and it uses the bound from Lemma 2 whose proof can be found in Section VI-A. ∎

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

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

to define different α intervals, while when $\Gamma \geq K-1$, we set $\alpha_{b,\eta} = 0$ reflecting the fact that no CSIT is needed to achieve the optimal performance.

Theorem 1: In the (K, M, N, α) cache-aided MISO BC with N files, K 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

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

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 V, while the corresponding gap to optimal is bounded in Section VI-A. ∎

The above is achieved with a general scheme whose caching phase is a function of α. We will henceforth 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), simpler expressions, 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 $\Gamma \in \{1, 2, \cdots, K\}$,

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

is achievable and has a gap from optimal

$$\frac{T}{T^*} < 4 \tag{11}$$

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

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

Proof: The scheme that achieves the above performance will be described later on as a special (simpler) case of the scheme corresponding to Theorem 1. The corresponding gap to optimal is bounded in Section VI-A. ∎

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 setting where $\frac{H_K - H_\Gamma}{\log(\frac{1}{\gamma})} = 1$.

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

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

and the derived DoF takes the form

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

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. [46]), 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.[4]

1) 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.[5]

For the specific case of $\alpha = 0$, we have the following.

Corollary 2b: In the (K, M, N) cache-aided MISO BC with $K \leq N$ users, and with $\Gamma \in \{1, 2, \cdots, K-1\}$, then

$$T = H_K - H_\Gamma \tag{15}$$

is achievable and has a gap-to-optimal

$$\frac{T}{T^*} < 4 \tag{16}$$

that is less than 4, for all K.

Proof: The scheme that achieves the above performance is presented in Section V, while the corresponding gap to the optimal performance is bounded in the appendix Section VI-A. ∎

Furthermore, the following corollary offers some insight by adopting the logarithmic approximation $H_n \approx \log\ (n)$ (which becomes tight as K increases).[6]

Corollary 2c: Under the logarithmic approximation, the above T takes the form

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

and the corresponding per-user DoF takes the form

$$d(\gamma) = \frac{1-\gamma}{\log(\frac{1}{\gamma})}. \tag{17}$$

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. [6]) 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.

B. Synergistic DoF Gains

We proceed to derive some insight from the above, and for this we look to the large K regime, where there is no ambiguity on which gains can be attributed solely to coded caching (in addition to possible DoF gains due to other resources such as feedback). In this regime, what the above says is that the gain that is directly attributed to caching

$$d(\gamma) - d^*(\gamma = 0) \to \frac{1-\gamma}{\log(\frac{1}{\gamma})} > \gamma, \quad \forall \gamma \in (0, 1]$$

can substantially exceed[7] the typical coded-caching (per-user DoF) gain γ.

What we also see, again for larger K, is that while the individual component settings/algorithms (MAT from [6], and the Maddah-Ali and Niesen (MN) algorithm from [1]) respectively provided individual DoF gains of the form $d_{\text{MAT}} = d^*(\gamma = 0) = \frac{1}{H_K}$ and $d_{\text{SS}}(\gamma) = \frac{1-\gamma}{\frac{K(1-\gamma)}{1+K\gamma}} = \gamma + \frac{1}{K}$

[4]We note that these gains are, as K increases, less and less a result of the extra performance boost directly from D-CSIT, because in the large K setting, the per-user DoF due to delayed feedback — without caching — is approximately $\frac{1}{\log K}$ which vanishes to zero.

[5]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.

[6]To avoid confusion, we clarify that the main theorem is simply a DoF-type result, that nothing but SNR scales to infinity, and the derived DoF holds for all K. The corollaries are simply the approximation of the above expression, under the logarithmic approximation, which becomes tight as K increases.

[7]In this larger K setting, we have $d_{\text{SS}}(\gamma) + d_{\text{MAT}} \to \gamma$. We clarify that this step is simply the result of a large-K approximation of the corresponding expression from the main theorem. In that sense, K scales after SNR does. We also recall from [6] that $d^*(\gamma = 0) = \frac{1}{H_K}$ which decreases with K.

(cf. [1]), the combination of these two components results in a synergistic

$$d(\gamma) > d_{\rm SS}(\gamma) + d_{\rm MAT}, \quad \forall \gamma \in [0,1]$$

that — for larger K — exceeds the sum of the two individual components. This is the first time that such synergistic gains have been recorded. The gains become very striking for smaller values of γ in which case we have that $\frac{1-\gamma}{\log(\frac{1}{\gamma})} \gg \gamma$.

1) Derivative Analysis for Understanding the Small-γ Gains Attributed to Caching: Let us fix K, and consider the derivative of the DoF gain attributed to caching

$$d(\gamma) - d(\gamma = 0) = \frac{1-\gamma}{H_K - H_{K\gamma}} - \frac{1}{H_K} \approx \frac{1-\gamma}{\log(1/\gamma)} - \frac{1}{H_K} \tag{18}$$

which takes the form

$$\frac{\delta\big(d(\gamma) - d(\gamma=0)\big)}{\delta\gamma} \approx \frac{\frac{1}{\gamma} - 1 - \log(\frac{1}{\gamma})}{(\log(\frac{1}{\gamma}))^2} \approx \frac{\frac{1}{\gamma}}{(\log(\frac{1}{\gamma}))^2} \tag{19}$$

which, when evaluated at $\gamma = 1/K$, gives

$$\frac{\delta\big(d(\gamma) - d(\gamma=0)\big)}{\delta\gamma}|_{\gamma=\frac{1}{K}} \approx \frac{K}{\log^2 K}$$

revealing a substantial DoF boost at the early stages[8] of γ.

These can be compared to linear gains where the derivative is constant

$$\frac{\delta\big(d(\gamma) - d(\gamma=0)\big)}{\delta\gamma} = \frac{\delta(\gamma)}{\delta\gamma} = 1, \quad \forall\gamma. \tag{20}$$

These gains in fact imply[9] an exponential (rather than linear) effect of coded caching, in the sense that now a microscopic $\gamma = e^{-G}$ can offer a very satisfactory

$$d(\gamma = e^{-G}) \approx \frac{1}{G} \tag{21}$$

which is only a factor G from the interference-free (cache-free) optimal $d = 1$. The above only needs that $K \geq e^G$ for any fixed $G \geq 1$. It does not require K to be asymptotically large. Naturally the higher the K, the more of these gains can be attributed solely to caching (rather than MAT). When the value of K is moderate, naturally MAT has an impact, in terms of per-user DoF.

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 feedback savings, let us consider

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

[8]Similarly for $\gamma = K^{-(1-\epsilon)}, \epsilon \in (0,1]$, we get $\frac{\delta\big(d(\gamma)-d(\gamma=0)\big)}{\delta\gamma}|_{\gamma=K^{-(1-\epsilon)}} \approx \frac{K^{1-\epsilon}}{(1-\epsilon)^2\log^2 K}$.

[9]Here we make the assumption that $1-\gamma \approx 1$, which is a soft approximation that allows for simplicity of expressions, and which reflects the reality of small γ (cf. [47]).

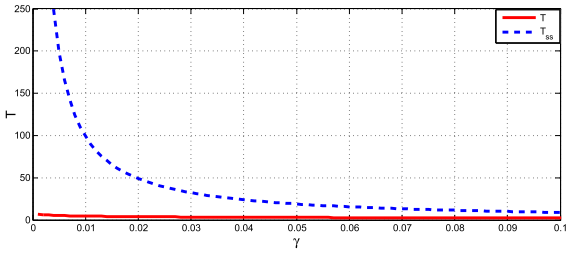

Fig. 2. Single stream $T_{\rm SS}$ (no delayed CSIT, dotted line) vs. T after the introduction of delayed CSIT. Plot holds even for very large K, and the main gains appear for smaller values of γ.

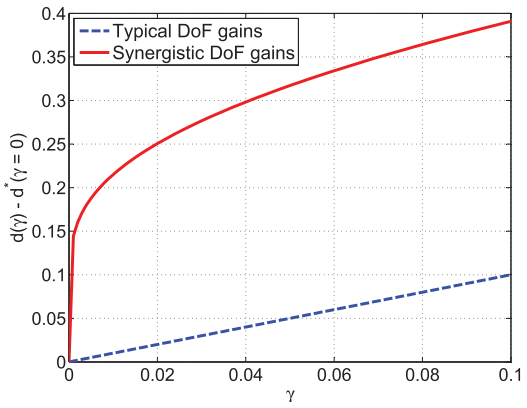

Fig. 3. Typical gain $d(\gamma) - d^*(\gamma = 0)$ attributed solely to coded caching (dotted line) vs. synergistic gains derived here. Plot holds for large K, and the main gains appear for smaller values of γ.

describing the *CSIT reduction due to caching*, down to an operational α. The proof is direct from Theorem 2.

Corollary 2d: In the (K, M, N, α) cache-aided MISO BC, caching can achieve a CSIT reduction

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

which, under the logarithmic approximation, takes the form

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

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

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

$$T^*(\gamma,\alpha) = T^*(\gamma,\alpha=1) = 1-\gamma \tag{22}$$

holds for any

$$\alpha \geq \alpha_{b,K-1} = \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)} \tag{23}$$

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 $T^*(\gamma,\alpha=1) = 1-\gamma$. ■

1) 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 $\gamma'_{\alpha=1/5} = e^{-5} = 0.0067 \approx 1/150$ which means that

$$d^*(\gamma = 0.0067, \alpha = 0) \geq d^*(\gamma = 0, \alpha = 1/5)$$

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. Examples of Schemes

The general scheme will be presented in Section V. To offer some intuition on the design, we provide here different examples (all for the case of $K = N = 3, M = 1$), first for the case of $\alpha = 1$, then for the case of $\alpha = 0$, then a third example for the general case of $\alpha \in (0, 1)$ corresponding to Theorem 2, and then a fourth example again for the general case of $\alpha \in (0, 1)$, now though for the case corresponding to Theorem 1, where the caching redundancy can increase with α.

In our examples here, for simplicity, the three distinct files in the library will be relabeled as $W_1 = A, W_2 = B, W_3 = C$, and we will assume the worst-case request where A, B, C are requested by user 1, 2, 3, respectively.

A. Scheme for $\alpha = 1$

We offer this very simple example as a warm up exercise. For cache placement, each user stores a fraction $\gamma = \frac{M}{N} = \frac{1}{3}$ of each file, and then (upon notification of the three requests) the remaining $f(1-\gamma) = \frac{2f}{3}$ bits of each desired file are delivered using interference-free zero-forcing (ZF) which employs perfect CSIT. Hence after an optimal duration $T = (1-\gamma) = \frac{2}{3}$, the transmitter delivers file A to user 1, B to user 2 and C to user 3. In this symmetric setting, the complete separation of signals due to perfect-CSIT, renders multicasting unnecessary, and the optimal performance reflects only local caching gains.

B. Scheme for $\alpha = 0$

Now we focus on the case where only delayed CSIT is available at the transmitter.

