# Coded Caching with Linear Coded Placement: Exact Tradeoff for the Three User Case

Yinbin Ma and Daniela Tuninetti
University of Illinois Chicago, Chicago, IL 60607, USA
Email:{yma52, danielat}@uic.edu

***Abstract*—This paper studies the optimal memory-load tradeoff in a coded caching system with $\mathsf{K} = 3$ users under the constraint that the contents in the local caches are the result of encoding the files by a linear code. This setting generalizes past work that had established the optimal tradeoff under uncoded placement. Let $\mathsf{N}$ be the number of files.**

**For $\mathsf{K} = \mathsf{N} = 3$ the optimal tradeoff under linear coded placement is shown to have a corner point in the low memory regime that was unknown before this work, which is actually optimal without any restrictions on the placement. For $\mathsf{K} = 3, \mathsf{N} \geq 4$, the optimal tradeoff under linear coded placement is shown to be attained by uncoded placement. As a consequence of this result together with past optimality results, it is an open question whether non-linear coded placement would outperform the tradeoff derived in this work for the memory regime $\mathsf{M} \in (1/2, 1)$ for $\mathsf{N} = 3$, and $\mathsf{M} \in (0, \mathsf{N}/3)$ for $\mathsf{K} = 3, \mathsf{N} \in \{4, 5\}$.**

***Index Terms*—Coded Caching; Converse Bound; Acheivable Scheme; Linear Coding Placement; Optimal Tradeoff.**

## I. Introduction

We consider an error-free broadcast network with $\mathsf{K}$ users, each equipped with a local cache of size $\mathsf{M}$ files, and a central server that stores $\mathsf{N}$ files. After the server has pushed content into the local caches and has received the file requests from the the users, it transmits coded multicast messages with the goal of reducing the network communication load by leveraging the locally cached contents. This technique, called coded caching, was originally introduced by Maddah-Ali and Niesen [1]. Coded caching has the potential to reduce the network load by trading local cache storage for network bandwidth [1].

*a) Relevant Past Work:* In [1], a cut-set converse bound and an achievable scheme (referred to as MAN in the following) where content is cached uncoded were proposed; while these bounds do not coincide in general, they are to within a constant factor of one another. Wan *et al.* in [2] derived a converse bound when the caches are populated with uncoded content, and shown it to be achievable when there are more files than users; it is also achievable for less files than users [3]. Yu *et al.* showed that coded placement can at most reduce the load under uncoded placement by a factor of two [4]. While the optimal placement is unknown for general system parameters, *linear coded placement (LinP)* is known to improve over uncoded placement in the low memory regime [1].

To the best our knowledge, the optimal memory-load tradeoff is known in the following cases.

- For $\min\{\mathsf{N}, \mathsf{K}\} = 1$ MAN, or uncoded transmission, is optimal [1].
- Two users:
  - the case $\mathsf{N} = \mathsf{K} = 2$ was already characterized in [1], showing that uncoded placement is insufficient but that LinP is optimal.
  - the case $\mathsf{N} > \mathsf{K} = 2$ was rather recently solved by Tian in [5], showing that MAN is optimal.
- Three users:
  - for $\mathsf{N} = 2$ and $\mathsf{K} = 3$, LinP is optimal [5].
  - for $\mathsf{N} = \mathsf{K} = 3$, LinP is known to be optimal, except for the regime $\mathsf{M} \in (1/3, 1)$ that is open at the time of writing this paper [5].
  - for $\mathsf{N} \in \{4, 5\}$ and $\mathsf{K} = 3$, MAN is optimal, except for the regime $\mathsf{M} \in (0, \mathsf{N}/3)$.
  - for $\mathsf{K} = 3$ and $\mathsf{N} \geq \mathsf{K}(\mathsf{K}+1)/2 = 6$, MAN is optimal for the whole memory regime [4].
- The case $\mathsf{N} = 2, \mathsf{K} \geq 4$ was partially characterized in [5], showing that MAN is optimal in the large memory regime $\mathsf{M} \geq 2(1 - 2/\mathsf{K})$.
- In general, MAN is optimal in the large memory regime $\mathsf{M} \geq \mathsf{N}(1 - 1/\mathsf{K})$ [1].
- In general, Chen *et al.* [6] showed that LinP is optimal for $\mathsf{N} \leq \mathsf{K}$ and $\mathsf{M} \leq 1/\mathsf{K}$.

From the above listed optimality results, one is tempted to conjecture that LinP is optimal, or at least for the few remaining open memory regimes in the $\mathsf{K} = 3$ user case. This is the focus of this paper.

*b) Main Contributions:* Inspired by a recent line of work by Yu and Jafar [7] on the capacity of the Linear Computation Broadcast Channel (LCBC) with three users, we derive a lower bound on the memory-load tradeoff under LinP for the coded caching problem with $\mathsf{K} = 3$ users. We show that this lower bound is achievable. Important observations from this result:

a) case $\mathsf{N} = 3$: a new optimal memory-load point in the low memory regime is discovered, which was unknown from the work in [5]. In this case, the optimal placement remains open for $\mathsf{M} \in (1/2, 1)$, where only a non-linear coded placement could possibly beat the optimal tradeoff we characterized under LinP in this work.

b) case $\mathsf{N} \geq 4$: uncoded placement is optimal under LinP. In this case, the optimal placement remains open for $\mathsf{N} \in \{4, 5\}$ when $\mathsf{M} \in (0, \mathsf{N}/3)$, where only a non-linear coded placement could possibly beat the optimal tradeoff we characterized under LinP in this work.

*c) Paper Organization:* This paper is organized as follows. Section II introduce the coded caching problem and summarizes relevant results. Section III presents our main result. Section IV introduces the LCBC model to re-derive the optimality of LinP for the case of $\mathsf{K} = 2$ users. Section V proves the optimal memory-load tradeoff under LinP for the case of $\mathsf{K} = 3$ users. Section VI concludes the paper. Some proofs can be found in Appendix.

