On the Optimality of Coded Caching With Heterogeneous User Profiles

Federico Brunero, Petros Elia

On the Optimality of Coded Caching With Heterogeneous User Profiles

Federico Brunero and Petros Elia
Communication Systems Department, EURECOM, Sophia Antipolis, France
Email: {brunero, elia}@eurecom.fr

***Abstract*—In this paper, we consider a coded caching scenario where users have heterogeneous interests. Taking into consideration the system model originally proposed by Wang and Peleato, for which the end-receiving users are divided into groups according to their file preferences, we develop a novel information-theoretic converse on the worst-case communication load under uncoded cache placement. Interestingly, the developed converse bound, jointly with one of the coded schemes proposed by Wang and Peleato, allows us to characterize the optimal worst-case communication load under uncoded prefetching within a constant multiplicative gap of 2. Although we restrict the caching policy to be uncoded, our work improves the previously known order optimality results for the considered caching problem.**

***Index Terms*—Coded caching, file popularity, heterogeneous profiles, information-theoretic converse, user preferences.**

I. Introduction

With the introduction of video streaming platforms and cloud computing services, such as Netflix, Amazon Prime and AWS, we witnessed in the recent years a significant rise in the network traffic. On the one hand, the appearance of data-intensive network applications clearly entails more services at disposal of the users, with consequent benefits for the latter. On the other, communication networks are constantly put under pressure to deliver increasingly larger volumes of data in a timely and efficient manner. As a consequence, the explosion of network traffic has sparked much interest in developing new communication techniques and, to this end, much research focused on the pivotal role that caching will play in the future.

The idea of caching simply consists of exploiting low-cost memories at end-receiving users to diminish significantly the network load from a centralized server to its cache-aided receiving users. The real challenge is to intelligently design the *placement phase* so as to minimize the volume of data that the server has to deliver during the *delivery phase*. The seminal work in [1] tackled this problem introducing the clever concept of coded caching, a major breakthrough which pushed even further the benefits of caching. More specifically, the Maddah-Ali and Niesen (MAN) coded scheme in [1] shed light on the actual information-theoretic gains that caching can provide if the placement phase is carefully designed so that, during the delivery phase, coding techniques can be employed to align the interference patterns. Current research on coded caching spans several topics such as the interplay between multiple antennas and caching [2]–[5], the construction of converse bounds [6], [7] and a variety of other scenarios [8]–[11].

This work was supported by the European Research Council (ERC) through the EU Horizon 2020 Research and Innovation Program under Grant 725929 (Project DUALITY).

A. Coded Caching With Hetereogeneous User Profiles

Since the introduction of coded caching, many works studied several variations of the original information-theoretic system model. In particular, much research aimed at understanding how caching policies should reflect the possibility that, in real scenarios, contents may have different degrees of popularity and users may have diverging interests.

On the one hand, considering that traditional caching techniques heavily rely on the fact that some contents might be more popular than others, the works in [12], [13] focused on the interplay between coded caching and file popularities.

On the other hand, some other works sought to explore the scenario where each user is not necessarily interested in the entire library of contents at the main server — as instead was implicitly assumed in [1]. For instance, the works in [14], [15] explored, for the case of $K = 2$ users, the performance of selfish coded caching in the presence of heterogeneous user profiles or, equivalently, heterogeneous file demand sets (FDSs), proving that, for the instances proposed therein, unselfish caching policies can do better than selfish ones. Later, the work in [16] analyzed for a unified setting the interplay between selfish caching policies and coded caching, providing the meaningful conclusion that unselfish coded caching can be unboundedly better the selfish caching. Subsequently, the recent work in [17] characterized the optimal memory-load tradeoff for a scenario where users are interested in a limited set of contents which depends on the location of the users themselves.

Recently, the authors in [18]–[20] considered coded caching with a very well-defined structure for the user profiles. Assuming that the files in the main library can be classified as either common files (files that can be requested by any user) or unique files (files that can be requested by groups of users only), the work in [18] proposed three different coded schemes for such scenario, providing a related analysis for their peak load performance. Then, these three schemes were studied also in [19] in terms of their average load performance, whereas the work in [20] provided, by means of a converse bound based on cut-set arguments, some order optimality results.

B. Main Contribution

Our work further explores the system model proposed in [18]. In particular, taking advantage of the genie-aided converse

bound idea from [7], our main result is a lower bound on the optimal worst-case communication load under uncoded prefetching. Interestingly, the derived converse, together with an already existing achievable scheme from [18], allows us to characterize the memory-load tradeoff under uncoded placement within a constant multiplicative factor of 2.

C. Paper Outline

The system model and related results are presented in Section II. Section III presents the information-theoretic converse and the order optimality result, whose proofs are provided in Section IV and Section V, respectively. Section VI concludes the paper.

D. Notation

We denote by $\mathbb{Z}^+$ the set of positive integers. For $n \in \mathbb{Z}^+$, we define $[n] \coloneqq \{1, 2, \dots, n\}$. If $a, b \in \mathbb{Z}^+$ such that $a < b$, then $[a : b] \coloneqq \{a, a+1, \dots, b-1, b\}$. For sets we use calligraphic symbols, whereas for vectors we use bold symbols. Given a finite set $\mathcal{A}$, we denote by $|\mathcal{A}|$ its cardinality. We denote by $\binom{n}{k}$ the binomial coefficient and we let $\binom{n}{k} = 0$ whenever $n < 0$, $k < 0$ or $n < k$. For $n \in \mathbb{Z}^+$, we denote by S_n the symmetric group defined over the set $[n]$.