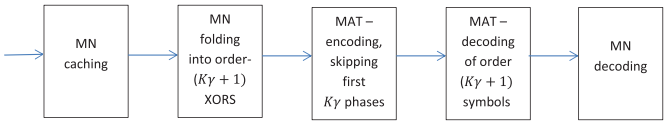

Fig. 4. Basic composition of scheme. 'MAT encoding/decoding' corresponds to the scheme in [6], while 'MN caching/folding' corresponds to the scheme in [1].

1) Key Idea Behind the Scheme: As Figure 4 implies, the scheme starts by first applying the Maddah-Ali and Niesen (MN) sub-packetization based scheme [1] for placing sub-packets in the caches, and for generating order-$(K\gamma+1)$ (i.e., order-2) messages — in the form of XORs of the sub-packets — where each XOR is meant for $K\gamma + 1 = 2$ users. These XORs are delivered by the well known MAT method [6], and in particular by the part of MAT which delivers order-$(K\gamma+1)$ (order-2) messages. This allows us to skip the first $K\gamma$ phases (i.e., to skip the first phase) of the MAT scheme, which happen to have the longest time duration (phase i has duration $1/i$). This gives us an idea as to why the impact of small caches (small γ) is substantial; even small caches can remove a large fraction of the communication duration. Upon MAT decoding, we simply proceed with decoding based on the algorithm in [1]. In the end, the key idea is 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 MAT scheme.

a) Placement phase: After splitting each file into three equally-sized subfiles as $A = (A_1, A_2, A_3), B = (B_1, B_2, B_3), C = (C_1, C_2, C_3)$, we fill the cache Z_k of each user k, as follows $Z_k = (A_k, B_k, C_k), k = 1, 2, 3$.

b) Delivery phase: To satisfy the requests A, B, C, we must deliver the following three XORs $A_2 \oplus B_1, A_3 \oplus C_1, B_3 \oplus C_2$, each having size $\frac{f}{3}$ bits, and each intended for two users (users 1-2, 1-3, and 2-3 respectively). These messages, which we respectively denote as AB, AC, BC, are delivered by employing the last two phases of the $(K = 3)$ MAT algorithm in [6].

Phase 2: Before transmission, we split each XOR into two *mini parts* as $AB = (AB_1, AB_2), AC = (AC_1, AC_2), BC = (BC_1, BC_2)$, where now each mini part has size $\frac{f}{6}$ bits. Then we form the following three vectors[10]

$$\boldsymbol{x}_1 = \begin{bmatrix} AB_1 \\ AB_2 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} AC_1 \\ AC_2 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix} BC_1 \\ BC_2 \\ 0 \end{bmatrix} \tag{24}$$

which we send in three sequential transmissions. This allows each user to get three different linear combinations of scalars, one per transmission, as we indicate and label below (ignoring noise)

user 1 gets: $L_1(AB_1, AB_2), L_4(AC_1, AC_2), L_7(BC_1, BC_2)$
user 2 gets: $L_2(AB_1, AB_2), L_5(AC_1, AC_2), L_8(BC_1, BC_2)$
user 3 gets: $L_3(AB_1, AB_2), L_6(AC_1, AC_2), L_9(BC_1, BC_2)$

where for example $L_1(AB_1, AB_2)$, $L_4(AC_1, AC_2)$, $L_7(BC_1, BC_2)$ denote the received observations of user 1,

[10]Here we assume a mapping from bits to QAM.

at time slot 1, 2, 3 respectively. The important thing to note is that — as we know from [6] — $L_7(BC_1, BC_2)$ is useful to both users 2 and 3 in decoding BC_1, BC_2, and similarly $L_5(AC_1, AC_2)$ is useful to both user 1 and 3, and $L_3(AB_1, AB_2)$ is useful to users 1 and 2. Recall that $|L_1(AB_1, AB_2)| = |AB_1| = \overline{6}^{f}$ bits, which means that $\text{dur}(\boldsymbol{x}_i) = \frac{1}{6}, i = 1, 2, 3$, and that this phase has duration $\frac{3}{6}$.

Phase 3: We now transmit, using one antenna only, first a linear combination

$$f_1\big(L_3(AB_1, AB_2),\ L_5(AC_1, AC_2),\ L_7(BC_1, BC_2)\big)$$

and then a second combination

$$f_2\big(L_3(AB_1, AB_2),\ L_5(AC_1, AC_2),\ L_7(BC_1, BC_2)\big)$$

both of which carry a fully-common message (i.e., a message of order 3) that is useful to all users. Thus after the sequential transmission of

$$\boldsymbol{x}_4 = [f_1, 0, 0]^T, \quad \boldsymbol{x}_5 = [f_2, 0, 0]^T \tag{25}$$

each user can decode. To see this, let us focus on user 1. Before transmission of $\boldsymbol{x}_4, \boldsymbol{x}_5$, user 1 had knowledge of $L_7(BC_1, BC_2)$ (as this was its received signal during the third transmission, for BC). Now, with $\boldsymbol{x}_4, \boldsymbol{x}_5$, user 1 has two observations regarding L_3, L_5, L_7. L_7 can be removed, hence now user 1 can resolve $L_3(AB_1, AB_2)$ and $L_5(AC_1, AC_2)$. Thus now user 1 can combine $L_1(AB_1, AB_2)$ and $L_3(AB_1, AB_2)$ to resolve AB_1 and AB_2, and recover $A_2 \oplus B_1$. Similarly user 1 can combine $L_4(AC_1, AC_2)$ and $L_5(AC_1, AC_2)$ to resolve AC_1 and AC_2, to thus recover $A_3 \oplus C_1$. User 1 can now combine $A_2 \oplus B_1$ with its cache (which includes B_1) to recover A_2, and can combine $A_3 \oplus C_1$ with its cache (which includes C_1) to recover A_3, and thus recover the desired A. Similarly users 2 and 3 can recover files B and C respectively.

With $\text{dur}(\boldsymbol{x}_4) = \text{dur}(\boldsymbol{x}_5) = \frac{1}{6}$, the third phase has duration $\frac{2}{6}$ and the total two-phase transmission has an overall duration

$$T = \frac{1}{2} + \frac{1}{3} = \frac{5}{6} \tag{26}$$

which matches the derived $T(\gamma) = T(\frac{1}{3}) = H_K - H_{K\gamma} = H_3 - H_2 = \frac{5}{6}$.

C. Scheme for $\alpha \in (0, 1)$

We proceed with the general case of $\alpha \in (0, 1)$, again focusing on the case corresponding to Theorem 2 where η is forced to be $\eta = \Gamma$, for all α (in this case, $\eta = 1$).

1) Placement Phase: The cache placement is the same as in the case of $\alpha = 0$ described above.

2) Data Folding: Recall that users 1,2,3 respectively request files A, B, C, which will be delivered with delay T. Also recall that now, user 1 requires subfiles A_2, A_3, user 2 subfiles B_1, B_3, and user 3 subfiles C_1, C_2. Each of these subfiles will be split into two mini parts as $A_2 = (A_2^f, A_2^{\bar{f}}), A_3 = (A_3^f, A_3^{\bar{f}})$, $B_1 = (B_1^f, B_1^{\bar{f}}), B_3 = (B_3^f, B_3^{\bar{f}})$ and $C_1 = (C_1^f, C_1^{\bar{f}}), C_2 = (C_2^f, C_2^{\bar{f}})$, such that $|A_2^{\bar{f}}| = |A_3^{\bar{f}}| = |B_1^{\bar{f}}| = |B_3^{\bar{f}}| = |C_1^{\bar{f}}| = |C_2^{\bar{f}}| = \frac{f\alpha T}{2}$ bits. Now the three generated XORs will be $AB \triangleq A_2^f \oplus B_1^f$, $AC \triangleq A_3^f \oplus C_1^f$, and $BC \triangleq B_3^f \oplus C_2^f$, and will deliver the 'folded' part of each missing subfile, while the remaining 'unfolded' parts $A_2^{\bar{f}}, A_3^{\bar{f}}, B_1^{\bar{f}}, B_3^{\bar{f}}, C_1^{\bar{f}}, C_2^{\bar{f}}$ will be delivered via a ZF component inside the employed QMAT scheme, as we describe below.

3) Transmission: We proceed to describe the transmission of the aforementioned folded and unfolded messages, employing the last two phases of the three-user QMAT algorithm from [46].

Phase 2: This phase will take as input the folded messages (which will be decoded upon completion of the third phase later on), and will also deliver some of the unfolded messages. For any $K\gamma$-length set (in this case, for any pair) $\psi \in \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$, and for $\bar{\psi} \triangleq \{1, 2, 3\} \backslash \psi$, the transmission takes the form

$$\boldsymbol{x}_t = \mathbf{G}_{\psi,t}\boldsymbol{x}_{\psi,t} + \boldsymbol{g}_{\bar{\psi},t}a^*_{\bar{\psi},t} + \boldsymbol{g}_{1,t}a_{1,t} + \boldsymbol{g}_{2,t}a_{2,t} + \boldsymbol{g}_{3,t}a_{3,t}. \tag{27}$$

where

- $\mathbf{G}_{\psi,t} \triangleq [\boldsymbol{g}_{\bar{\psi},t}, \mathbf{U}_{\psi,t}]$, for $\boldsymbol{g}_{\bar{\psi},t}$ being simultaneously orthogonal to the channel estimates of the two users in ψ, and for $\mathbf{U}_\psi \in \mathbb{C}^{3\times 2}$ being a randomly chosen sub-unitary matrix
- $\boldsymbol{g}_{k,t}$ is orthogonal to the current estimates of the channels to users $\{1, 2, 3\}\backslash k$
- $\boldsymbol{x}_{\psi,t} = [x_{\psi,1}, x_{\psi,2}, 0]^T$ is a $(K = 3)$-length vector with $K - K\gamma = 2$ non-zero scalar entries, where each scalar has rate $(1-\alpha)\log P$ bits per unit time, and where each scalar has power $\mathbb{E}[|x_{\psi,1}|^2] \doteq P$, $\mathbb{E}[|x_{\psi,2}|^2] \doteq P^{1-\alpha}$
- the bits in XOR AB ($\psi = \{1, 2\}$) are split evenly between $x_{\{1,2\},1}$ and $x_{\{1,2\},2}$, the bits in XOR AC ($\psi = \{1, 3\}$) are split evenly between $x_{\{1,3\},1}$ and $x_{\{1,3\},2}$, and the bits in XOR BC ($\psi = \{2, 3\}$) are split evenly between $x_{\{2,3\},1}$ and $x_{\{2,3\},2}$
- irrespective of ψ, $a_{k,t}$ is the ZF symbol carrying the unfolded messages for user $k \in \{1, 2, 3\}$, with the rate $\alpha \log P$ bits per unit time, and with power P^α
- $a^*_{\bar{\psi},t}$ is an auxiliary symbol intended for user $\bar{\psi}$, carrying residual interference.[11] The symbol has power P, and rate $\min\ (1-\alpha, \alpha)\log P$ bits per unit time. Auxiliary symbols allow for the simultaneous delivery of private data (unfolded messages) and higher-order data (XORs)
- Transmission $\boldsymbol{x}_t$ is sequential: first for $\psi = \{1, 2\}$, then for $\psi = \{1, 3\}$, and then for $\psi = \{2, 3\}$.