*d) Notation Convention:* We use the following notations:

- Calligraphic symbols denote sets, bold lowercase symbols vectors, bold uppercase symbols matrices, and sans-serif symbols system parameters.
- $|\cdot|$ is the cardinality of a set or the length of a vector.
- For integers $a$ and $b$, $\binom{a}{b}$ is the binomial coefficient, or 0 if $a \geq b \geq 0$ does not hold.
- For an integer $b$, we let $[b] := \{1, \ldots, b\}$.
- For sets $\mathcal{S}$ and $\mathcal{Q}$, we let $\mathcal{S} \setminus \mathcal{Q} := \{k : k \in \mathcal{S}, k \notin \mathcal{Q}\}$.
- For a collection $\{Z_1, \ldots Z_n\}$ and a index set $\mathcal{S} \subseteq [n]$, we let $Z_{\mathcal{S}} := \{Z_i : i \in \mathcal{S}\}$.
- For a set $\mathcal{G}$ and an integer $t$, we let $\Omega^t_{\mathcal{G}} := \{\mathcal{T} \subseteq \mathcal{G} : |\mathcal{T}| = t\}$.

## II. Problem Formulation and Known Results

### A. Problem Formulation

A $(\mathsf{N}, \mathsf{K})$ coded caching system includes a server, $\mathsf{K}$ users, and $\mathsf{N}$ files. Each file has $\mathsf{B}$ symbols, which are uniformly and independently distributed over $\mathbb{F}_{\mathsf{q}}$, where $\mathsf{q}$ is a prime-power number. Files are denoted as $F_i \in \mathbb{F}_{\mathsf{q}}^{\mathsf{B}}, i \in [\mathsf{N}]$. All users are connected to the server via an error-free shared link. The coded caching system has two phases: placement and delivery. Each user has a cache, which stores no more than $\mathsf{MB}$ symbols; we refer to $\mathsf{M}$ as the *memory size*. Caches are filled during the placement phase without knowledge of future demands. During the delivery phase, users communicates to the server their demands and the server transmits a message $X$ of no more than $\mathsf{RB}$ symbols; we refer to $\mathsf{R}$ as the *load*.

Mathematically, the cache content of user $k$ is denoted by $Z_k \in \mathbb{F}_{\mathsf{q}}^{\mathsf{MB}}$ and satisfies

$$H(Z_k|F_1, \ldots, F_{\mathsf{N}}) = 0, \quad \forall k \in [\mathsf{K}]. \tag{1}$$

In the delivery phase, each user demands a file from the server. Denote the demand of user $k$ as $d_k \in [\mathsf{N}]$, $k \in [\mathsf{K}]$. After the demands are known, the server transmits a message $X \in \mathbb{F}_{\mathsf{q}}^{\mathsf{RB}}$ to the users, where

$$H(X|d_1, \ldots, d_{\mathsf{K}}, F_1, \ldots, F_{\mathsf{N}}) = 0. \tag{2}$$

All users must decode their desired file correctly from the local cached content and the transmitted message, i.e.,

$$H(F_{d_k}|X, Z_k) = 0, \quad \forall k \in [\mathsf{K}]. \tag{3}$$

The goal is to characterize the *worst-case load* (or simply load in the following), defined as

$$\mathsf{R}^\star(\mathsf{M}) = \limsup_{\mathsf{B},\mathsf{q}} \min_{X, Z_1, \ldots, Z_{\mathsf{K}}} \max_{d_1, \ldots, d_{\mathsf{K}}} \{\mathsf{R} : \text{is achievable with cache size } \mathsf{M}\}, \ \mathsf{M} \in [0, \mathsf{N}]. \tag{4}$$

### B. Linear Coded Placement

Denote by $F := [F_1; \ldots; F_{\mathsf{N}}] \in \mathbb{F}_{\mathsf{q}}^{\mathsf{NB}}$ the column vector that contains all the symbols of all the files. In this work we consider *linear coding placement* (LinP), that is, in (1) we restrict the cached contents to be of the form

$$\text{LinP:} \quad Z_k = \tilde{\mathbf{E}}_k F \in \mathbb{F}_{\mathsf{q}}^{\mathsf{MB}}, \ \forall k \in [\mathsf{K}], \tag{5}$$

where $\tilde{\mathbf{E}}_k \in \mathbb{F}_{\mathsf{q}}^{\mathsf{MB} \times \mathsf{NB}}$ is the *cache encoding matrix* for user $k$. We do not restrict the encoding for the delivery message in (2) or the decoding in (3) to be linear.

The optimal load in this case is defined as in (4) but with the LinP constraint in (5) (instead of (1)), and is denoted as $\mathsf{R}^\star_{\text{LinP}}(\mathsf{M})$. Trivially,

$$\mathsf{R}^\star(\mathsf{M}) \leq \mathsf{R}^\star_{\text{LinP}}(\mathsf{M}) \leq \mathsf{R}_{\text{YMA}}(\mathsf{M}), \tag{6}$$

where $\mathsf{R}_{\text{YMA}}$ is the optimal load under uncoded placement [3], i.e., each row in each cache encoding matrix in (5) has at most one non-zero entry. The YMA scheme is introduced next.

### C. YMA Scheme and Optimality Under Uncoded Placement

Fix $t \in [0 : \mathsf{K}]$ and partition each file into $\binom{\mathsf{K}}{t}$ equal-size subfiles as

$$F_i = \{F_{i,\mathcal{W}} \in \mathbb{F}_{\mathsf{q}}^{\mathsf{B}/\binom{\mathsf{K}}{t}} : \mathcal{W} \in \Omega^t_{[\mathsf{K}]}\}, \quad \forall i \in [\mathsf{N}]. \tag{7}$$

The cache contents are

$$Z_k = \{F_{i,\mathcal{W}} : i \in [\mathsf{N}], \mathcal{W} \in \Omega^t_{[\mathsf{K}]}, k \in \mathcal{W}\}, \quad \forall k \in [\mathsf{K}]. \tag{8}$$

The memory size is $\mathsf{M} = \mathsf{N}\binom{\mathsf{K}-1}{t-1}/\binom{\mathsf{K}}{t} = \mathsf{N}t/\mathsf{K}$.