II. System Model and Related Results

We consider the coded caching setting where there is a single server connected to K users through an error-free broadcast channel. The server has access to a central library that contains N files of B bits each. Each user in the system is equipped with a cache of size MB bits (or, equivalently, M files). According to the system model in [18], the K users are split in G groups, where each group consists of K/G users sharing the same interests. Furthermore, the files in the library are divided in two categories, i.e., common files and unique files. There are N_c common files $\{W_n^\mathrm{c} : n \in [N_\mathrm{c}]\}$, where each of them is of interest to every user in the system. Then, for each group $g \in [G]$, there are N_u files $\{W_n^{\mathrm{u},g} : n \in [N_\mathrm{u}]\}$, where each of them is of interest to the users belonging to the group $g \in [G]$ only. Assuming that $\{W_n^\mathrm{c} : n \in [N_\mathrm{c}]\} \cap \{W_n^{\mathrm{u},g} : n \in [N_\mathrm{u}]\} = \emptyset$ for each $g \in [G]$, and that $\{W_n^{\mathrm{u},g_1} : n \in [N_\mathrm{u}]\} \cap \{W_n^{\mathrm{u},g_2} : n \in [N_\mathrm{u}]\} = \emptyset$ for each $g_1, g_2 \in [G]$ with $g_1 \neq g_2$, we have $N = N_\mathrm{c} + GN_\mathrm{u}$ files in total. Deviating from standard notation practices, we will use $W_{d_k}^{f_k, g(k)}$ to denote the file requested by user $k \in [K]$, where $f_k \in \{\mathrm{c}, \mathrm{u}\}$, $d_k \in [N_{f_k}]$ and $g(k)$ is an abuse of notation to denote the group which user k belongs to, i.e., $g(k) \in [G]$ for each $k \in [K]$. We further assume that $W_{d_k}^{\mathrm{c}, g(k)} = W_{d_k}^\mathrm{c}$, since common files do not depend on the group $g \in [G]$. In addition, we let $\boldsymbol{d} = ((d_1, f_1), \dots, (d_K, f_K))$ be the demand vector and we denote by $\mathcal{D}$ the set of all possible demand vectors with distinct requested files, i.e., $W_{d_{k_1}}^{f_{k_1}, g(k_1)} \neq W_{d_{k_2}}^{f_{k_2}, g(k_2)}$ for each $k_1, k_2 \in [K]$ with $k_1 \neq k_2$. Finally, we assume $N_\mathrm{c} \geq K$ and $N_\mathrm{u} \geq K/G$.

The caching problem consists of two phases. During the placement phase, users have access to the main library, and so each user fills their cache using the library. Here, we focus on *uncoded* caching policies according to the following definition.

Definition 1 (Uncoded Prefetching)**.** A cache placement is *uncoded* if each user $k \in [K]$ simply copies in their cache a total of (at most) MB bits from the library. Consequently, the files are partitioned as

$$W_n^\mathrm{c} = \{W_{n,\mathcal{T}}^\mathrm{c} : \mathcal{T} \subseteq [K]\}, \quad \forall n \in [N_\mathrm{c}] \tag{1}$$

$$W_n^{\mathrm{u},g} = \{W_{n,\mathcal{T}}^{\mathrm{u},g} : \mathcal{T} \subseteq [K]\}, \quad \forall n \in [N_\mathrm{u}], \quad \forall g \in [G] \tag{2}$$

where $W_{n,\mathcal{T}}^\mathrm{c}$ and $W_{n,\mathcal{T}}^{\mathrm{u},g}$ represent the bits of W_n^c and $W_n^{\mathrm{u},g}$, respectively, which are cached only by the users in $\mathcal{T}$.

During the delivery phase, the demand vector $\boldsymbol{d} = ((d_1, f_1), \dots, (d_K, f_K))$ is revealed to the server. Denoting by $\mathcal{X}$ the set of caching schemes with uncoded placement, the server transmits a message X of $R(\boldsymbol{d}, \chi, M)B$ bits for a given demand $\boldsymbol{d} \in \mathcal{D}$, a given uncoded cache placement $\chi \in \mathcal{X}$ and some given memory M. The quantity $R(\boldsymbol{d}, \chi, M)$ is called *load* and our goal is to characterize the optimal worst-case communication load under uncoded cache placement, namely, we aim to characterize the quantity given by

$$R^\star(M) = \min_{\chi \in \mathcal{X}} \max_{\boldsymbol{d} \in \mathcal{D}} R(\boldsymbol{d}, \chi, M). \tag{3}$$

In the following, the dependency on M will be implied for the sake of simplicity.

A. An Existing Achievable Scheme

The authors in [18] proposed for the aforementioned setting a coded scheme — referred to as Scheme 2 in [18] — which treats separately the caching and the delivery of common and unique files.

1) Placement Phase: First, the cache of each user is split in two parts for some $0 \leq \beta \leq 1$, so that βM is the part of cache that is devoted to store common files and $(1-\beta)M$ is the part of cache that is devoted to store unique files. Then, common files $\{W_n^\mathrm{c} : n \in [N_\mathrm{c}]\}$ are stored across the K users using the MAN cache placement with memory βM. Similarly, unique files $\{W_n^{\mathrm{u},g} : n \in [N_\mathrm{u}]\}$ are stored across the K/G users in group $g \in [G]$ using the MAN algorithm with memory $(1-\beta)M$.