For any given ψ, the received signal (noise is removed) $y_{k,t}$ of each desired user $k \in \psi$, takes the form

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\mathbf{G}_{\psi,t}\boldsymbol{x}_{\psi,t}}_{\triangleq L_{\psi,k},\ \text{power}\doteq P} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{\bar{\psi},t}a^*_{\bar{\psi},t}}_{\doteq P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{\doteq P^\alpha} \tag{28}$$

while the received signal for the other user $\bar{\psi}$, takes the form

$$y_{\bar{\psi},t} = \underbrace{\boldsymbol{h}^T_{\bar{\psi},t}\boldsymbol{g}_{\bar{\psi},t}a^*_{\bar{\psi},t}}_{\text{power}\ \doteq P} + \underbrace{\boldsymbol{h}^T_{\bar{\psi},t}\mathbf{G}_{\psi,t}\boldsymbol{x}_{\psi,t}}_{\triangleq i_{\psi,\bar{\psi}},\ P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{\bar{\psi},t}\boldsymbol{g}_{\bar{\psi},t}a_{\bar{\psi},t}}_{P^\alpha}. \tag{29}$$

[11] The interference is carried over from a previous round of the QMAT scheme. We spare the reader some of the details regarding rounds, and consider the scheme for just one round. The rounds are linked via the auxiliary variables, and having more than one round simply guarantees the QMAT DoF optimality, as it minimizes the cost of initialization.

It is easy to see that at the end of this phase, each user in ψ needs one more observation to resolve $x_{\psi,1}, x_{\psi,2}$, hence the overheard messages $i_{\psi,\bar{\psi}} \triangleq \boldsymbol{h}_{\psi,t}^T \mathbf{G}_{\psi,t} \boldsymbol{x}_{\psi,t}$ will be quantized and placed into a message that will be meant for $K\gamma + 1 = 3$ users, and which will be delivered in the next phase.

To calculate the duration of this second phase, we recall that the phase terminates when we transmit $AB = A_2^f \oplus B_1^f$, $AC = A_3^f \oplus C_1^f$, and $BC = B_3^f \oplus C_2^f$. Given that $|A_2^{\bar{f}}| = |A_3^{\bar{f}}| = |B_1^{\bar{f}}| = |B_3^{\bar{f}}| = |C_1^{\bar{f}}| = |C_2^{\bar{f}}| = \frac{f\alpha T}{2}$ bits (by design, as we have seen above), then $|AB| = |AC| = |BC| = \frac{1}{3} - \frac{f\alpha T}{2}$. Given that the rate of each transmitted scalar $x_{\psi,1}$ and $x_{\psi,2}$ is $(1-\alpha)\log P$ bits per unit time, and given that we are transmitting all three XORs AB, AC, BC, then the duration of the second phase is $3\frac{\frac{1}{3} - \frac{f\alpha T}{2}}{2(1-\alpha)}$.

Phase 3: We use $\bar{i}_{\psi,\bar{\psi}}$ to denote the quantized version of $i_{\psi,\bar{\psi}}$ (from phase 2) which can be reconstructed by the transmitter at the beginning of phase 3. Instead of creating two linear combinations as in (25), here we use the XOR operator to combine messages: we create $f_1 \triangleq \bar{i}_{12,3} \oplus \bar{i}_{13,2}$ and $f_2 \triangleq \bar{i}_{13,2} \oplus \bar{i}_{23,1}$. The transmission then takes the form

$$\boldsymbol{x}_t = [x_{c,t}, 0, 0]^T + \sum_{k=1}^{3} \boldsymbol{g}_{k,t} a_{k,t} \tag{30}$$

where $x_{c,t}$ carries information from f_1 and f_2, has power P, rate $(1-\alpha)\log P$ bits per unit time, and finally where $a_{k,t}$ is again the ZF symbols with power P^α and rate $\alpha \log P$ bits per unit time, as before.

4) Decoding: We first start by noting that by design of the QMAT algorithm (cf. [46]), each user $k \in \psi$ — using information from the previous round — can decode the auxiliary variable $a^*_{\psi,t}$ from (28) and remove it from its received signal. Similarly the undesired user $\bar{\psi}$ can employ successive interference cancelation (cf. (29)) to again remove $a^*_{\psi,t}$. Thus we can proceed from (28) and (29), without having to consider the auxiliary variables.

Now each user can decode $x_{c,t}$ and its own private (ZF) messages in phase 3, and can then go back to phase 2 to decode the XORs and the private messages. To see this, we again focus on user 1 who already knows $L_{\{1,2\},1}, L_{\{1,3\},1}, \bar{i}_{\{2,3\},1}$ (cf. (28)), where $L_{\psi,k} \triangleq \boldsymbol{h}_{k,t}^T \mathbf{G}_{\psi,t} \boldsymbol{x}_{\psi,t}$ (here, since we focus on user 1, we set $k = 1$).

From the common message $x_{c,t}$, user 1 knows $f_1 = \bar{i}_{\{1,2\},3} \oplus \bar{i}_{\{1,3\},2}$ and $f_2 = \bar{i}_{\{1,3\},2} \oplus \bar{i}_{\{2,3\},1}$. Using its knowledge of $\bar{i}_{\{2,3\},1}$ and f_2, user 1 can get $\bar{i}_{\{1,3\},2}$, and then from f_1 the user can also have $\bar{i}_{\{1,2\},3}$. Now user 1 combines $\bar{i}_{\{1,2\},3}$ with $L_{\{1,2\},1}$ to get $x_{\{1,2\},1}$ and $x_{\{1,2\},2}$ which allows for resolving $AB = A_2^f \oplus B_1^f$. Then user 1 combines $\bar{i}_{\{1,3\},2}$ with $L_{\{1,3\},1}$ to get $x_{\{1,3\},1}$ and $x_{\{1,3\},2}$ which allows for resolving $AC = A_3^f \oplus C_1^f$. Using B_1^f from Z_1 yields A_2^f, and using C_1^f from Z_1 yields A_3^f. Combined with the ZF-transmitted private data which delivers $A_2^{\bar{f}}$ and $A_3^{\bar{f}}$, completes the delivery of A_2, A_3 and thus of A.

To calculate the duration of the third phase, we recall that $x_{c,t}$ carries f_1 and f_2, each of size $|f_1| = |f_2| = \frac{\frac{1}{3} - \frac{f\alpha T}{2}}{2}$. Thus $x_{c,t}$ carries a total of $\frac{1}{3} - \frac{f\alpha T}{2}$ bits, at a rate of $(1-\alpha)\log P$ bits per unit time. Hence the total duration of phase 3 is $\frac{\frac{1}{3} - \frac{f\alpha T}{2}}{1-\alpha}$. Combined with the duration $3\frac{\frac{1}{3} - \frac{f\alpha T}{2}}{2(1-\alpha)}$ of phase 2, implies a total duration of

$$T = \frac{10}{12 + 3\alpha} \tag{31}$$

which matches the derived expression $T = \frac{(1-\gamma)(H_K - H_\Gamma)}{\alpha(H_K - H_\Gamma) + (1-\alpha)(1-\gamma)} = \frac{\frac{2}{3}(H_3 - H_1)}{\alpha(H_3 - H_1) + (1-\alpha)\frac{2}{3}} = \frac{10}{12+3\alpha}$ from Theorem 2.

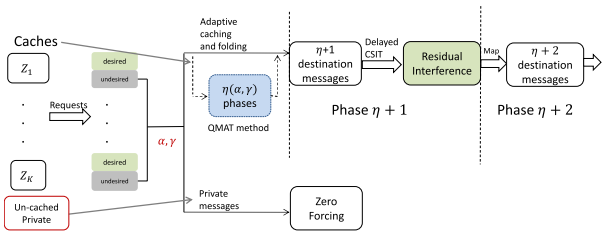

Fig. 5. Cache-aided retrospective communications scheme.

D. Scheme Which Adapts the Caching Redundancy η to α

In the previous example, we considered the case corresponding to Theorem 2 where the caching redundancy η is fixed as $\eta = K\gamma$. This incurs a certain (albeit small) degree of suboptimality, because as we argue in the next section, a higher α can allow for higher caching redundancy because more private messages means reduced multicasting, which allows some of the data to remain uncached, which in turn allows for more copies of the same information across different users' caches. Here we give an example of the general scheme that captures this interplay between η and α, and show the corresponding improvement that is found in Theorem 1 over Theorem 2. The description of the example focuses on showing how we calibrate — as a function of α — the cache placement and the process of creating the XORs. This is again presented for the case of $K = N = 3$, $M = 1$.

Using the defined breaking points $\alpha_{b,\eta} = \frac{\eta - \Gamma}{\Gamma(H_K - H_{\eta-1}) + \eta}$, $\eta = \lceil \Gamma \rceil, \ldots, K-1$ in (8), the range of α is split into two intervals. The first interval is $\alpha \in [0, \alpha_{b,2}) = [0, \frac{3}{4}]$ during which the scheme remains the same as in the previous example, where we chose $\eta = K\gamma = 1$ to create XORs that were meant for two users at a time. In the second interval $\alpha \geq \alpha_{b,2} = \frac{3}{4}$, we set $\eta = 2$, and each XOR is meant for $\eta + 1 = 3$ users, allowing us to skip the second phase of the previous example.

1) Placement Phase: We first split each file as

$$A = (A^c, A^{\bar{c}}), \quad B = (B^c, B^{\bar{c}}), \quad C = (C^c, C^{\bar{c}}), \tag{32}$$

where A^c, B^c, C^c denote the cached parts, and $A^{\bar{c}}, B^{\bar{c}}, C^{\bar{c}}$ the parts that are not cached. The split is such that $|A^c| = |B^c| = |C^c| = f\frac{K\gamma}{\eta} = \frac{f}{2}$. Then each cached part is again divided evenly into $\binom{K}{\eta} = 3$ mini parts as $A^c = (A_{12}, A_{13}, A_{23})$, $B^c = (B_{12}, B_{13}, B_{23})$, $C^c = (C_{12}, C_{13}, C_{23})$,

and then the caches are filled as

$$Z_1 = A_{12}, A_{13}, B_{12}, B_{13}, C_{12}, C_{13},$$
$$Z_2 = A_{12}, A_{23}, B_{12}, B_{23}, C_{12}, C_{23},$$
$$Z_3 = A_{13}, A_{23}, B_{13}, B_{23}, C_{13}, C_{23}. \quad (33)$$

Now, to satisfy the requests A, B, C for users $1, 2, 3$ respectively, we must send $A_{23} \oplus B_{13} \oplus C_{12}$ as well as the uncached messages $A^{\bar{c}}, B^{\bar{c}}, C^{\bar{c}}$. This will be achieved by employing phase $\eta + 1 = 3$ of the scheme we saw in the previous example, where the transmission again takes the form $\boldsymbol{x}_t = [x_{c,t}, 0, 0]^T + \sum_{k=1}^{3} \boldsymbol{g}_{k,t} a_{k,t}$, where $x_{c,t}$ carries $A_{23} \oplus B_{13} \oplus C_{12}$ (again with power P and rate $(1-\alpha)\log P$ bits per unit time), while $a_{1,t}, a_{2,t}, a_{3,t}$ respectively carry $A^{\bar{c}}, B^{\bar{c}}, C^{\bar{c}}$ (each with power P^{α} and rate $\alpha \log P$ bits per unit time).