Given demands $(d_1, \ldots, d_{\mathsf{K}})$, the server constructs the coded multicast messages

$$X_{\mathcal{S}} := \sum_{k \in \mathcal{S}} \alpha_{\mathcal{S},k} F_{d_k, \mathcal{S}\setminus\{k\}}, \ \forall \mathcal{S} \in \Omega^{t+1}_{[\mathsf{K}]}, \tag{9}$$

where $\alpha_{\mathcal{S},k} \in \mathbb{F}_{\mathsf{q}}$ is a coefficient chosen as in [8], [9]. Some of the multicast messages in (9) are linearly dependent on the others when a file is requested by multiple users [3], [8], [9]. Let $\mathcal{L} \subseteq [\mathsf{K}]$ be the set of *leader users*, which contains one user per each demanded file. The server sends

$$X = \{X_{\mathcal{S}} : \mathcal{S} \in \Omega^{t+1}_{[\mathsf{K}]}, \mathcal{S} \cap \mathcal{L} \neq \emptyset\} \cup \{d_1, \ldots, d_{\mathsf{K}}, \mathcal{L}\}, \tag{10}$$

which enables successful decoding at each user [3].

Thus, the points

$$(\mathsf{M}_t, \mathsf{R}_t)_{\text{YMA}} := \left( \mathsf{N}\frac{\binom{\mathsf{K}-1}{t-1}}{\binom{\mathsf{K}}{t}}, \ \frac{\binom{\mathsf{K}}{t+1} - \binom{\mathsf{K}-\min(\mathsf{K},\mathsf{N})}{t+1}}{\binom{\mathsf{K}}{t}} \right), \tag{11}$$

for $t \in [0 : \mathsf{K}]$ are achievable. When $\mathsf{K} \leq \mathsf{N}$, (11) reduces to the MAN scheme [1], namely

$$(\mathsf{M}_t, \mathsf{R}_t)_{\text{MAN}} := \left( \mathsf{N}\frac{t}{\mathsf{K}}, \ \frac{\mathsf{K}-t}{1+t} \right), \quad t \in [0 : \mathsf{K}]. \tag{12}$$

**Theorem II.1** (From [3])**.** The optimal tradeoff for the coded caching problem under uncoded placement is the lower convex envelope of the points in (11). □

### D. Known Optimality Results for $\mathsf{K} = 2$ Users

**Theorem II.2** (From [1])**.** Any optimal memory-load tradeoff pair for $\mathsf{N} = \mathsf{K} = 2$ satisfies

$$2\mathsf{M} + \mathsf{R} \geq 2, \quad 2\mathsf{M} + 2\mathsf{R} \geq 3, \quad \mathsf{M} + 2\mathsf{R} \geq 2. \tag{13}$$

The first non-trivial corner point in (13) is $(1/2, 1)$ and is attained by LinP; while the second and last non-trivial corner point is $(1, 1/2)$ attained by MAN. □

The first non-trivial corner point in (13) is a special case of the following general result by Chen *at al* [6].

**Theorem II.3** (From [6])**.** For $\mathsf{N} \leq \mathsf{K}$, the segment connecting points $(\mathsf{M}, \mathsf{R})_{\text{trivial}} = (\mathsf{N}, 0)$ and

$$(\mathsf{M}, \mathsf{R})_{\text{CFL}} = (1/\mathsf{K}, \mathsf{N}(1 - 1/\mathsf{K})), \tag{14}$$

is optimal and is achieved by LinP. □

**Theorem II.4** (From [5])**.** Any optimal memory-load tradeoff pair for $\mathsf{N} > \mathsf{K} = 2$ satisfies

$$3\mathsf{M} + \mathsf{NR} \geq 2\mathsf{N}, \quad \mathsf{M} + \mathsf{NR} \geq \mathsf{N}. \tag{15}$$

The only non-trivial corner point in (15) is $(\mathsf{N}/2, 1)$ and is attained by MAN. □

### E. Known Optimality Results for $\mathsf{K} = 3$ Users

**Theorem II.5** (From [5])**.** Any optimal memory-load tradeoff pair for $\mathsf{N} = 2 < \mathsf{K} = 3$ satisfies

$$2\mathsf{M} + \mathsf{R} \geq 2, \quad 3\mathsf{M} + 3\mathsf{R} \geq 5, \quad \mathsf{M} + 2\mathsf{R} \geq 2. \tag{16}$$

The first non-trivial corner point in (16) is attained by LinP in Theorem II.3, while the remaining one by MAN. □

**Theorem II.6** (From [5]–converse without any restrictions on the placement)**.** Any memory-load tradeoff pair for $\mathsf{N} = \mathsf{K} = 3$ must satisfy

$$3\mathsf{M} + \mathsf{R} \geq 3, \quad 6\mathsf{M} + 3\mathsf{R} \geq 8, \tag{17a}$$
$$\mathsf{M} + \mathsf{R} \geq 2, \tag{17b}$$
$$2\mathsf{M} + 3\mathsf{R} \geq 5, \quad \mathsf{M} + 3\mathsf{R} \geq 3. \quad \square \tag{17c}$$

**Remark 1.** The first non-trivial corner point in (17) is attained by LinP in Theorem II.3 and gives optimality for $\mathsf{M} \leq 1/3$; MAN is optimal for $\mathsf{M} \geq 1$. The second non-trivial corner point in (17) is $(2/3, 4/3)$, and is the only one that provably cannot be achieved by LinP [5]. Thus, based on past work, the optimal scheme for for $\mathsf{N} = \mathsf{K} = 3$ in the memory regime $\mathsf{M} \in (1/3, 1)$ is at present unknown and may require non-linear coded placement.