2) Delivery Phase: It was shown in [18] that, when there are α users per group requesting unique files, the optimal worst-case load can be upper bounded as

$$R^\star \leq \min_\beta \max_\alpha R(\beta, \alpha) \tag{4}$$

where $R(\beta, \alpha)$ is defined as

$$R(\beta, \alpha) \coloneqq \frac{\binom{K}{t_\mathrm{c}+1} - \binom{G\alpha}{t_\mathrm{c}+1}}{\binom{K}{t_\mathrm{c}}} + G\frac{\binom{K/G}{t_\mathrm{u}+1} - \binom{K/G-\alpha}{t_\mathrm{u}+1}}{\binom{K/G}{t_\mathrm{u}}} \tag{5}$$

with $t_\mathrm{c} \coloneqq K\beta M/N_\mathrm{c}$ and $t_\mathrm{u} \coloneqq K(1-\beta)M/GN_\mathrm{u}$.

Since the works in [18], [20] treated the variables K, G, N_c, N_u and $t \coloneqq KM/N$ as continuous[1], we do the same here for the sake of simplicity. Further, we extend the Scheme 2 in [18] to the entire memory regime $0 \leq M \leq N_\mathrm{c} + N_\mathrm{u}$, using the Gamma function whenever the binomial coefficients in (5) have non-integer arguments.

[1]Indeed, if the quantities K, G, N_c and N_u are large enough, the rounding errors due to integer effects during calculations can be neglected.

B. A Genie-Aided Converse Bound

We will provide our converse bound on the optimal worst-case load under uncoded prefetching using the genie-aided approach in [7]. Consider a demand vector $\boldsymbol{d} \in \mathcal{D}$ and let $\boldsymbol{u} = (u_1, \dots, u_K) \in S_K$ be a permutation of the set $[K]$. Denoting by Z_k the cache content of user $k \in [K]$, we can construct a genie-aided user with the following cache content

$$Z' = \left(Z_{u_k} \setminus \left(\bigcup_{i \in [k-1]} Z_{u_i} \cup W_{d_{u_i}}^{f_{u_i}, g(u_i)} \right) : k \in [K] \right) \quad (6)$$

which is enough for such genie-aided user to inductively decode all the requested files from (X, Z'). Consequently, the following

$$\begin{aligned} R(\boldsymbol{d}, \chi) B &\geq H(X) && (7)\\ &\geq H(X \mid Z') && (8)\\ &\geq I\left(\left\{ W_{d_{u_k}}^{f_{u_k}, g(u_k)} \right\}_{k \in [K]} ; X \mid Z' \right) && (9)\\ &= H\left(\left\{ W_{d_{u_k}}^{f_{u_k}, g(u_k)} \right\}_{k \in [K]} \mid Z' \right) && (10)\\ &= \sum_{k \in [K]} \sum_{\mathcal{T} \in ([K] \setminus \{u_1, \dots, u_k\})} \left| W_{d_{u_k}, \mathcal{T}}^{f_{u_k}, g(u_k)} \right| && (11) \end{aligned}$$

holds, which means that we have the following lower bound

$$R(\boldsymbol{d}, \chi) \geq \sum_{k \in [K]} \sum_{\mathcal{T} \in ([K] \setminus \{u_1, \dots, u_k\})} \frac{\left| W_{d_{u_k}, \mathcal{T}}^{f_{u_k}, g(u_k)} \right|}{B} \quad (12)$$

on the communication load for a given[2] $\boldsymbol{d} \in \mathcal{D}$ and $\chi \in \mathcal{X}$. Since it will be of use later, we define the following

$$R_{\text{LB}}(\boldsymbol{d}, \boldsymbol{u}, \chi) \coloneqq \sum_{k \in [K]} \sum_{\mathcal{T} \in ([K] \setminus \{u_1, \dots, u_k\})} \frac{\left| W_{d_{u_k}, \mathcal{T}}^{f_{u_k}, g(u_k)} \right|}{B}. \quad (13)$$

III. Main Results

The first result provides a converse bound on the optimal worst-case load under uncoded prefetching. The proof is presented in Section IV.

Theorem 1. *For the coded caching problem with heterogeneous user profiles presented in Section II, the optimal worst-case load under uncoded cache placement is lower bounded as*

$$R^\star \geq \min_{\beta} \frac{1}{2} \left(\frac{\binom{K}{t_{\text{c}}+1}}{\binom{K}{t_{\text{c}}}} + G \frac{\binom{K/G}{t_{\text{u}}+1}}{\binom{K/G}{t_{\text{u}}}} \right) \quad (14)$$

where $t_{\text{c}} = K\beta M / N_{\text{c}}$ *and* $t_{\text{u}} = K(1-\beta)M/GN_{\text{u}}$.

If we compare the achievable performance in (4) with the converse in Theorem 1, we can provide the following optimality result, whose proof is described in Section V.

Theorem 2. *The achievable load in* (4) *is order optimal within a multiplicative factor of* 2.

[2]We recall that the dependency on M is implied for the sake of simplicity.

Remark 1. The result in Theorem 2 improves the previously known order optimality results presented in [20]. Indeed, even though the work in [20] provided a converse bound without constraining the placement to be uncoded, the smallest gap to optimality therein was a constant factor 8 for the limited memory regime $N/K \leq M \leq N/2G$. Moreover, the achievable performance in (4) was shown to be within a multiplicative factor of $8+8K/G$ from optimal for the memory regime $G(N_{\text{c}} + N_{\text{u}})/K \leq M \leq N/2G$. Here, although our converse holds under the assumption of uncoded placement, we provide a gap to optimality which is a constant multiplicative factor of 2 for the entire[3] memory regime $0 \leq M \leq N_{\text{c}} + N_{\text{u}}$.