This adaptation of η as a function of α provides for a slightly better performance, which now — for any $\alpha \geq \frac{3}{4}$ — takes the form $T = 1 - \gamma = \frac{2}{3}$, which is the interference-free optimal, despite having imperfect CSIT, as this was discussed in Corollary 2e. This performance is an improvement over the previously derived $T = \frac{10}{12+3\alpha}$ for all $\alpha \geq \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)} = \frac{3}{4}$.

V. Cache-Aided Retrospective Communications: The General Case

We proceed to describe the general communication scheme, and in particular the process of placement, folding-and-delivery, and decoding that achieve the performance described in Theorem 1. In the end we calculate the achievable duration T. We remind the reader that, in the following, forcing η to the (slightly) suboptimal $\eta = \Gamma$, delivers the result in Theorem 2 (for which we saw some examples above).

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 transmission algorithm (cf. [46]) (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.

The intuition is that as α increases, we can have more private data, which means that there is less to be cached, which means that caching can have higher redundancy, which implies XORs of higher order, which means that we can multicast to more users at a time, which in turn means that we can skip more phases of QMAT. The intensity of the impact of small values of γ relates to the fact that the early phases of QMAT are the longest. So while a small γ can only skip a few phases, it nonetheless manages to substantially reduce delay.

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.

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

$$W_n = (W_n^c, W_n^{\overline{c}}) \quad (34)$$

where W_n^c (c for 'cached') will be placed into one or more caches, while the content of $W_n^{\overline{c}}$ ($\overline{c}$ 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

$$|W_n^c| = \frac{KMf}{N\eta} \quad (35)$$

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. As we will see later, η will increase with α, and it will reflect the degree of caching redundancy; cached content will appear in $\eta \geq \Gamma$ caches.

Now for any specific η, we equally divide W_n^c into $\binom{K}{\eta}$ subfiles $\{W_{n,\tau}^c\}_{\tau \in \Psi_\eta}$,

$$W_n^c = \{W_{n,\tau}^c\}_{\tau \in \Psi_\eta} \quad (36)$$

where[12]

$$\Psi_\eta \triangleq \{\tau \subset [K] \ : \ |\tau| = \eta\} \quad (37)$$

where each subfile has size

$$|W_{n,\tau}^c| = \frac{KMf}{N\eta\binom{K}{\eta}} = \frac{Mf}{N\binom{K-1}{\eta-1}} \text{ bits.} \quad (38)$$

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

$$Z_k = \{W_{n,\tau}^c\}_{n \in [N], \tau \in \Psi_\eta^{(k)}} \quad (39)$$

where

$$\Psi_\eta^{(k)} \triangleq \{\tau \in \Psi_\eta \ : \ k \in \tau\}. \quad (40)$$

Hence each subfile $W_{n,\tau}^c$ is stored in Z_k as long as $k \in \tau$, which means that each $W_{n,\tau}^c$ (and thus each part of W_n^c) is repeated η times in the caches.

B. Data Folding

At this point, the transmitter becomes aware of the file requests $R_k, k = 1, \ldots, K$, and must now deliver each requested file W_{R_k}, by delivering the constituent subfiles $\{W_{R_k,\tau}^c\}_{\tau \in \Psi_\eta \backslash \Psi_\eta^{(k)}}$ as well as $W_{R_k}^{\overline{c}}$, all to the corresponding receiver k. We quickly recall that:

1) subfiles $\{W_{R_k,\tau}^c\}_{\tau \in \Psi_\eta^{(k)}}$ are already in Z_k;
2) subfiles $\{W_{R_k,\tau}^c\}_{\tau \in \Psi_\eta \backslash \Psi_\eta^{(k)}}$ are directly requested by user k, but are not cached in Z_k;
3) subfiles $Z_k \backslash \{W_{R_k,\tau}^c\}_{\tau \in \Psi_\eta^{(k)}} = Z_k \backslash W_{R_k}^c$ 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 $W_{R_k,\tau}^c$,

[12] 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.

$\tau \in \Psi_\eta \backslash \Psi_\eta^{(k)}$ (each of size $|W^c_{R_k,\tau}| = \frac{Mf}{N\binom{K-1}{\eta-1}}$ as we saw in (38)) into

$$W^c_{R_k,\tau} = [W^{c,f}_{R_k,\tau} \quad W^{c,\overline{f}}_{R_k,\tau}] \tag{41}$$

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 $W^{c,\overline{f}}_{R_k,\tau}$ corresponds to information that is cached somewhere but that will not be folded with other information. The split yields

$$|W^{c,\overline{f}}_{R_k,\tau}| = \frac{f\alpha T - f(1-\frac{KM}{N\eta})}{\binom{K-1}{\eta}} \tag{42}$$

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 $|W^{c,f}_{R_k,\tau}| = |W^c_{R_k,\tau}| - |W^{c,\overline{f}}_{R_k,\tau}|$.

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], the idea is that if we deliver

$$W^{c,f}_{R_k,\tau} \oplus (\oplus_{k'\in\tau} \underbrace{W^{c,f}_{R_{k'},P_{k,k'}(\tau)}}_{\in Z_k}) \tag{43}$$

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. (38)), 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*

$$X_\psi \triangleq \oplus_{k\in\psi} W^{c,f}_{R_k,\psi\backslash\{k\}}, \quad \psi \in \Psi_{\eta+1} \tag{44}$$

— which is of the same form as in (43), 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 $W^{\overline{c}}_{R_k}$, $k = 1, \ldots, K$
- cached but unfolded messages $\{W^{c,\overline{f}}_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\eta+1}}$, $k = 1, \ldots, K$
- and the entire set

$$\mathcal{X}_\Psi \triangleq \{X_\psi = \oplus_{k\in\psi} W^{c,f}_{R_k,\psi\backslash\{k\}}\}_{\psi\in\Psi_{\eta+1}} \tag{45}$$

consisting of

$$|\mathcal{X}_\Psi| = \binom{K}{\eta+1} \tag{46}$$

folded messages of order-$(\eta+1)$, each of size (cf. (42),(38))

$$\begin{aligned} |X_\psi| &= |W^{c,f}_{R_k,\tau}| = |W^c_{R_k,\tau}| - |W^{c,\overline{f}}_{R_k,\tau}| \\ &= \frac{f(1-\gamma-\alpha T)}{\binom{K-1}{\eta}} \text{ (bits).} \end{aligned} \tag{47}$$

C. Transmission

We proceed to describe the transmission of the aforementioned messages by adapting the QMAT algorithm from [46], with delay T.

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

$$\boldsymbol{x}_t = \mathbf{G}_{c,t}\boldsymbol{x}_{c,t} + \sum_{\ell\in\bar{\mathcal{S}}} \boldsymbol{g}_{\ell,t} a^*_{\ell,t} + \sum_{k=1}^K \boldsymbol{g}_{k,t} a_{k,t} \tag{48}$$

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

$$\hat{\boldsymbol{h}}^T_{k',t}\boldsymbol{g}_{k,t} = 0, \quad \forall k' \in [K]\backslash k. \tag{49}$$

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.[13]

Throughout communication

- we will allocate power such that

$$\begin{aligned} &\mathbb{E}\{|\boldsymbol{x}_{c,t}|_1^2\} \doteq \mathbb{E}\{|a^*_{k,t}|^2\} \doteq P, \\ &\mathbb{E}\{|\boldsymbol{x}_{c,t}|_{i\neq 1}^2\} \doteq P^{1-\alpha}, \quad \mathbb{E}\{|a_{k,t}|^2\} \doteq P^\alpha \end{aligned} \tag{50}$$

  where $|\boldsymbol{x}_{c,t}|_i, i = 1, 2, \cdots, K$, denotes scalar i in vector $\boldsymbol{x}_{c,t}$, and we will allocate rate such that

[13] Whenever possible, we will henceforth avoid going into the details of the Q-MAT scheme. Some aspects of this scheme are similar to MAT, and a main new element is that Q-MAT applies digital transmission of interference, and a double-quantization method that collects and distributes residual interference across different rounds (this is here carried by $a^*_{k,t}$), in a manner that allows for ZF and MAT to coexist at maximal rates. Some of the details of this scheme are 'hidden' behind the choice of $\mathbf{G}_{c,t}$ and behind the loading of the MAT-type symbols $\boldsymbol{x}_{c,t}$ and additional auxiliary symbols $a^*_{k,t}$. The important element for the decoding part later on, will be how to load the symbols, the rate of each symbol, and the corresponding allocated power. An additional element that is hidden from the presentation here is that, while the Q-MAT 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.

- each $\boldsymbol{x}_{c,t}$ carries $f(1-\alpha)$ bits per unit time,
- each $a^*_{\ell,t}$ carries $\min(f\ (1-\alpha), f\alpha)$ bits per unit time.
- and each $a_{k,t}$ carries $f\alpha$ bits per unit time.

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

$$\eta = \arg \max_{\eta' \in [\Gamma, K-1] \cap \mathbb{Z}} \{\eta' \ : \ \alpha_{b,\eta'} \leq \alpha\} \tag{51}$$

for

$$\alpha_{b,\eta} = \frac{\eta - \Gamma}{\Gamma(H_K - H_\eta - 1) + \eta}. \tag{52}$$

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

Remark 2: As we have argued, η (which increases with α) reflects the caching redundancy, and thus the minimum degree of multicasting; instead of content appearing in Γ different caches, it now appears in $\eta \geq \Gamma$ caches instead, which will translate into multicast messages that are intended for more receivers, which will eventually result in reduced delay of delivery. The above equation (52) simply defines the rule that relates α to the degree of multicasting η. This transition, from one η to the next, happens in steps, at the different break-point α values $\alpha_{b,\eta}$ (b here stands for 'break-point').

In all cases,

- all of $\{X_\psi\}_{\psi \in \Psi_{\eta+1}}$ (which are functions of the cached-and-to-be-folded $\{W^{c,f}_{R_k,\tau}\}_{\tau \in \Psi_\eta \backslash \Psi^{(k)}_\eta}$) 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 $\{W^{c,\overline{f}}_{R_k,\tau}\}_{\tau \in \Psi_\eta \backslash \Psi^{(k)}_\eta}$ will be exclusively carried by $a_{k,t}, t \in [0, T]$.