**Theorem II.7** (From [4])**.** The optimal tradeoff for $\mathsf{N} \geq 6$ and $\mathsf{K} = 3$ is attained by MAN.

**Remark 2.** For the $\mathsf{K} = 3$ user case, Theorems II.5, II.6 and II.7 do not cover the whole memory regime when there are $\mathsf{N} = 4$ or $\mathsf{N} = 5$ files. Thus, based on past work, the cases $\mathsf{N} \in \{4, 5\}$ files for $\mathsf{K} = 3$ users are at present open in the memory regime $\mathsf{M} < \mathsf{N}/3$.

## III. Main Results

With reference to Remarks 1 and 2, the goal of this paper is to shed some light into the optimal placement in the open memory regimes for the coded caching problem with $\mathsf{K} = 3$ users. In particular, *we aim to characterize what can be ultimately attained by LinP, and compare it with what is attainable with uncoded placement in Theorem II.1.*

To address this question, in Section IV, we first revisit the case of $\mathsf{K} = 2$ users and re-derive the results in Theorems II.2 and II.4 by leveraging the recent line of work in [10] on the capacity of the LCBC with two users. Our aim is to introduce the methodology and the notation in a case where the notation is easier to grasp. In Section V, we leverage the result in [7] on the capacity of the LCBC with three users, and then to show the following optimal tradeoff under LinP.

**Theorem III.1** (New result: Optimal Tradeoff under LinP for $\mathsf{N} = \mathsf{K} = 3$)**.** Any memory-load tradeoff pair for $\mathsf{N} = \mathsf{K} = 3$ under LinP satisfies

$$3\mathsf{M} + \mathsf{R} \geq 3, \quad 6\mathsf{M} + 3\mathsf{R} \geq 8, \tag{18a}$$
$$4\mathsf{M} + 3\mathsf{R} \geq 7, \quad \text{(new bound)} \tag{18b}$$
$$2\mathsf{M} + 3\mathsf{R} \geq 5, \quad \mathsf{M} + 3\mathsf{R} \geq 3. \tag{18c}$$

The first non-trivial corner point in (18) is $(1/3, 2)$ which is attained by LinP in Theorem II.3; the second non-trivial corner point is $(1/2, 5/3)$ which is attained by a novel LinP scheme described in Section V; the last two non-trivial corner points are attained by MAN. □

**Theorem III.2** (New result: Optimal Tradeoff under LinP for $\mathsf{N} > \mathsf{K} = 3$)**.** The optimal tradeoff for $\mathsf{N} > \mathsf{K} = 3$ under LinP is attained by MAN. □

**Remark 3.** For the case of $\mathsf{N} = \mathsf{K} = 3$, the difference between the converse bound in Theorem II.6 and our optimal tradeoff under LinP in Theorem III.1 is the 'middle bound' (namely, (17b) vs (18b)). We stress that the point $(1/2, 5/3)$ in (18) is not contained in Theorem II.6 and is not achieved by Theorem II.3 either. In addition, this point is actually optimal without any restrictions on the placement, as point $(1/2, 5/3)$ meets with equality the converse bound $6\mathsf{M}+3\mathsf{R} \geq 8$ in (17a).

*The discovery of the new optimal point* $(1/2, 5/3)$ *is a major contribution of this work. This discovery shrinks the memory range for which optimality is unknown from* $\mathsf{M} \in (1/3, 1)$ *to* $\mathsf{M} \in (1/2, 1)$. At this point, for the case of $\mathsf{N} = \mathsf{K} = 3$, we conclude that no known achievable scheme meets the converse bound in Theorem II.6 for $\mathsf{M} \in (1/2, 1)$. Any improvement on our LinP optimality result in Theorem III.1 for $\mathsf{M} \in (1/2, 1)$ could only come from non-linear coded placement.

Similarly, for the case of $\mathsf{K} = 3 < \mathsf{N} \leq 5$, any improvement on our LinP optimality result in Theorem III.2 for $\mathsf{M} \in (0, \mathsf{N}/3)$ could only come from non-linear coded placement. □

## IV. Warm-up: Proof of Theorems II.2 and II.4

Here we leverage the results of [10] for the LCBC with two users to re-derive the results in Theorems II.2 and II.4. We start by introducing matrix notations from [7], [10].

### A. Some Matrix Notation

In the rest of the paper, given a matrix $\mathbf{M}$, we let $\langle \mathbf{M} \rangle$ denote the subspace spanned by the columns of $\mathbf{M}$, and $\mathsf{rk}(\mathbf{M})$ the rank of $\mathbf{M}$.

Given two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we use a Matlab-like notation where $[\mathbf{M}_1, \mathbf{M}_2]$ denotes the concatenated matrix which can be partitioned column-wise into $\mathbf{M}_1$ and $\mathbf{M}_2$, while $[\mathbf{M}_1; \mathbf{M}_2]$ denotes the concatenated matrix which can be partitioned row-wise into $\mathbf{M}_1$ and $\mathbf{M}_2$. In addition, $\mathbf{M}_1 \cap \mathbf{M}_2$ denotes a matrix whose columns span $\langle \mathbf{M}_1 \rangle \cap \langle \mathbf{M}_2 \rangle$; and $\mathbf{M}_1 \setminus \mathbf{M}_2$ is the subspace in $\langle \mathbf{M}_1 \rangle$ but not in $\langle \mathbf{M}_2 \rangle$