IV. Converse Bound Proof

We recall that our goal is to lower bound the quantity

$$R^\star = \min_{\chi \in \mathcal{X}} \max_{\boldsymbol{d} \in \mathcal{D}} R(\boldsymbol{d}, \chi). \quad (15)$$

where again the dependency on M is implied to simplify the notation. Denote by $\mathcal{D}_{\text{c}}$ the subset of $\mathcal{D}$ that contains all demands for which users make requests only from common files, which implies $\boldsymbol{d} = ((d_1, \text{c}), \dots, (d_K, \text{c}))$ for each $\boldsymbol{d} \in \mathcal{D}_{\text{c}}$. Similarly, denote by $\mathcal{D}_{\text{u}}$ the subset of $\mathcal{D}$ for which users make requests only from unique files, which implies $\boldsymbol{d} = ((d_1, \text{u}), \dots, (d_K, \text{u}))$ for each $\boldsymbol{d} \in \mathcal{D}_{\text{u}}$. One can see that $|\mathcal{D}_{\text{c}}| = \binom{N_{\text{c}}}{K} K!$ and $|\mathcal{D}_{\text{u}}| = \left(\binom{N_{\text{u}}}{K/G} (K/G)! \right)^G$. Then, we proceed to lower bound the optimal worst-case load as follows

$$\begin{aligned} R^\star &= \min_{\chi \in \mathcal{X}} \max_{\boldsymbol{d} \in \mathcal{D}} R(\boldsymbol{d}, \chi) && (16)\\ &\geq \min_{\chi \in \mathcal{X}} \max \left(\max_{\boldsymbol{d} \in \mathcal{D}_{\text{c}}} R(\boldsymbol{d}, \chi), \max_{\boldsymbol{d} \in \mathcal{D}_{\text{u}}} R(\boldsymbol{d}, \chi) \right) && (17)\\ &\geq \min_{\chi \in \mathcal{X}} \frac{1}{2} \left(\max_{\boldsymbol{d} \in \mathcal{D}_{\text{c}}} R(\boldsymbol{d}, \chi) + \max_{\boldsymbol{d} \in \mathcal{D}_{\text{u}}} R(\boldsymbol{d}, \chi) \right) && (18)\\ &\geq \min_{\chi \in \mathcal{X}} \frac{1}{2} \left(\frac{1}{|\mathcal{D}_{\text{c}}|} \sum_{\boldsymbol{d} \in \mathcal{D}_{\text{c}}} R(\boldsymbol{d}, \chi) + \frac{1}{|\mathcal{D}_{\text{u}}|} \sum_{\boldsymbol{d} \in \mathcal{D}_{\text{u}}} R(\boldsymbol{d}, \chi) \right) && (19)\\ &= \min_{\chi \in \mathcal{X}} \frac{1}{2} \left(R_{\text{c}}(\chi) + R_{\text{u}}(\chi) \right) && (20) \end{aligned}$$

where $R_{\text{c}}(\chi)$ and $R_{\text{u}}(\chi)$ are defined as

$$R_{\text{c}}(\chi) \coloneqq \frac{1}{|\mathcal{D}_{\text{c}}|} \sum_{\boldsymbol{d} \in \mathcal{D}_{\text{c}}} R(\boldsymbol{d}, \chi) \quad (21)$$

$$R_{\text{u}}(\chi) \coloneqq \frac{1}{|\mathcal{D}_{\text{u}}|} \sum_{\boldsymbol{d} \in \mathcal{D}_{\text{u}}} R(\boldsymbol{d}, \chi). \quad (22)$$

Notice that (17) holds because $(\mathcal{D}_{\text{c}} \cup \mathcal{D}_{\text{u}}) \subset \mathcal{D}$, whereas both (18) and (19) follow from the fact that the maximum can be lower bounded by the average.

We proceed now to lower bound separately $R_{\text{c}}(\chi)$ and $R_{\text{u}}(\chi)$ by means of the genie-aided approach in Section II-B.

[3]The bound in Theorem 1 becomes 0 only when it holds $t_{\text{c}} = K$ and $t_{\text{u}} = K/G$ simultaneously. This happens when $\beta = N_{\text{c}}/M$ and $(1-\beta) = N_{\text{u}}/M$, which implies $0 \leq M \leq N_{\text{c}} + N_{\text{u}}$. In addition, we recall that the Scheme 2 in [18] is extended to the entire memory regime $0 \leq M \leq N_{\text{c}} + N_{\text{u}}$.