1) Transmission of $\{X_\psi\}_{\psi \in \Psi_{\eta+1}}$: From [46], we know that the transmission relating to $\boldsymbol{x}_{c,t}$ can be treated independently from that of $a_{k,t}$, simply because — as we will further clarify later on — the $a_{k,t}$ 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 [46] corresponding to having the ZF symbols $a_{k,t}$ set to zero. For ease of notation, we will label these phases starting from phase $\eta+1$ and terminating in phase K. The total duration is the desired T. Each phase $j = \eta+1, \ldots, K$ aims to deliver order-j folded messages (cf. (45)), and will do so gradually: phase j will try to 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 linear combinations of received signals from $j+1$ different users, and where these N_{j+1} signals will be conveyed in the next phase $j+1$. During the last phase $j = K$, the transmitter will send fully common symbols that are useful and decoded by all users, thus allowing each user to go back and retroactively decode the information of phase $j = K-1$, which will then be used to decode the information in phase $j = K-2$ and so on, until they reach phase $j = \eta+1$ (first transmission phase) which will complete the task. We proceed to describe these phases. We will use T_j to denote the duration of phase j.

2) 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

$$\boldsymbol{x}_\psi = [x_{\psi,1}, \ldots, x_{\psi,K-\eta}, 0, \ldots, 0]^T \tag{53}$$

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

$$\frac{|X_\psi|}{(K-\eta)} = \frac{f(1-\gamma-\alpha T)}{\binom{K-1}{\eta}(K-\eta)} \tag{54}$$

bits (cf. (47)). 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

$$\text{dur}(\boldsymbol{x}_\psi) = \frac{|X_\psi|}{(K-\eta)(1-\alpha)f}. \tag{55}$$

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

$$T_{\eta+1} = \binom{K}{\eta+1}\text{dur}(\boldsymbol{x}_\psi) = \frac{\binom{K}{\eta+1}|X_\psi|}{(K-\eta)(1-\alpha)f}. \tag{56}$$

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

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\mathbf{G}_{c,t}\boldsymbol{x}_{c,t}}_{L_{\psi,k}, \text{power } \doteq P} + \underbrace{\boldsymbol{h}^T_{k,t}\sum_{\ell \in \bar{\psi}} \boldsymbol{g}_{\ell,t}a^*_{\ell,t}}_{\doteq P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{P^\alpha} \tag{57}$$

where $\bar{\psi} \triangleq [K] \backslash \psi$, while the received signal for the other users $k \in \bar{\psi}$ takes the form

$$y_{k,t} = \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a^*_{k,t}}_{\text{power } \doteq P} + \underbrace{\boldsymbol{h}^T_{k,t}\sum_{\substack{\ell \in \bar{\psi} \\ \ell \neq k}} \boldsymbol{g}_{\ell,t}a^*_{\ell,t} + \underbrace{\boldsymbol{h}^T_{k,t}\mathbf{G}_{c,t}\boldsymbol{x}_{c,t}}_{L_{\psi,k}, \doteq P^{1-\alpha}}}_{i_{\psi,k}, \doteq P^{1-\alpha}} + \underbrace{\boldsymbol{h}^T_{k,t}\boldsymbol{g}_{k,t}a_{k,t}}_{P^\alpha} \tag{58}$$

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 $j = \eta + 2$.

3) 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

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

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

$$\Psi_{\eta+2} = \{\psi \in [K] \ :\ |\psi| = \eta + 2\} \tag{59}$$

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

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

where $\bar{i}_{\psi\backslash k,k}$ is the quantization of $i_{\psi\backslash k,k}$ 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, each of which can be considered 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 calculate that

$$T_{\eta+2} = \binom{K}{\eta+2} \mathrm{dur}(\boldsymbol{x}_\psi) = T_{\eta+1} \frac{\eta+1}{\eta+2}. \tag{60}$$

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.

4) Phase j ($\eta+3 \leq j \leq K$): Generalizing the described approach to any phase $j \in [\eta+3, \ldots, K]$, 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

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

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

$$T_j = T_{\eta+1} \frac{\eta+1}{j}, \quad j = \eta+3, \ldots, K. \tag{61}$$

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 immediately 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 $W_{R_k}^{\overline{c}}$ and all of the unfolded $\{W_{R_k,\tau}^{c,\overline{f}}\}_{\tau\in\Psi_\eta\backslash\Psi_\eta^{(k)}}$.

D. Decoding

The whole transmission lasts $K - \eta$ phases. For each phase j, $j = \eta+1, \cdots, K$ 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 (57), while the received signal for the other users $k \in [K]\backslash\psi$ takes the same form as in (58). As we see in [46], 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 will remove $\boldsymbol{h}_{k,t}^T \mathbf{G}_{c,t} \boldsymbol{x}_{c,t}$ 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 $\{W_{R_k,\psi\backslash\{k\}}^{c,\overline{f}}\}_{\psi\in\Psi_{\eta+1}}$ and uncached $W_{R_k}^{\overline{c}}$. 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 (61) that

$$T = \sum_{j=\eta+1}^{K} T_j = T_{\eta+1} \sum_{j=\eta+1}^{K} \frac{\eta+1}{j} = (\eta+1)(H_K - H_\eta) T_{\eta+1} \tag{62}$$

which combines with (54) and (56) to give

$$T = \frac{(K-\Gamma)(H_K - H_\eta)}{(K-\eta) + \alpha(\eta + K(H_K - H_\eta - 1))} \tag{63}$$

as stated in Theorem 1. The bound by $T = 1-\gamma$ seen in the theorem, corresponds to the fact that the above expression (63) 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 $\alpha = \frac{K(1-\gamma)-1}{(K-1)(1-\gamma)}$.

VI. Conclusions

This work studied the previously unexplored interplay between coded-caching and CSIT feedback quality and timeliness, and identified the optimal cache-aided DoF within a multiplicative factor of 4. This work 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.

The work also revealed interesting connections between retrospective transmission schemes which alleviate the effect of the delay in knowing the channel, and coded caching schemes which alleviate the effect of the delay in knowing the content destination. These connections are at the core of the coded caching paradigm, and their applicability can extend to different settings. The result also implies that a very modest amount of caching can have a substantial impact on performance, as well as can go a long way toward removing the burden of acquiring timely CSIT.

Appendix A
Proof of Lemma 1 (Lower Bound on T^*)

We here note that for the outer (lower) bound to hold, we will make the common assumption that the current channel state must be independent of the previous channel-estimates and estimation errors, *conditioned on the current estimate* (there is no need for the channel to be i.i.d. in time). We will also make the common assumption that the channel is drawn from a continuous ergodic distribution such that all the channel matrices and all their sub-matrices are full rank almost surely.

To lower bound T, we first consider the easier problem where we want to serve $s \leq K$ different files to s users, each with access to all caches. We also consider that we repeat this (easier) last experiment $\lfloor \frac{N}{s} \rfloor$ times, thus spanning a total duration of $T\lfloor \frac{N}{s} \rfloor$ (and up to $\lfloor \frac{N}{s} \rfloor s$ files delivered). At this point, we transfer to the equivalent setting of the s-user MISO BC with delayed CSIT and imperfect current CSIT, and a side-information multicasting link to the receivers, of capacity d_m (files per time slot). Under the assumption that in this latter setting, decoding happens at the end of communication, and once we set

$$d_m T \lfloor \frac{N}{s} \rfloor = sM \tag{64}$$

(which guarantees that the side information from the side link, throughout the communication process, matches the maximum amount of information in the caches), we have that

$$T \lfloor \frac{N}{s} \rfloor d'_\Sigma(d_m) \geq \lfloor \frac{N}{s} \rfloor s \tag{65}$$

where $d'_\Sigma(d_m)$ is any sum-DoF upper bound on the above s-user MISO BC channel with delayed CSIT and the aforementioned side link. Using the bound

$$d'_\Sigma(d_m) = s\alpha + \frac{s}{H_s}(1 - \alpha + d_m)$$

from Lemma 2 (see below), and applying (64), we get

$$T \lfloor \frac{N}{s} \rfloor \big(s\alpha + \frac{s}{H_s}(1 - \alpha + \frac{sM}{T\lfloor \frac{N}{s} \rfloor}) \big) \geq \lfloor \frac{N}{s} \rfloor s \tag{66}$$

and thus that

$$T \geq \frac{1}{(H_s\alpha + 1 - \alpha)} (H_s - \frac{Ms}{\lfloor \frac{N}{s} \rfloor}) \tag{67}$$

which implies a lower bound on the original s-user problem. Maximization over all s, gives the desired bound on the optimal T^*

$$T^* \geq \max_{s \in \{1,\ldots,\min(N,K)\}} \frac{1}{(H_s\alpha + 1 - \alpha)} (H_s - \frac{Ms}{\lfloor \frac{N}{s} \rfloor}) \tag{68}$$

required for the original K-user problem. This concludes the proof of Lemma 1.

A. Sum-DoF Bound for the s-User MISO BC, With Delayed CSIT, α-Quality Current CSIT, and Additional Side Information

We begin with the statement of the lemma, which we prove immediately below.

Lemma 2: For the s-user MISO BC, with delayed CSIT, α-quality current CSIT, and an additional parallel side-link of capacity that scales as $d_m \log P$, the sum-DoF is upper bounded as

$$d_\Sigma(d_m) \leq s\alpha + \frac{s}{H_s}(1 - \alpha + d_m). \tag{69}$$

Proof: Our proof traces the proof of [48], adapting for the additional α-quality current CSIT.

Consider a permutation π of the set $\mathcal{E} = \{1, 2, \cdots, s\}$. For any user $k, k \in \mathcal{E}$, we provide the received signals $y_k^{[n]}$ as well as the message W_k of user k to user $k+1, k+2, \cdots, s$.