Given two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we define the *conditional rank* as follows [10]

$$\mathsf{rk}(\mathbf{M}_1 \mid \mathbf{M}_2) := \mathsf{rk}([\mathbf{M}_1; \mathbf{M}_2]) - \mathsf{rk}(\mathbf{M}_2) \tag{19}$$

$$= \mathsf{rk}(\mathbf{M}_1) - \mathsf{rk}(\mathbf{M}_1 \cap \mathbf{M}_2). \tag{20}$$

Given three matrices $\mathbf{M}_1, \mathbf{M}_2$ and $\mathbf{M}_3$, suppose $\{i, j, k\}$ is a permutation of $[3]$, we define the following [7]

$$\mathbf{M}_{123} := \mathbf{M}_1 \cap \mathbf{M}_2 \cap \mathbf{M}_3, \tag{21}$$

$$\mathbf{M}_{ij} := \mathbf{M}_i \cap \mathbf{M}_j, \tag{22}$$

$$\mathbf{M}_{i(j,k)} := \mathbf{M}_i \cap [\mathbf{M}_j \cup \mathbf{M}_k]. \tag{23}$$

### B. LCBC Model

Since we leverage results for LCBC model [10], we briefly describe the LCBC problem. A general LCBC model is specified by the parameters $(\mathsf{q}, \mathsf{r}, \mathsf{K}, \mathbf{E}_{[\mathsf{K}]}, \mathbf{D}_{[\mathsf{K}]})$, where a server has $\mathsf{r}$ uniformly and independently distributed data blocks from $\mathbb{F}_\mathsf{q}$ and serves $\mathsf{K}$ users. We denote the concatenation of data blocks as $\mathbf{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_\mathsf{r}] \in \mathbb{F}_\mathsf{q}^{\mathsf{r} \times 1}$. For every user $j \in [\mathsf{K}]$, we denote the "cache projection matrix" as $\mathbf{E}_j \in \mathbb{F}_\mathsf{q}^{m_j \times \mathsf{r}}$, and the "demand projection matrix" as $\mathbf{D}_j \in \mathbb{F}_\mathsf{q}^{n_j \times \mathsf{r}}$, where $m_j$ and $n_j$ are non-negative integers, which means that user $j$ has a side-information $\mathbf{S}_j = \mathbf{E}_j \mathbf{X} \in \mathbb{F}_\mathsf{q}^{m_j \times 1}$ and wants $\mathbf{W}_j = \mathbf{D}_j \mathbf{X} \in \mathbb{F}_\mathsf{q}^{n_j \times 1}$. A valid scheme consists of an encoding function $\Psi_0$, and decoding functions $\Psi_1, \ldots, \Psi_\mathsf{K}$. The server sends $\Psi_0(\mathbf{X}) \in \mathbb{F}_\mathsf{q}^{\Delta \times 1}$ to the users and user $j \in [\mathsf{K}]$ decodes $y_j := \Psi_j(\Psi_0(\mathbf{X}), \mathbf{S}_j) : H(\mathbf{W}_j | y_j) = 0$. The optimal number of transmission is $\Delta^\star(\mathbf{E}_1, \ldots, \mathbf{E}_\mathsf{K}; \mathbf{D}_1, \ldots, \mathbf{D}_\mathsf{K})$ and is the smallest $\Delta$ such as all listed requirements are met, where symbol and field extensions are possible [10].

### C. Bounds for coded caching from the LCBC model

We can obtain a lower bound on the load for the coded caching model with LinP from LCBC as follows. Let

$$\mathsf{r} := \mathsf{N}\mathsf{B}, \tag{24a}$$

$$m_1 = m_2 = \ldots = m_\mathsf{K} := \mathsf{M}\mathsf{B}, \tag{24b}$$

$$n_1 = n_2 = \ldots = n_\mathsf{K} := \mathsf{B}. \tag{24c}$$

The demand projection matrices correspond to single file retrieval, thus they are of the of form

$$\mathbf{D}_j = \mathbf{e}_{d_j} \otimes \mathbf{I}_\mathsf{B}, \ d_j \in [\mathsf{N}], \ j \in [\mathsf{K}], \tag{24d}$$

where $\otimes$ is the Kronecker product and $\mathbf{e}_i$ the $i$-th standard basis vector. We have

$$\mathsf{BR}^\star_{\text{LinP}} \geq \min_{\mathbf{E}_1, \ldots, \mathbf{E}_\mathsf{K}} \max_{d_1, \ldots, d_\mathsf{K}} \Delta^\star(\mathbf{E}_1, \ldots, \mathbf{E}_\mathsf{K}; \mathbf{D}_1, \ldots, \mathbf{D}_\mathsf{K}). \tag{24e}$$

### D. The two-user Coded Caching Problem with LinP

We will leverage the following LCBD result.

**Theorem IV.1** (From [10])**.** For the LCBC with $\mathsf{K} = 2$ users, side information matrices $(\mathbf{E}_1, \mathbf{E}_2)$ and demand matrices $(\mathbf{D}_1, \mathbf{D}_2)$, the optimal number of transmissions is

$$\Delta^\star(\mathbf{E}_1, \mathbf{E}_2; \mathbf{D}_1, \mathbf{D}_2) \tag{25}$$

$$= \max_{(i,j) \in \{(1,2),\ (2,1)\}} \big( \mathsf{rk}(\mathbf{U}_i \mid \mathbf{E}_i) + \mathsf{rk}(\mathbf{U}_j \mid \mathbf{U}_i, \mathbf{E}_1, \mathbf{E}_2) \big)$$

where $\mathbf{U}_j := [\mathbf{E}_j, \mathbf{D}_j], j \in [2]$. □

We now show how to leverage Theorem IV.1 for the coded caching problem with LinP with $\mathsf{K} = 2$ users. Following [10], we partition the cache encoding matrices $\tilde{\mathbf{E}}_1$ and $\tilde{\mathbf{E}}_2$ in (5) into orthogonal submatrices $\mathbf{E}_1, \mathbf{E}_{12}, \mathbf{E}_2$, where