A. Lower Bounding $R_{\mathrm{c}}(\chi)$

As we observed in Section II-B, the communication load can be lower bounded, for a given demand $\boldsymbol{d}$ and a given caching scheme χ, as in (12). Hence, if we construct the inequality in (12) for each demand $\boldsymbol{d} \in \mathcal{D}_{\mathrm{c}}$ and for each permutation of users $\boldsymbol{u} \in S_K$, and then we sum together all such inequalities, we obtain

$$K! \sum_{\boldsymbol{d} \in \mathcal{D}_{\mathrm{c}}} R(\boldsymbol{d}, \chi) \geq \sum_{(\boldsymbol{d},\boldsymbol{u}) \in (\mathcal{D}_{\mathrm{c}}, S_K)} R_{\mathrm{LB}}(\boldsymbol{d}, \boldsymbol{u}, \chi) \tag{23}$$

which can be further rewritten as

$$R_{\mathrm{c}}(\chi) \geq \frac{1}{K!|\mathcal{D}_{\mathrm{c}}|} \sum_{(\boldsymbol{d},\boldsymbol{u}) \in (\mathcal{D}_{\mathrm{c}}, S_K)} R_{\mathrm{LB}}(\boldsymbol{d}, \boldsymbol{u}, \chi) \tag{24}$$

recalling that $\boldsymbol{d} = ((d_1, \mathrm{c}), \ldots, (d_K, \mathrm{c}))$ for each $\boldsymbol{d} \in \mathcal{D}_{\mathrm{c}}$ and that $W_n^{\mathrm{c},g} = W_n^{\mathrm{c}}$ for each $n \in [N_{\mathrm{c}}]$ and for each $g \in [G]$. Now, towards simplifying the expression in (24), we proceed by counting how many times each subfile $W_{n,\mathcal{T}}^{\mathrm{c}}$ — for any given $n \in [N_{\mathrm{c}}]$ and $\mathcal{T} \subseteq [K]$ — appears in the RHS of (24).

First, we focus on the subfile $W_{n,\mathcal{T}}^{\mathrm{c}}$ for some $n \in [N_{\mathrm{c}}]$ and $\mathcal{T} \subseteq [K]$ such that $|\mathcal{T}| = t'$ with $t' \in [0:K]$. Next, we denote by $\mathcal{D}_{\mathrm{c},n,k}$ the subset of demands in $\mathcal{D}_{\mathrm{c}}$ for which the file W_n^{c} is requested by some specific user $k \in ([K] \setminus \mathcal{T})$. We can see that $|\mathcal{D}_{\mathrm{c},n,k}| = \binom{N_{\mathrm{c}}}{K} K!/N_{\mathrm{c}} = |\mathcal{D}_{\mathrm{c}}|/N_{\mathrm{c}}$. Then, we observe that, for each $\boldsymbol{d} \in \mathcal{D}_{\mathrm{c},n,k}$, all permutations of users $\boldsymbol{u} \in S_K$ are considered. Nevertheless, we can notice from the construction of (12) that, for each $\boldsymbol{d} \in \mathcal{D}_{\mathrm{c},n,k}$, the subfile $W_{n,\mathcal{T}}^{\mathrm{c}}$ appears in the RHS of (24) only for those permutations of users where k appears before the elements from the set $\mathcal{T}$ in the permutation vector $\boldsymbol{u}$. Since there is a total of $(K - 1 - t')!t'!\binom{K}{t'+1}$ such vectors, we can conclude that the subfile $W_{n,\mathcal{T}}^{\mathrm{c}}$ appears in the RHS of (24) a total of $|\mathcal{D}_{\mathrm{c}}|(K - 1 - t')!t'!\binom{K}{t'+1}/N_{\mathrm{c}}$ times when we consider the demands in $\mathcal{D}_{\mathrm{c},n,k}$ only. Then, since the reasoning above holds for each user $k \in ([K] \setminus \mathcal{T})$, we can conclude that the subfile $W_{n,\mathcal{T}}^{\mathrm{c}}$ appears in the RHS of (24) a total of $|\mathcal{D}_{\mathrm{c}}|(K - t')!t'!\binom{K}{t'+1}/N_{\mathrm{c}}$ times. Moreover, we considered a generic subfile $W_{n,\mathcal{T}}^{\mathrm{c}}$, so the above holds for any $n \in [N_{\mathrm{c}}]$ and for any $\mathcal{T} \subseteq [K]$. Therefore, we can rewrite the RHS of (24) as

$$\frac{1}{K!|\mathcal{D}_{\mathrm{c}}|} \sum_{t' \in [0:K]} |\mathcal{D}_{\mathrm{c}}|(K - t')!t'! \binom{K}{t'+1} x_{t'}^{\mathrm{c}} \tag{25}$$

where $x_{t'}^{\mathrm{c}}$ is defined as

$$0 \leq x_{t'}^{\mathrm{c}} \coloneqq \sum_{n \in [N_{\mathrm{c}}]} \sum_{\mathcal{T} \subseteq [K]: |\mathcal{T}| = t'} \frac{\left| W_{n,\mathcal{T}}^{\mathrm{c}} \right|}{BN_{\mathrm{c}}}. \tag{26}$$

After some algebraic manipulations, we can rewrite (24) as

$$R_{\mathrm{c}}(\chi) \geq \sum_{t' \in [0:K]} f_{\mathrm{c}}(t') x_{t'}^{\mathrm{c}} \tag{27}$$

where $f_{\mathrm{c}}(t')$ is defined as

$$f_{\mathrm{c}}(t') \coloneqq \frac{\binom{K}{t'+1}}{\binom{K}{t'}}. \tag{28}$$

B. Lower Bounding $R_{\mathrm{u}}(\chi)$

Applying as before the genie-aided approach from Section II-B, we obtain the following inequality

$$R_{\mathrm{u}}(\chi) \geq \frac{1}{K!|\mathcal{D}_{\mathrm{u}}|} \sum_{(\boldsymbol{d},\boldsymbol{u}) \in (\mathcal{D}_{\mathrm{u}}, S_K)} R_{\mathrm{LB}}(\boldsymbol{d}, \boldsymbol{u}, \chi) \tag{29}$$

recalling that now $\boldsymbol{d} = ((d_1, \mathrm{u}), \ldots, (d_K, \mathrm{u}))$ for each $\boldsymbol{d} \in \mathcal{D}_{\mathrm{u}}$. Once again, towards simplifying the expression in (29), we proceed by counting how many times each subfile $W_{n,\mathcal{T}}^{\mathrm{u},g}$ — for any given $n \in [N_{\mathrm{u}}]$, $g \in [G]$ and $\mathcal{T} \subseteq [K]$ — appears in the RHS of (29).