We use $y_0^{[n]}$ to denote the output of the side-link and we also define the following notations

$$\begin{aligned}
&\boldsymbol{\Omega}^{[n]} := \{\boldsymbol{h}_k^{[n]}\}_{k=1}^{s}, \quad \hat{\boldsymbol{\Omega}}^{[n]} := \{\hat{\boldsymbol{h}}_k^{[n]}\}_{k=1}^{s}, \ \mathcal{U}^{[n]} := \{\boldsymbol{\Omega}^{[n]}, \hat{\boldsymbol{\Omega}}^{[n]}\},\\
&\boldsymbol{h}_k^{[t]} := \{\boldsymbol{h}_k^{(i)}\}_{i=1}^{t}, \quad y_k^{[t]} := \{y_k^{(i)}\}_{i=1}^{t}, \ t = 1, 2, \cdots, n,\\
&W_{[k]} := \{W_1, W_2, \cdots, W_k\}, \quad y_{[k]}^{[n]} := \{y_1^{[n]}, y_2^{[n]}, \cdots, y_k^{[n]}\}.
\end{aligned}$$

Then for $k = 1, 2, \cdots, s$, we have

$$\begin{aligned}
&n\ (R_k - \epsilon_n)\\
&\le I(W_k; y_{[k]}^{[n]}, y_0^{[n]}, W_{[k-1]}|\mathcal{U}^{[n]}) && (70)\\
&= I(W_k; y_{[k]}^{[n]}, y_0^{[n]}|W_{[k-1]}, \mathcal{U}^{[n]}) && (71)\\
&= I(W_k; y_{[k]}^{[n]}|W_{[k-1]}, \mathcal{U}^{[n]}) + I(W_k; y_0^{[n]}|y_{[k]}^{[n]}, W_{[k-1]}, \mathcal{U}^{[n]})\\
&= h(y_{[k]}^{[n]}|W_{[k-1]}, \mathcal{U}^{[n]}) - h(y_{[k]}^{[n]}|W_{[k]}, \mathcal{U}^{[n]})\\
&\quad + h(y_0^{[n]}|y_{[k]}^{[n]}, W_{[k-1]}, \mathcal{U}^{[n]}) - h(y_0^{[n]}|y_{[k]}^{[n]}, W_{[k]}, \mathcal{U}^{[n]}) && (72)
\end{aligned}$$

where (70) follows from Fano's inequality, where (71) holds due to the fact that the messages are independent, and where the last two steps use the basic chain rule. Note that $W_0 = 0$.

$$\begin{aligned}
&\sum_{k=1}^{s-1}\Big(\frac{h(y_{[k+1]}^{[n]}|W_{[k]}, \mathcal{U}^{[n]})}{k+1} - \frac{h(y_{[k]}^{[n]}|W_{[k]}, \mathcal{U}^{[n]})}{k}\Big)\\
&= \sum_{t=1}^{n}\sum_{k=1}^{s-1}\Big(\frac{h(y_1^{(t)}, \cdots, y_{k+1}^{(t)}|y_1^{[t-1]}, \cdots, y_{k+1}^{[t-1]}, W_{[k]}, \mathcal{U}^{[n]})}{k+1}\\
&\quad - \frac{h(y_1^{(t)}, \cdots, y_k^{(t)}|y_1^{[t-1]}, \cdots, y_k^{[t-1]}, W_{[k]}, \mathcal{U}^{[n]})}{k}\Big) && (73)\\
&= \sum_{t=1}^{n}\sum_{k=1}^{s-1}\Big(\frac{h(y_1^{(t)}, \cdots, y_{k+1}^{(t)}|y_1^{[t-1]}, \cdots, y_{k+1}^{[t-1]}, W_{[k]}, \mathcal{U}^{[t]})}{k+1}\\
&\quad - \frac{h(y_1^{(t)}, \cdots, y_k^{(t)}|y_1^{[t-1]}, \cdots, y_k^{[t-1]}, W_{[k]}, \mathcal{U}^{[t]})}{k}\Big) && (74)\\
&\le \sum_{t=1}^{n}\sum_{k=1}^{s-1}\Big(\frac{h(y_1^{(t)}, \cdots, y_{k+1}^{(t)}|y_1^{[t-1]}, \cdots, y_{k+1}^{[t-1]}, W_{[k]}, \mathcal{U}^{[t]})}{k+1}\\
&\quad - \frac{h(y_1^{(t)}, \cdots, y_k^{(t)}|y_1^{[t-1]}, \cdots, y_k^{[t-1]}, W_{[k]}, \mathcal{U}^{[t]})}{k}\Big) && (75)\\
&\le \sum_{t=1}^{n}\sum_{k=1}^{s-1}\frac{1}{k+1}\alpha\log P + n\cdot o(\log P) && (76)\\
&= n(H_s - 1)\alpha\log P + n\cdot o(\log P) && (77)
\end{aligned}$$

where (73) follows from the linearity of the summation, where (74) holds since the received signal is independent of the future channel state information, where (75) uses the fact that conditioning reduces entropy, and where (77) is from the fact that Gaussian distribution maximizes differential entropy under the covariance constraint and from [9, Lemma 2]. From (72), we then have

$$\begin{aligned}
&\sum_{k=1}^{s}\frac{n\ (R_k - \epsilon_n)}{k}\\
&\le \sum_{k=1}^{s-1}\Big(\frac{h(y_{[k+1]}^{[n]}|W_{[k]}, \mathcal{U}^{[n]})}{k+1} - \frac{h(y_{[k]}^{[n]}|W_{[k]}, \mathcal{U}^{[n]})}{k}\Big)\\
&\quad + h(y_1^{[n]}|\mathcal{U}^{[n]}) - \frac{1}{s}h(y_{[s]}^{[n]}|W_{[s]}, \mathcal{U}^{[n]})\\
&\quad + \sum_{k=1}^{s-1}\Big(\frac{h(y_0^{[n]}|y_{[k+1]}^{[n]}, W_{[k]}, \mathcal{U}^{[n]})}{k+1} - \frac{h(y_0^{[n]}|y_{[k]}^{[n]}, W_{[k]}, \mathcal{U}^{[n]})}{k}\Big)\\
&\quad + h(y_0^{[n]}|y_1^{[n]}, \mathcal{U}^{[n]}) - \frac{1}{s}h\ (y_0^{[n]}|y_{[s]}^{[n]}, W_{[s]}, \mathcal{U}^{[n]})\\
&\le n\ (H_s - 1)\alpha\log P + \underbrace{h(y_1^{[n]}|\mathcal{U}^{[n]})}_{\le n\log P} + \underbrace{h(y_0^{[n]}|y_1^{[n]}, \mathcal{U}^{[n]})}_{\le n\cdot d_m\log P}\\
&\quad + \sum_{k=1}^{s-1}\Big((\frac{1}{k+1} - \frac{1}{k})h(y_0^{[n]}|y_{[k]}^{[n]}, W_{[k]}, \mathcal{U}^{[n]}\Big) + n\cdot o(\log P)\\
&\le n(H_s - 1)\alpha\log P + n\log P + n\cdot d_m\log P + n\cdot o(\log P). && (78)
\end{aligned}$$

where the second inequality is from (77), where the last inequality follows from the fact that entropy is non-negative and $\frac{1}{k+1} - \frac{1}{k} \le 0$. Dividing by $n\log P$ and letting $P \to \infty$ gives

$$\sum_{k=1}^{s}\frac{d_k}{k} \le (H_s - 1)\alpha + 1 + d_m \tag{79}$$

which implies that

$$d_\Sigma(d_m) \le s\alpha + \frac{s}{H_s}(1 - \alpha + d_m) \tag{80}$$

which completes the proof of Lemma 2. ∎

Appendix B
Bounding the Gap to Optimal

This section presents the proof that the gap $\frac{T(\gamma)}{T^*(\gamma)}$, between the achievable $T(\gamma)$ and the optimal $T^*(\gamma)$, is always upper bounded by 4, which also serves as the proof of identifying the optimal $T^*(\gamma)$ within a factor of 4.

We first begin with the case of $\alpha = 0$.

B. Gap for $\alpha = 0$

First recall from Corollary 2b that

$$T(\gamma) = H_K - H_{K\gamma}$$

and from Lemma 1 that

$$T^*(\gamma) \ge \max_{s\in\{1,\ldots,\lfloor\frac{N}{M}\rfloor\}} H_s - \frac{Ms}{\lfloor\frac{N}{s}\rfloor}.$$

We want to prove that

$$\frac{T(\gamma)}{T^*(\gamma)} < 4, \quad \forall K,\ \forall\Gamma = 1, 2, \ldots, K-1 \tag{81}$$

and the proof will be split into three cases: case 1 for $\gamma \in [\frac{1}{K}, \frac{1}{36}]$, case 2 for $\gamma \in [\frac{1}{36}, \frac{1}{2}]$, and case 3 for $\gamma \in [\frac{1}{2}, \frac{K-1}{K}]$. Recall that γ is bounded as $\gamma \ge \frac{1}{K}$.

1) Case 1 ($\gamma \le \frac{1}{36}$): First note that having $\gamma \le \frac{1}{36}$ implies $K \ge 36$. To prove (81), we see that

$$\begin{aligned}
\frac{T}{T^*} &\le \max_{\gamma\in[\frac{1}{K}, \frac{1}{36}]\cap(\mathbb{Z}/K)} \frac{H_K - H_{K\gamma}}{\max\limits_{s\in[1,K]\cap\mathbb{Z}} H_s(1 - \frac{Ms}{H_s\lfloor\frac{N}{s}\rfloor})} && (82)\\
&\le \max_{\gamma\in[\frac{1}{K}, \frac{1}{36}]\cap(\mathbb{Z}/K)} \frac{H_K - H_{K\gamma}}{\max\limits_{s\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} (H_s - \frac{Ms}{\lfloor\frac{N}{s}\rfloor})} && (83)
\end{aligned}$$

$$\leq \max_{\gamma\in[\frac{1}{K},\frac{1}{36}]} \frac{H_K - H_{K\gamma}}{\max_{s\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} (H_s - \frac{Ms}{\lfloor\frac{N}{s}\rfloor})} \tag{84}$$

$$\leq \max_{\gamma\in[\frac{1}{K},\frac{1}{36}]} \frac{\log(\frac{1}{\gamma}) + \epsilon_{36} - \epsilon_\infty}{\max_{s\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} \log(s) + \epsilon_\infty - \gamma s^2 \frac{7}{6}} \tag{85}$$

where (83) holds because $H_s - \frac{Ms}{\lfloor\frac{N}{s}\rfloor} < 0$ when $s > \lfloor\frac{1}{\gamma}\rfloor$ and because we reduced the maximizing region for s, where (84) holds because we increased the maximizing region for γ, and where (85) holds because ϵ_K decreases with K, because $H_K - \log(K) \leq \epsilon_{36}$, $H_{K\gamma} - \log(K\gamma) > \epsilon_\infty$, $H_s > \log(s) + \epsilon_\infty$, and because $(\lfloor\frac{N}{s}\rfloor)/\frac{N}{s} \geq \frac{6}{7}$, $s \leq \frac{N}{6}$ (recall that $s \leq \lfloor\sqrt{K}\rfloor \leq \frac{K}{6} \leq \frac{N}{6}$). Continuing from (85), we have that

$$\frac{T}{T^*} \leq \max_{s_c\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} \max_{\gamma\in[\frac{1}{(s_c+1)^2},\frac{1}{s_c^2}]} \frac{\log(\frac{1}{\gamma}) + \epsilon_{36} - \epsilon_\infty}{\log(s_c) + \epsilon_\infty - \gamma s_c^2 \frac{7}{6}} \tag{86}$$

because

$$\max_{s\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} \log(s) + \epsilon_\infty - \gamma s^2 \frac{7}{6} \geq \log(s_c) + \epsilon_\infty - \gamma s_c^2 \frac{7}{6}$$