$$\mathbf{E}_{12} := \tilde{\mathbf{E}}_1 \cap \tilde{\mathbf{E}}_2, \tag{26a}$$

$$\mathbf{E}_1 := \tilde{\mathbf{E}}_1 \setminus \mathbf{E}_{12}, \tag{26b}$$

$$\mathbf{E}_2 := \tilde{\mathbf{E}}_2 \setminus \mathbf{E}_{12}. \tag{26c}$$

Without loss of generality, we write $\mathbf{E}_\mathcal{S}$, $\mathcal{S} \subseteq [2]$, as

$$\begin{bmatrix} \mathbf{P}^\mathcal{S}_{\{1\},1} & 0 & \ldots & 0 & 0 \\ 0 & \mathbf{P}^\mathcal{S}_{\{2\},2} & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{P}^\mathcal{S}_{\{\mathsf{N}-1\},\mathsf{N}-1} & 0 \\ 0 & 0 & \ldots & 0 & \mathbf{P}_{\{\mathsf{N}\},\mathsf{N}} \\ \mathbf{P}^\mathcal{S}_{\{1,2\},1} & \mathbf{P}^\mathcal{S}_{\{1,2\},2} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{P}^\mathcal{S}_{\{\mathsf{N}-1,\mathsf{N}\},\mathsf{N}-1} & \mathbf{P}^\mathcal{S}_{\{\mathsf{N}-1,\mathsf{N}\},\mathsf{N}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}^\mathcal{S}_{[\mathsf{N}],1} & \mathbf{P}^\mathcal{S}_{[\mathsf{N}],2} & \ldots & \mathbf{P}^\mathcal{S}_{[\mathsf{N}],\mathsf{N}-1} & \mathbf{P}^\mathcal{S}_{[\mathsf{N}],\mathsf{N}} \end{bmatrix}, \tag{27}$$

where $\mathbf{P}^\mathcal{S}_{\mathcal{T},n}$, for $\mathcal{T} \subseteq [\mathsf{N}]$ and $n \in \mathcal{T}$, can be thought of as a linear encoding matrix involving $\mathcal{T}$ files for the $n^{\text{th}}$ file. We next leverage two symmetry properties for the coded caching problem: file symmetry and user symmetry [5]: even if we shuffle the indices of users or of the files, the load is unchanged. Thus, for $\mathcal{S} \subseteq [2], \mathcal{T} \subseteq [\mathsf{N}], n \in \mathcal{T}$, we can set the rank of $\mathbf{P}^\mathcal{S}_{\mathcal{T},n}$ as,

$$r_{i,j}\mathsf{B} := \mathsf{rk}(\mathbf{P}^\mathcal{S}_{\mathcal{T},n}) : \ i = |\mathcal{T}|, \ j = |\mathcal{S}|. \tag{28}$$

With this we write

$$\mathsf{rk}(\mathbf{E}_1) = \mathsf{rk}(\mathbf{E}_2) = \sum_{i\in[\mathsf{N}]} \binom{\mathsf{N}}{i} r_{1,i}\,\mathsf{B}, \tag{29}$$

$$\mathsf{rk}(\mathbf{E}_{12}) = \sum_{i\in[\mathsf{N}]} \binom{\mathsf{N}}{i} r_{2,i}\,\mathsf{B}. \tag{30}$$

In addition, we write the cache and the file constraints as

$$\text{(cache)} \qquad \sum_{i\in[\mathsf{N}]} \binom{\mathsf{N}}{i} (r_{1,i} + r_{2,i}) \le \mathsf{M}, \tag{31}$$

$$\text{(file)} \qquad \sum_{i\in[\mathsf{N}]} \binom{\mathsf{N}-1}{i-1} (2r_{1,i} + r_{2,i}) \le 1. \tag{32}$$

Assume now that the two users have different demands, for example $d_1 = 1, d_2 = 2$ (and similarly for all other demands), then we can write

$$\mathbf{U}_1 = [\mathbf{E}_1, \mathbf{e}_1 \otimes \mathbf{I}_{\mathsf{B}}] = \begin{bmatrix} \mathbf{I} & 0 & \dots & 0 & 0 \\ 0 & \mathbf{P}^{\mathcal{S}}_{\{2\},2} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{P}^{\mathcal{S}}_{\{\mathsf{N}-1\},\mathsf{N}-1} & 0 \\ 0 & 0 & \dots & 0 & \mathbf{P}_{\{\mathsf{N}\},\mathsf{N}} \\ 0 & \mathbf{P}^{\mathcal{S}}_{\{1,2\},2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{P}^{\mathcal{S}}_{\{\mathsf{N}-1,\mathsf{N}\},\mathsf{N}-1} & \mathbf{P}^{\mathcal{S}}_{\{\mathsf{N}-1,\mathsf{N}\},\mathsf{N}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \mathbf{P}^{\mathcal{S}}_{[\mathsf{N}],2} & \dots & \mathbf{P}^{\mathcal{S}}_{[\mathsf{N}],\mathsf{N}-1} & \mathbf{P}^{\mathcal{S}}_{[\mathsf{N}],\mathsf{N}} \end{bmatrix}, \tag{33}$$

and thus compute

$$\frac{1}{\mathsf{B}}\mathsf{rk}(\mathbf{U}_1 \mid \mathbf{E}_1) = 1 - r_{1,1} - r_{2,1}, \tag{34}$$

$$\frac{1}{\mathsf{B}}\mathsf{rk}(\mathbf{U}_2 \mid \mathbf{U}_1, \mathbf{E}_1, \mathbf{E}_2) = 1 - 2r_{1,1} - r_{2,1} - 2r_{1,2} - r_{2,2}. \tag{35}$$