First, we focus on the subfile $W_{n,\mathcal{T}_i}^{\mathrm{u},g}$ for a given $g \in [G]$, $n \in [N_{\mathrm{u}}]$ and $\mathcal{T}_i \subseteq [K]$ with $|\mathcal{T}_i| = t'$ for some $t' \in [0:K]$, where i denotes the number of users from group g that appear in the set $\mathcal{T}_i$, namely, $i = |\{k \in \mathcal{T}_i : g(k) = g\}|$. Then, we let k be one of the $(K/G - i)$ users from group g that do not appear in $\mathcal{T}_i$ and we further assume that the file $W_n^{\mathrm{u},g}$ is requested by such user k. If we denote by $\mathcal{D}_{\mathrm{u},n,k}^{g}$ the subset of demands in $\mathcal{D}_{\mathrm{u}}$ for which the file $W_n^{\mathrm{u},g}$ is requested by this user k, we can see that $|\mathcal{D}_{\mathrm{u},n,k}^{g}| = \left(\binom{N_{\mathrm{u}}}{K/G}(K/G)!\right)^G / N_{\mathrm{u}} = |\mathcal{D}_{\mathrm{u}}|/N_{\mathrm{u}}$. In addition, for each $\boldsymbol{d} \in \mathcal{D}_{\mathrm{u},n,k}^{g}$, all permutations $\boldsymbol{u} \in S_K$ are considered. Nevertheless, as already observed, the subfile $W_{n,\mathcal{T}_i}^{\mathrm{u},g}$ appears in the RHS of (29), for each $\boldsymbol{d} \in \mathcal{D}_{\mathrm{u},n,k}^{g}$, only when k is located before the elements from $\mathcal{T}_i$ in the permutation vector $\boldsymbol{u}$. Since there is a total of $(K-1-t')!t'!\binom{K}{t'+1}$ such vectors, the subfile $W_{n,\mathcal{T}_i}^{\mathrm{u},g}$ appears $|\mathcal{D}_{\mathrm{u}}|(K-1-t')!t'!\binom{K}{t'+1}/N_{\mathrm{u}}$ times in the RHS of (29) when only the demands $\mathcal{D}_{\mathrm{u},n,k}^{g}$ are considered. The same reasoning holds for any of the $(K/G - i)$ users from group g not appearing in $\mathcal{T}_i$, so the subfile $W_{n,\mathcal{T}_i}^{\mathrm{u},g}$ appears $|\mathcal{D}_{\mathrm{u}}|(K/G - i)(K-1-t')!t'!\binom{K}{t'+1}/N_{\mathrm{u}}$ in the RHS of (29). In addition, since this reasoning holds for each $g \in [G]$, $n \in [N_{\mathrm{u}}]$ and $\mathcal{T}_i \subseteq [K]$ where $i \in [\max(0, |\mathcal{T}_i| - K + K/G), \min(|\mathcal{T}_i|, K/G)]$, after some algebraic manipulations we can rewrite the RHS of (29) as

$$\sum_{t' \in [0:K]} \sum_{i = \max(0, t' - K + K/G)}^{\min(t', K/G)} \frac{(K/G - i)}{t' + 1} x_{t',i}^{\mathrm{u}} \tag{30}$$

where $x_{t',i}^{\mathrm{u}}$ is defined as

$$0 \leq x_{t',i}^{\mathrm{u}} \coloneqq \sum_{g \in [G]} \sum_{n \in [N_{\mathrm{u}}]} \sum_{\substack{\mathcal{T}_i \subseteq [K]: |\mathcal{T}_i| = t', \\ |\{k \in \mathcal{T}_i : g(k) = g\}| = i}} \frac{\left| W_{n,\mathcal{T}_i}^{\mathrm{u},g} \right|}{BN_{\mathrm{u}}}. \tag{31}$$

We can further lower bound the RHS of (29) as follows

$$\sum_{t' \in [0:K]} \sum_{i = \max(0, t' - K + K/G)}^{\min(t', K/G)} \frac{(K/G - i)}{t' + 1} x_{t',i}^{\mathrm{u}} \tag{32}$$

$$\geq \sum_{t' \in [0:K]} \sum_{i = \max(0, t' - K + K/G)}^{\min(t', K/G)} \frac{(K/G - \min(t', K/G))}{t' + 1} x_{t',i}^{\mathrm{u}} \tag{33}$$

$$= \sum_{t' \in [0:K]} \frac{(K/G - \min(t', K/G))}{t' + 1} \sum_{i = \max(0, t' - K + K/G)}^{\min(t', K/G)} x_{t',i}^{\mathrm{u}} \tag{34}$$

$$= \sum_{t' \in [0:K]} G \frac{\binom{K/G}{t'+1}}{\binom{K/G}{t'}} x^{\mathrm{u}}_{t'} \tag{35}$$

where $x^{\mathrm{u}}_{t'}$ is defined as

$$0 \leq x^{\mathrm{u}}_{t'} \coloneqq \sum_{i=\max(0,t'-K+K/G)}^{\min(t',K/G)} \frac{x^{\mathrm{u}}_{t',i}}{G}. \tag{36}$$