for any γ and for any $s_c \in [6, \lfloor\sqrt{K}\rfloor] \cap \mathbb{Z}$. The split of the maximization $\max_{\gamma\in[\frac{1}{K},\frac{1}{36}]}$ into the double maximization $\max_{s_c\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} \max_{\gamma\in[\frac{1}{(s_c+1)^2},\frac{1}{s_c^2}]}$ reflects the fact that we heuristically choose[14] $s = s_c \in \mathbb{Z}$ when $\gamma \in [\frac{1}{(s_c+1)^2}, \frac{1}{s_c^2}]$. Now we perform a simple change of variables, introducing a real valued s' ($s' \triangleq \sqrt{\frac{1}{\gamma}}$) such that $\gamma = \frac{1}{s'^2}$. Hence, a γ range of $\gamma \in [\frac{1}{(s_c+1)^2}, \frac{1}{s_c^2}]$, corresponds to an s' range of $s' \in [s_c, s_c+1]$. Hence we rewrite (86) using this change of variables, to get

$$\frac{T}{T^*} \leq \max_{s_c\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} \max_{s'\in[s_c,s_c+1]} \frac{\log(s'^2) + \epsilon_{36} - \epsilon_\infty}{\log(s_c) + \epsilon_\infty - \frac{7}{6}\frac{s_c^2}{s'^2}} \tag{87}$$

$$\leq \max_{s_c\in[6,\lfloor\sqrt{K}\rfloor]\cap\mathbb{Z}} \underbrace{\frac{\log(s_c+1)^2 + \epsilon_{36} - \epsilon_\infty}{\log(s_c) + \epsilon_\infty - \frac{7}{6}}}_{f(s_c)} \tag{88}$$

$$\leq \frac{2 * \log(7) + \epsilon_{36} - \epsilon_\infty}{\log(6) + \epsilon_\infty - \frac{7}{6}} < 4 \tag{89}$$

where (88) holds because $\frac{s_c^2}{s'^2} \leq 1$, where (89) holds because $f(s_c)$ is decreasing in s_c.

2) Case 2 ($\gamma \in [\frac{1}{36}, \frac{1}{2}]$): In the maximization of the lower bound, we will now choose $s = 1$.

For $K \geq 2$, we have

$$\frac{T}{T^*} \leq \frac{\log(\frac{1}{\gamma}) + \epsilon_2 - \epsilon_\infty}{1-\gamma} =: f(\gamma) \tag{90}$$

because $H_K - \log(K) \leq \epsilon_2$, $H_{K\gamma} - \log(K\gamma) > \epsilon_\infty, \forall K \geq 2$.

[14]Essentially we choose an s that is approximately equal to $\lfloor\sqrt{\frac{1}{\gamma}}\rfloor$, and while this choice does not guarantee the exact maximizing s, it does manage to sufficiently raise the resulting lower bound.

For the above defined $f(\gamma)$, we calculate the derivative to take the form

$$\frac{df(\gamma)}{d\gamma} = \frac{\overbrace{1 - \gamma^{-1} - \log(\gamma) + \epsilon_2 - \epsilon_\infty}^{f'_N(\gamma)}}{\underbrace{(1-\gamma)^2}_{f'_D(\gamma)}}$$

where $f'_N(\gamma)$, $f'_D(\gamma)$ respectively denote the numerator and denominator of this derivative. Since $f'_D(\gamma) > 0, \forall \gamma < 1$, and since

$$\frac{df'_N(\gamma)}{d\gamma} = \gamma^{-2} - \gamma^{-1} \geq 0, \quad \forall \gamma \in [\frac{1}{36}, \frac{1}{2}].$$

To prove this, we use the following lemma, which we prove in Section VI-B.4 below.

Lemma 3: Let $g'_N(\gamma)$ and $g'_D(\gamma)$ respectively denote the numerator and the denominator of the derivative $\frac{dg(\gamma)}{d\gamma}$ of some function $g(\gamma)$. If in the range $\gamma \in [\gamma_1, \gamma_2]$, $g'_N(\gamma)$ increases in γ, and if $g'_D(\gamma) > 0$, then

$$\max_{\gamma\in[\gamma_1,\gamma_2]} g(\gamma) = \max\{g(\gamma = \gamma_1), g(\gamma = \gamma_2)\}. \tag{91}$$

We now continue with the main proof, and apply Lemma 3, to get

$$\max_{\gamma\in[\frac{1}{36},\frac{1}{2}]} f(\gamma) = \max\{f(\frac{1}{2}), f(\frac{1}{36})\} < 4 \tag{92}$$

which directly shows the desired $\frac{T}{T^*} < 4$ for $\frac{1}{36} \leq \gamma \leq \frac{1}{2}$, $K \geq 2$.

3) Case 3 ($\gamma \in [\frac{1}{2}, \frac{K-1}{K}]$): In the maximization of the lower bound, we will again choose $s = 1$. Considering that now γ takes the values $\gamma = \frac{j}{K}$, $j \in [\frac{K}{2}, K-1] \cap \mathbb{Z}$, we have

$$\begin{aligned}
\frac{T}{T^*} &\leq \frac{H_K - H_{K\gamma}}{1-\gamma} = \frac{H_K - H_{(K-j)}}{j/K} \\
&= \frac{1}{j}(\frac{K}{K-j+1} + \frac{K}{K-j+2} + \cdots + 1) \\
&= \frac{1}{j}(1 + \frac{j-1}{K-(j-1)} + 1 + \frac{j-2}{K-(j-2)} + \cdots + 1) \\
&= 1 + \frac{1}{j}(\frac{j-1}{K-(j-1)} + \frac{j-2}{K-(j-2)} + \cdots + \frac{1}{K-1}) \\
&< 2
\end{aligned} \tag{93}$$

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

This completes the proof for the entire case where $\Gamma = 1, 2, \ldots, K-1$.

4) Proof of Lemma 3 : We first note that the condition $\frac{dg'_N(\gamma)}{d\gamma} \geq 0$ implies that $g'_N(\gamma)$ is increasing in γ. We also note that $g'_D(\gamma) \geq 0, \gamma \in [\gamma_1, \gamma_2]$ where naturally $\gamma_1 \leq \gamma_2$. We consider the following three cases.

a) Case 1 ($g'_N(\gamma_1) \geq 0$): If $g'_N(\gamma_1) \geq 0$ then $g'_N(\gamma) \geq 0$ for any $\gamma \in [\gamma_1, \gamma_2]$, which in turn means that $\frac{dg(\gamma)}{d\gamma} = \frac{g'_N(\gamma)}{g'_D(\gamma)} \geq 0, \gamma \in [\gamma_1, \gamma_2]$. This gives the desired

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

b) Case 2 ($g'_N(\gamma_1) < 0$ & $g'_N(\gamma_2) \leq 0$): For any $\gamma \in [\gamma_1, \gamma_2]$, then if $g'_N(\gamma_1) < 0$ & $g'_N(\gamma_2) \leq 0$ then $g'_N(\gamma) \leq 0$, thus $\frac{dg(\gamma)}{d\gamma} \leq 0$, which gives the desired

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

c) Case 3 ($g'_N(\gamma_1) < 0$ & $g'_N(\gamma_2) > 0$): For any $\gamma \in [\gamma_1, \gamma_2]$, then if $g'_N(\gamma_1) < 0$ & $g'_N(\gamma_2) > 0$, there exists a unique $\gamma = \gamma' \in [\gamma_1, \gamma_2]$ such that $g'_N(\gamma') = 0$. Hence $\frac{dg(\gamma)}{d\gamma} \leq 0, \forall \gamma \in [\gamma_1, \gamma']$ and $\frac{dg(\gamma)}{d\gamma} \geq 0, \forall \gamma \in [\gamma', \gamma_2]$. Consequently we have the desired

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

Combining the above three cases, yields the derived $\max_{\gamma \in [\gamma_1, \gamma_2]} g(\gamma) = \max\{g(\gamma_1), g(\gamma_2)\}$ which completes the proof for the case of $\alpha = 0$.

C. Gap for $\alpha > 0$

Our aim here is to show that

$$\frac{T(\gamma, \alpha > 0)}{T^*(\gamma, \alpha > 0)} < 4$$

and we will do so by showing that the above gap is smaller than the gap we calculated above for $\alpha = 0$, which was again bounded above by 4. For this, we will use the expression[15]

$$T(\gamma, \alpha > 0) = \frac{(1-\gamma)(H_K - H_\Gamma)}{\alpha(H_K - H_\Gamma) + (1-\alpha)(1-\gamma)} \tag{94}$$

from Theorem 2, and the expression

$$T^*(\gamma, \alpha > 0) \geq \max_{s \in \{1, \ldots, \lfloor \frac{N}{M} \rfloor\}} \frac{1}{(H_s \alpha + 1 - \alpha)} (H_s - \frac{Ms}{\lfloor \frac{N}{s} \rfloor})$$

from Lemma 1. Hence we have

$$\frac{T}{T^*} \leq \frac{\frac{(1-\gamma)(H_K - H_{K\gamma})}{\alpha(H_K - H_{K\gamma}) + (1-\alpha)(1-\gamma)}}{\max_{s \in \{1, \ldots, \lfloor \frac{N}{M} \rfloor\}} \frac{1}{(H_s \alpha + 1 - \alpha)} (H_s - \frac{Ms}{\lfloor \frac{N}{s} \rfloor})} \tag{95}$$

$$\leq \underbrace{\frac{\frac{(1-\gamma)(H_K - H_{K\gamma})}{\alpha(H_K - H_{K\gamma}) + (1-\alpha)(1-\gamma)}}{\frac{1}{(H_{s_c} \alpha + 1 - \alpha)} (H_{s_c} - \frac{Ms_c}{\lfloor \frac{N}{s_c} \rfloor})}}_{g(s_c, \gamma)} \tag{96}$$

where $s = s_c \in \{1, \ldots, \lfloor \frac{N}{M} \rfloor\}$, but where this s_c will be chosen here to be exactly the same as in the case of $\alpha = 0$. This will be useful because, for that case of $\alpha = 0$, we have already proved that the same specific s_c guarantees that

$$\frac{H_K - H_{K\gamma}}{H_{s_c} - \frac{Ms_c}{\lfloor \frac{N}{s_c} \rfloor}} < 4 \tag{97}$$

for the appropriate ranges of γ. This will apply towards bounding (96).

The proof is broken in two cases, corresponding to $\gamma \in [\frac{1}{36}, \frac{K-1}{K}]$, and $\gamma \in [0, \frac{1}{36}]$.

[15] We note that the here derived upper bound on the gap corresponding to the T in Theorem 2, automatically applies as an upper bound to the gap corresponding to the T from Theorem 1, because the latter T is smaller than the former.

1) Case 1 ($\alpha > 0, \gamma \in [\frac{1}{36}, \frac{K-1}{K}]$): As when $\alpha = 0$ (cf. [50]), we again set $s = 1$, which reduces (96) to

$$\frac{T(\alpha > 0, \gamma)}{T^*(\alpha > 0, \gamma)} \leq \frac{\frac{(1-\gamma)(H_K - H_{K\gamma})}{\alpha(H_K - H_{K\gamma}) + (1-\alpha)(1-\gamma)}}{1-\gamma}.$$

For this case — when α was zero, and when we chose the same $s = 1$ — we have already proved that $\frac{T(\alpha=0,\gamma)}{1-\gamma} < 4$. As a result, since $T(\alpha > 0, \gamma) < T(\alpha = 0, \gamma)$, and since $1 - \gamma \leq T^*$, we conclude that $\frac{T(\alpha>0,\gamma)}{T^*} < 4$, $\gamma \in [\frac{1}{36}, \frac{K-1}{K}]$ which completes this part of the proof.