Finally, we write the optimal number of transmissions in (25) as the following Linear Program (LP)

$$\mathsf{R}^{\star}_{\text{LinP}} \ge \min(2 - 3r_{1,1} - 2r_{1,2} - 2r_{2,1} - r_{2,2}) \tag{36a}$$

$$\text{s.t.} \quad (r_{1,1} + r_{2,1}) + \frac{\mathsf{N}-1}{2}(r_{1,2} + r_{2,2}) \le \frac{\mathsf{M}}{\mathsf{N}}, \tag{36b}$$

$$(2r_{1,1} + r_{2,1}) + (\mathsf{N}-1)(2r_{1,2} + r_{2,2}) \le 1. \tag{36c}$$

In Tables I and II we report the optimal values for the LP variables in (36) for the case of $\mathsf{N} = 2$ and $\mathsf{N} \ge 3$, respectively. Note that variables whose optimal value is always zero are not listed. When $\mathsf{N} = 2$, LinP is needed for $\mathsf{M} \in [0, 1]$ (i.e., non zero value for $r_{1,2}$) and the optimal load is as stated in Theorem II.2. When $\mathsf{N} \ge 3$, MAN is optimal (i.e., all $r_{\cdot,2}$ are zero) and the optimal load is as stated in Theorem II.4.

TABLE I: Optimal LP values for $\mathsf{K} = 2 = \mathsf{N}$.

| $\mathsf{M}$ | $[0, \frac{1}{2}]$ | $[\frac{1}{2}, 1]$ | $[1, 2]$ |
|---|---|---|---|
| $r_{1,1}$ | $0$ | $1 - \frac{\mathsf{M}}{2}$ | $\mathsf{M} - \frac{1}{2}$ |
| $r_{1,2}$ | $\mathsf{M}$ | $1 - \mathsf{M}$ | $0$ |
| $r_{2,1}$ | $0$ | $0$ | $\mathsf{M} - 1$ |
| $\mathsf{R}$ | $2 - 2\mathsf{M}$ | $\frac{3-2\mathsf{M}}{2}$ | $1 - \frac{\mathsf{M}}{2}$ |

TABLE II: Optimal LP values for $\mathsf{K} = 2 < \mathsf{N}$.

| $\frac{\mathsf{M}}{\mathsf{N}}$ | $[0, \frac{1}{2}]$ | $[\frac{1}{2}, 1]$ |
|---|---|---|
| $r_{1,1}$ | $\frac{\mathsf{M}}{\mathsf{N}}$ | $1 - \frac{\mathsf{M}}{\mathsf{N}}$ |
| $r_{2,1}$ | $0$ | $2\frac{\mathsf{M}}{\mathsf{N}} - 1$ |
| $\mathsf{R}$ | $2 - 3\frac{\mathsf{M}}{\mathsf{N}}$ | $1 - \frac{\mathsf{M}}{\mathsf{N}}$ |

## V. Proof of Theorems III.1 and III.2

Next we leverage the result of [7] for the LCBC with three users to prove Theorems III.1 and III.2. The approach is similar to the one in Section IV, except for the complexity because now we deal with three users.

**Theorem V.1** (From [7])**.** For the LCBC with $\mathsf{K} = 3$ users, side information matrices $(\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3)$ and demand matrices $(\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$, the optimal number of transmissions is

$$\Delta^{\star}(\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3; \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3) \tag{37a}$$

$$= \mathsf{rk}(\mathbf{U}_1 \mid \mathbf{E}_1) + \mathsf{rk}(\mathbf{U}_2 \mid \mathbf{E}_2) + \mathsf{rk}(\mathbf{U}_3 \mid \mathbf{E}_3) \tag{37b}$$

$$- \max\{2\lambda_{123} + \lambda_{12} + \lambda_{13} + \lambda_{23} + \lambda\}, \tag{37c}$$

where the maximization is subject to the following constraints

$$\lambda_{123} \le \mathsf{rk}(\mathbf{U}_{123} \mid \mathbf{E}_1), \tag{37d}$$

$$\lambda_{123} \le \mathsf{rk}(\mathbf{U}_{123} \mid \mathbf{E}_2), \tag{37e}$$

$$\lambda_{123} \le \mathsf{rk}(\mathbf{U}_{123} \mid \mathbf{E}_3), \tag{37f}$$

$$\lambda_{12} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{12} \mid \mathbf{E}_1), \tag{37g}$$

$$\lambda_{12} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{12} \mid \mathbf{E}_2), \tag{37h}$$

$$\lambda_{13} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{13} \mid \mathbf{E}_1), \tag{37i}$$

$$\lambda_{13} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{13} \mid \mathbf{E}_3), \tag{37j}$$

$$\lambda_{23} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{23} \mid \mathbf{E}_2), \tag{37k}$$

$$\lambda_{23} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{23} \mid \mathbf{E}_3), \tag{37l}$$

$$\lambda_{12} + \lambda_{13} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{12}, \mathbf{U}_{13} \mid \mathbf{E}_1), \tag{37m}$$

$$\lambda_{12} + \lambda_{23} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{12}, \mathbf{U}_{23} \mid \mathbf{E}_2), \tag{37n}$$

$$\lambda_{13} + \lambda_{23} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{13}, \mathbf{U}_{23} \mid \mathbf{E}_3), \tag{37o}$$

$$\lambda + \lambda_{12} + \lambda_{13} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{1(2,3)} \mid \mathbf{E}_1), \tag{37p}$$

$$\lambda + \lambda_{12} + \lambda_{23} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{2(1,3)} \mid \mathbf{E}_2), \tag{37q}$$

$$\lambda + \lambda_{13} + \lambda_{23} + \lambda_{123} \le \mathsf{rk}(\mathbf{U}_{3(1,2)} \mid \mathbf{E}_3). \tag{37r}$$