After the passages above, we can rewrite (29) as

$$R_{\mathrm{u}}(\chi) \geq \sum_{t' \in [0:K]} f_{\mathrm{u}}(t') x^{\mathrm{u}}_{t'} \tag{37}$$

where $f_{\mathrm{u}}(t')$ is defined as

$$f_{\mathrm{u}}(t') \coloneqq G \frac{\binom{K/G}{t'+1}}{\binom{K/G}{t'}}. \tag{38}$$

C. Lower Bounding $R^\star$

Finally, we can lower bound the optimal worst-case load $R^\star$. Indeed, we have the following

$$R^\star \geq \min_{\chi \in \mathcal{X}} \frac{1}{2} \left(R_{\mathrm{c}}(\chi) + R_{\mathrm{u}}(\chi) \right) \tag{39}$$

$$\geq \min_{\chi \in \mathcal{X}} \frac{1}{2} \left(\sum_{t' \in [0:K]} f_{\mathrm{c}}(t') x^{\mathrm{c}}_{t'} + \sum_{t' \in [0:K]} f_{\mathrm{u}}(t') x^{\mathrm{u}}_{t'} \right). \tag{40}$$

Moreover, for any uncoded cache placement $\chi \in \mathcal{X}$ and for some $0 \leq \beta \leq 1$, the following

$$\sum_{t' \in [0:K]} x^{\mathrm{c}}_{t'} = 1 \tag{41}$$

$$\sum_{t' \in [0:K]} t' x^{\mathrm{c}}_{t'} \leq \frac{K \beta M}{N_{\mathrm{c}}} \tag{42}$$

holds for common files, whereas we have the following

$$\sum_{t' \in [0:K]} x^{\mathrm{u}}_{t'} = 1 \tag{43}$$

$$\sum_{t' \in [0:K]} t' x^{\mathrm{u}}_{t'} \leq \frac{K(1-\beta) M}{G N_{\mathrm{u}}} \tag{44}$$

for unique files. This means that we can consider $\boldsymbol{x}^{\mathrm{c}} = (x^{\mathrm{c}}_0, \ldots, x^{\mathrm{c}}_K)$ and $\boldsymbol{x}^{\mathrm{u}} = (x^{\mathrm{u}}_0, \ldots, x^{\mathrm{u}}_K)$ as probability mass functions with constraints in (42) and (44) on the first moment, where such constraints simply represent the maximum memory that is available across the caches of all users for common files and unique files, respectively. In light of the above, we can write

$$R^\star \geq \min_{\chi \in \mathcal{X}} \frac{1}{2} \left(\sum_{t' \in [0:K]} f_{\mathrm{c}}(t') x^{\mathrm{c}}_{t'} + \sum_{t' \in [0:K]} f_{\mathrm{u}}(t') x^{\mathrm{u}}_{t'} \right) \tag{45}$$

$$= \min_{\beta, \boldsymbol{x}^{\mathrm{c}}, \boldsymbol{x}^{\mathrm{u}}} \frac{1}{2} \left(\mathbb{E}_{\boldsymbol{x}^{\mathrm{c}}} \left[f_{\mathrm{c}}(t') \right] + \mathbb{E}_{\boldsymbol{x}^{\mathrm{u}}} \left[f_{\mathrm{u}}(t') \right] \right) \tag{46}$$

$$\geq \min_{\beta, \boldsymbol{x}^{\mathrm{c}}, \boldsymbol{x}^{\mathrm{u}}} \frac{1}{2} \left(f_{\mathrm{c}}(\mathbb{E}_{\boldsymbol{x}^{\mathrm{c}}} \left[t' \right]) + f_{\mathrm{u}}(\mathbb{E}_{\boldsymbol{x}^{\mathrm{u}}} \left[t' \right]) \right) \tag{47}$$

$$\geq \min_{\beta} \frac{1}{2} \left(f_{\mathrm{c}}(t_{\mathrm{c}}) + f_{\mathrm{u}}(t_{\mathrm{u}}) \right) \tag{48}$$

$$= \min_{\beta} \frac{1}{2} \left(\frac{\binom{K}{t_{\mathrm{c}}+1}}{\binom{K}{t_{\mathrm{c}}}} + G \frac{\binom{K/G}{t_{\mathrm{u}}+1}}{\binom{K/G}{t_{\mathrm{u}}}} \right). \tag{49}$$

Notice that, since both $f_{\mathrm{c}}(t')$ and $f_{\mathrm{u}}(t')$ are convex and decreasing in t', we have (47) and (48) from Jensen's inequality and the constraints on the first moment, respectively. The proof is complete. □