2) Case 2 ($\alpha > 0, \gamma \in [0, \frac{1}{36}]$): Going back to (96), we now aim to bound

$$g(s_c, \gamma) \triangleq \frac{\frac{(1-\gamma)(H_K - H_{K\gamma})}{\alpha(H_K - H_{K\gamma}) + (1-\alpha)(1-\gamma)}}{\frac{1}{(H_{s_c} \alpha + 1 - \alpha)} (H_{s_c} - \frac{Ms_c}{\lfloor \frac{N}{s_c} \rfloor})} < 4. \tag{98}$$

We already know from the case of $\alpha = 0$ (cf. (97)) that

$$\frac{H_K - H_{K\gamma}}{H_{s_c} - \frac{Ms_c}{\lfloor \frac{N}{s_c} \rfloor}} < 4 \tag{99}$$

holds. Hence we will prove that

$$g(s_c, \gamma) \leq \frac{H_K - H_{K\gamma}}{H_{s_c} - \frac{Ms_c}{\lfloor \frac{N}{s_c} \rfloor}} \tag{100}$$

to guarantee the bound. We note that (100) is implied by

$$H_{s_c} \leq \frac{H_K - H_{K\gamma}}{1-\gamma} \tag{101}$$

which is implied by

$$\log(s_c) \leq \frac{\log(1/\gamma)}{1-\gamma} - \epsilon_6, \quad \epsilon_6 = H_6 - \log(6) \tag{102}$$

because $H_{s_c} \leq \log(s_c) + \epsilon_6, \forall s_c \geq 6, \forall \gamma \in [0, \frac{1}{36}], \forall K$. Furthermore (102) is implied by

$$\frac{1}{2} \log(\frac{1}{\gamma}) \leq \frac{\log(1/\gamma)}{1-\gamma} - \epsilon_6 \tag{103}$$

because $\gamma \in [\frac{1}{(s_c+1)^2}, \frac{1}{s_c^2}]$ means that $s_c \leq \sqrt{\frac{1}{\gamma}}$. Since $\frac{1}{1-\gamma} \geq 1$, then (103) is implied by

$$\frac{1}{2} \log(\frac{1}{\gamma}) \leq \log(\frac{1}{\gamma}) - \epsilon_6. \tag{104}$$

It is obvious that (104) holds since $\gamma \leq \frac{1}{36}$. Towards this, by proving (104), we guarantee (98) and the desired bound. This completes the proof.

REFERENCES

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, 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] A. G. Davoodi and S. A. Jafar. (Mar. 2014). "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." [Online]. Available: https://arxiv.org/abs/1403.1541

[5] 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.

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

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

[8] J. Chen and P. Elia, "Degrees-of-freedom region of the MISO broadcast channel with general mixed-CSIT," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2013, pp. 1–10.

[9] P. de Kerret, X. Yi, and D. Gesbert. (Jan. 2013). "On the degrees of freedom of the K-user time correlated broadcast channel with delayed CSIT." [Online]. Available: https://arxiv.org/abs/1301.2138

[10] J. Chen, S. Yang, and P. Elia, "On the fundamental feedback-vs-performance tradeoff over the MISO-BC with imperfect and delayed CSIT," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 997–1001.

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

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

[13] N. Lee and R. W. Heath, Jr., "Not too delayed CSIT achieves the optimal degrees of freedom," in *Proc. 50th Annu. Allerton Conf. Commun., Control Comput.*, Oct. 2012, pp. 1262–1269.

[14] C. Hao and B. Clerckx, "Imperfect and unmatched CSIT is still useful for the frequency correlated MISO broadcast channel," in *Proc. IEEE Int. Conf. Commun.(ICC)*, Budapest, Hungary, Jun. 2013, pp. 3181–3186.

[15] C. S. Vaze and M. K. Varanasi. (Dec. 2010). "The degrees of freedom regions of two-user and certain three-user MIMO broadcast channels with delayed CSIT." [Online]. Available: https://arxiv.org/abs/1101.0306

[16] M. J. Abdoli, A. Ghasemi, and A. K. Khandani, "On the degrees of freedom of three-user MIMO broadcast channel with delayed CSIT," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2011, pp. 209–213.

[17] M. Torrellas, A. Agustin, and J. Vidal, "Retrospective interference alignment for the MIMO interference broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1492–1496.

[18] A. Bracher and M. A. Wigger, "Feedback and partial message side-information on the semideterministic broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2495–2499.

[19] S. Lashgari, R. Tandon, and S. Avestimehr, "Three-user MISO broadcast channel: How much can CSIT heterogeneity help?" in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 4187–4192.

[20] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, "Placement algorithms for hierarchical cooperative caching," in *Proc. ACM-SIAM SODA*, Jan. 1999, pp. 586–595.

[21] B.-J. Ko and D. Rubenstein, "Distributed self-stabilizing placement of replicated resources in emerging networks," *IEEE/ACM Trans. Netw.*, vol. 13, no. 3, pp. 476–487, Jun. 2005.

[22] Y. Birk and T. Kol, "Coding on demand by an informed source (ISCOD) for efficient broadcast of different supplemental data to caching clients," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2825–2830, Jun. 2006.

[23] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

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

[25] M. A. Maddah-Ali and U. Niesen. (Jan. 2013). "Decentralized coded caching attains order-optimal memory-rate tradeoff." [Online]. Available: https://arxiv.org/abs/1301.5848

[26] M. Ji, A. M. Tulino, J. Llorca, and G. Caire. (Feb. 2014). "Caching and coded multicasting: Multiple groupcast index coding." [Online]. Available: https://arxiv.org/abs/1402.4572

[27] H. Ghasemi and A. Ramamoorthy. (Jan. 2015). "Improved lower bounds for coded caching." [Online]. Available: https://arxiv.org/abs/1501.06003

[28] C.-Y. Wang, S. H. Lim, and M. Gastpar. (Apr. 2015). "Information-theoretic caching: Sequential coding for computing." [Online]. Available: https://arxiv.org/abs/1504.00553

[29] N. Ajaykrishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze. (Jan. 2015). "Critical database size for effective caching." [Online]. Available: https://arxiv.org/abs/1501.02549

[30] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang. (Apr. 2015). "The performance analysis of coded cache in wireless fading channel." [Online]. Available: https://arxiv.org/abs/1504.01452

[31] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," in *Proc. IEEE Int. Symp. Wireless Commun. (ISWCS)*, Aug. 2015, pp. 201–205.

[32] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 809–813.

[33] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[34] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte. (Mar. 2015). "Caching at the edge: A green perspective for 5G networks." [Online]. Available: https://arxiv.org/abs/1503.05365

[35] M. Ji *et al.*, "On the fundamental limits of caching in combination networks," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2015, pp. 695–699.

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

[37] E. Baştuğ, M. Bennis, and M. Debbah. (Mar. 2015). "A transfer learning approach for cache-enabled wireless networks." [Online]. Available: https://arxiv.org/abs/1503.05448

[38] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji. (May 2014). "Caching eliminates the wireless bottleneck in video-aware wireless networks." [Online]. Available: https://arxiv.org/abs/1405.5864

[39] J. Hachem, N. Karamchandani, and S. N. Diggavi. (Apr. 2014). "Content caching and delivery over heterogeneous wireless networks." [Online]. Available: https://arxiv.org/abs/1404.6560

[40] J. Hachem, N. Karamchandani, and S. Diggavi, "Effect of number of users in multi-level coded caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, 2015, pp. 1701–1705.

[41] K. Shanmugam, M. Ji, A. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, Oct. 2016.

[42] M. Deghel, E. Bastug, M. Assaad, and M. Debbah, "On the benefits of edge caching for MIMO interference alignment," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2015, pp. 655–659.

[43] A. Ghorbel, M. Kobayashi, and S. Yang. (Sep. 2015). "Cache-enabled broadcast packet erasure channels with state feedback." [Online]. Available: https://arxiv.org/abs/1509.02074

[44] 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, Sep. 2013.

[45] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *Proc. 53rd Annu. Allerton Conf. Commun., Control Comput.*, Monticello, IL, USA, Sep. 2015, pp. 1099–1105.

[46] P. de Kerret, D. Gesbert, J. Zhang, and P. Elia. (Apr. 2016). "Optimal DoF of the K-user broadcast channel with delayed and imperfect current CSIT." [Online]. Available: https://arxiv.org/abs/1604.01653

[47] S.-E. Elayoubi and J. Roberts, "Performance and cost effectiveness of caching in mobile access networks," in *Proc. 2nd Int. Conf. Inf.-Centric Netw. (ICN)*, 2015, pp. 79–88.

[48] J. Chen, S. Yang, and A. Ozgür, and A. Goldsmith. (Sep. 2014). "Achieving full DoF in heterogeneous parallel broadcast channels with outdated CSIT." [Online]. Available: https://arxiv.org/abs/1409.6808

[49] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," Dept. Commun. Syst., EURECOM, Tech. Rep. RR-15-307, Aug. 2015. [Online]. Available: http://www.eurecom.fr/publication/4723

[50] J. Zhang and P. Elia. (Apr. 2016). "The synergistic gains of coded caching and delayed feedback." [Online]. Available: https://arxiv.org/abs/1604.06531

Jingjing Zhang received the B.Sc. degree from Harbin Institute of Technology in 2010 and the M.Sc. degree from Beijing University of Posts and Telecommunications in 2013, both in Electrical Engineering. Currently, she is pursuing the Ph.D. degree at Communication Systems Department, EURECOM, Sophia Antipolis, France. Her research interests include caching, feedback, topological networks, interference management for multiuser communications and information theory.

Petros Elia received the M.Sc. and Ph.D. in electrical engineering from the University of Southern California (USC), Los Angeles, in 2001 and 2006 respectively. Since February 2008 he has been an Assistant Professor with the Department of Communication Systems at EURECOM in Sophia Antipolis, France. His research interests include combining approaches from different sciences, such as mathematics, physics, and from information theory, complexity theory, and game theory, toward analysis and algorithmic design for distributed and decentralized communication networks. His latest research deals with MIMO, cooperative and multiple access protocols and transceivers, complexity of communication, isolation and connectivity in dense networks, queuing theory and cross-layer design, coding theory, information theoretic limits in cooperative communications, and surveillance networks. He is a Fulbright scholar, the co-recipient of the NEWCOM++ distinguished achievement award 2008- 2011 for a sequence of publications on the topic of complexity in wireless communications, and the recipient of the ERC Consolidator Grant 2016 on cache-aided wireless communications.