V. Order Optimality Proof

From Theorem 1 we have

$$R^\star \geq \min_{\beta} \frac{1}{2} \left(\frac{\binom{K}{t_{\mathrm{c}}+1}}{\binom{K}{t_{\mathrm{c}}}} + G \frac{\binom{K/G}{t_{\mathrm{u}}+1}}{\binom{K/G}{t_{\mathrm{u}}}} \right) \tag{50}$$

$$= \frac{1}{2} \left(\frac{\binom{K}{t^\star_{\mathrm{c}}+1}}{\binom{K}{t^\star_{\mathrm{c}}}} + G \frac{\binom{K/G}{t^\star_{\mathrm{u}}+1}}{\binom{K/G}{t^\star_{\mathrm{u}}}} \right) \tag{51}$$

where $t^\star_{\mathrm{c}} = K\beta^\star M / N_{\mathrm{c}}$ and $t^\star_{\mathrm{u}} = K(1-\beta^\star)M / G N_{\mathrm{u}}$ for some optimal $\beta^\star$. Further, from [18] we have

$$R^\star \leq \min_{\beta} \max_{\alpha} R(\beta, \alpha) \tag{52}$$

$$\leq \max_{\alpha} R(\beta^\star, \alpha) \tag{53}$$

$$= \max_{\alpha} \frac{\binom{K}{t^\star_{\mathrm{c}}+1} - \binom{G\alpha}{t^\star_{\mathrm{c}}+1}}{\binom{K}{t^\star_{\mathrm{c}}}} + G \frac{\binom{K/G}{t^\star_{\mathrm{u}}+1} - \binom{K/G-\alpha}{t^\star_{\mathrm{u}}+1}}{\binom{K/G}{t^\star_{\mathrm{u}}}} \tag{54}$$

$$\leq \frac{\binom{K}{t^\star_{\mathrm{c}}+1}}{\binom{K}{t^\star_{\mathrm{c}}}} + G \frac{\binom{K/G}{t^\star_{\mathrm{u}}+1}}{\binom{K/G}{t^\star_{\mathrm{u}}}} \tag{55}$$

where the inequality in (53) holds since the optimal value $\beta^\star$, which minimizes the lower bound in Theorem 1, is not necessarily the optimal memory splitting for the scheme in Section II-A. To conclude, we have

$$\frac{1}{2} \left(\frac{\binom{K}{t^\star_{\mathrm{c}}+1}}{\binom{K}{t^\star_{\mathrm{c}}}} + G \frac{\binom{K/G}{t^\star_{\mathrm{u}}+1}}{\binom{K/G}{t^\star_{\mathrm{u}}}} \right) \leq R^\star \leq \frac{\binom{K}{t^\star_{\mathrm{c}}+1}}{\binom{K}{t^\star_{\mathrm{c}}}} + G \frac{\binom{K/G}{t^\star_{\mathrm{u}}+1}}{\binom{K/G}{t^\star_{\mathrm{u}}}} \tag{56}$$

which implies that the coded scheme in Section II-A is order optimal within a constant multiplicative factor of 2. The proof is complete. □

VI. Conclusion

In this paper, we considered a coded caching setting with heterogeneous user profiles. Under the system model originally proposed in [18], we constructed a novel information-theoretic converse on the worst-case communication load under uncoded prefetching. We developed the lower bound by taking advantage of the genie-aided approach introduced in [7]. Interestingly, the proposed converse bound, jointly with the Scheme 2 from [18], allows us to characterize the optimal worst-case load under uncoded prefetching within a constant multiplicative factor of 2. Although the converse in Theorem 1 holds under the constraint of uncoded placement, the result in Theorem 2, which provides a constant order optimality factor independent of all system parameters, improves the previously known order optimality results in [20]. Possible extensions could include the study of other (maybe more complex) heterogeneous user profiles as well as establishing the exact fundamental limits of the setting considered in this paper.

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] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, jun 2018.

[3] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.

[4] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2331–2348, Apr. 2022.

[5] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.

[6] K. Wan, D. Tuninetti, and P. Piantanida, "An index coding approach to caching with uncoded cache placement," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1318–1332, Mar. 2020.

[7] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1281–1296, Feb. 2018.

[8] J. Hachem, N. Karamchandani, and S. Diggavi, "Coded caching for multi-level popularity and access," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3108–3141, May 2017.

[9] B. Serbetci, E. Parrinello, and P. Elia, "Multi-access coded caching: gains beyond cache-redundancy," in *2019 IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.

[10] P. N. Muralidhar, D. Katyal, and B. S. Rajan, "Maddah-Ali-Niesen scheme for multi-access coded caching," in *2021 IEEE Inf. Theory Workshop (ITW)*, Oct. 2021, pp. 1–6.

[11] F. Brunero and P. Elia, "Fundamental limits of combinatorial multi-access caching," *IEEE Trans. Inf. Theory*, Jul. 2022, early access. doi: 10.1109/TIT.2022.3193723.

[12] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.

[13] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 349–366, Jan. 2018.

[14] C.-H. Chang and C.-C. Wang, "Coded caching with heterogeneous file demand sets — the insufficiency of selfish coded caching," in *2019 IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1–5.

[15] C.-H. Chang, C.-C. Wang, and B. Peleato, "On coded caching for two users with overlapping demand sets," in *ICC 2020 - 2020 IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[16] F. Brunero and P. Elia, "Unselfish coded caching can yield unbounded gains over selfish caching," *IEEE Trans. Inf. Theory*, Aug. 2022, early access. doi: 10.1109/TIT.2022.3195345.

[17] K. Wan, M. Cheng, M. Kobayashi, and G. Caire, "On the optimal memory-load tradeoff of coded caching for location-based content," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3047–3062, Mar. 2022.

[18] S. Wang and B. Peleato, "Coded caching with heterogeneous user profiles," in *2019 IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2619–2623.

[19] C. Zhang and B. Peleato, "On the average rate for coded caching with heterogeneous user profiles," in *ICC 2020 - 2020 IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[20] C. Zhang, S. Wang, V. Aggarwal, and B. Peleato, "Coded caching with heterogeneous user profiles," *IEEE Trans. Inf. Theory*, Jun. 2022, early access. doi: 10.1109/TIT.2022.3186210.