□

**Remark 4.** In (37), each $\lambda$-parameter can be viewed as a coding gain for different multicast messages, as illustarted in Fig. 1. The term $\mathsf{rk}(\mathbf{U}_1 \mid \mathbf{E}_1) + \mathsf{rk}(\mathbf{U}_2 \mid \mathbf{E}_2) + \mathsf{rk}(\mathbf{U}_3 \mid \mathbf{E}_3)$ in (37b) represent the load for uncoded transmission when users are served sequentially one by one. $\lambda_{123}$ represents an

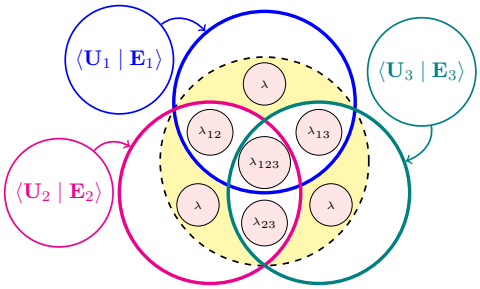



Fig. 1: Intuitive understanding for (37) from [7, Fig. 3].

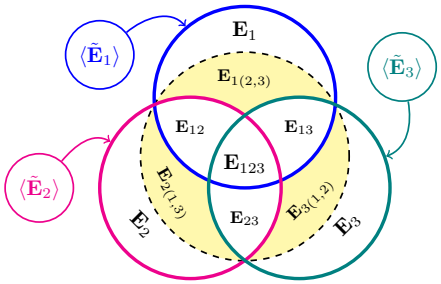



Fig. 2: The decomposition of $\langle \tilde{\mathbf{E}}_1 \rangle, \langle \tilde{\mathbf{E}}_2 \rangle, \langle \tilde{\mathbf{E}}_3 \rangle$ into subspaces, labeled by corresponding bases

amount of information that benefits all users; so its transmission reduces the load by $2\lambda_{123}$ in (37c). Similarly, $\lambda_{ij}$ represents an amount of information that benefits users $i$ and $j$, for $\{i,j\} \subseteq [3]$; so their transmission reduces the load by $\lambda_{12}+\lambda_{13}+\lambda_{23}$ in (37c). The yellow regions in Fig. 1 are labeled by $\lambda$ and are somewhat special; the corresponding subspaces are mutually disjoint but any two of them contain the remaining one [7]; its transmission reduces the load of uncoded transmission by $\lambda$ in (37c). □

### A. Proof of Theorem III.1

*Converse:* We leverage Theorem V.1 for the coded caching problem with $\mathsf{K} = 3$ users. From [7], we partition $\tilde{\mathbf{E}}_1, \tilde{\mathbf{E}}_2, \tilde{\mathbf{E}}_3$ as shows in Fig. 2. Note that $\mathbf{E}_{1(2,3)}, \mathbf{E}_{2(1,3)}, \mathbf{E}_{3(1,2)}$ are mutually disjoint but any two of them contain the remaining one. Without loss of generality, for every $\mathcal{S} \subseteq [3]$, we write $\mathbf{E}_{\mathcal{S}}$ as

$$\mathbf{E}_{\mathcal{S}} = \begin{bmatrix} \mathbf{P}^{\mathcal{S}}_{\{1\},1} & 0 & 0 \\ 0 & \mathbf{P}^{\mathcal{S}}_{\{2\},2} & 0 \\ 0 & 0 & \mathbf{P}^{\mathcal{S}}_{\{3\},3} \\ \mathbf{P}^{\mathcal{S}}_{\{1,2\},1} & \mathbf{P}^{\mathcal{S}}_{\{1,2\},2} & 0 \\ \mathbf{P}^{\mathcal{S}}_{\{1,3\},1} & 0 & \mathbf{P}^{\mathcal{S}}_{\{1,3\},3} \\ 0 & \mathbf{P}^{\mathcal{S}}_{\{2,3\},2} & \mathbf{P}^{\mathcal{S}}_{\{2,3\},3} \\ \mathbf{P}^{\mathcal{S}}_{\{1,2,3\},1} & \mathbf{P}^{\mathcal{S}}_{\{1,2,3\},2} & \mathbf{P}^{\mathcal{S}}_{\{1,2,3\},3} \end{bmatrix}, \tag{38}$$

where $\mathbf{P}^{\mathcal{S}}_{\mathcal{T},n}$, for $\mathcal{T} \subseteq [\mathsf{N}]$ and $n \in \mathcal{T}$, can be thought of as a linear encoding matrix involving $\mathcal{T}$ files for the $n^{\mathsf{th}}$ file. By symmetry, for $\mathcal{S} \subseteq [3], \mathcal{T} \subseteq [\mathsf{N}], n \in \mathcal{S}$, we can set the rank of $\mathbf{P}^{\mathcal{S}}_{\mathcal{T},n}$ as,

$$\mathsf{rk}(\mathbf{P}^{\mathcal{S}}_{\mathcal{T},n}) = r_{i,j}\mathsf{B}, \quad i = |\mathcal{T}|,\ j = |\mathcal{S}|. \tag{39}$$

Similarly, for $i \in [3]$ and $\{j,\ell\} = [3] \setminus \{i\}$, without loss of generality we can write

$$\mathbf{E}_{i(j,\ell)} = \begin{bmatrix} \mathbf{Q}^{i}_{\{1\},1} & 0 & 0 \\ 0 & \mathbf{Q}^{i}_{\{2\},2} & 0 \\ 0 & 0 & \mathbf{Q}^{i}_{\{3\},3} \\ \mathbf{Q}^{i}_{\{1,2\},1} & \mathbf{Q}^{i}_{\{1,2\},2} & 0 \\ \mathbf{Q}^{i}_{\{1,3\},1} & 0 & \mathbf{Q}^{i}_{\{1,3\},3} \\ 0 & \mathbf{Q}^{\mathcal{S}}_{\{2,3\},2} & \mathbf{Q}^{i}_{\{2,3\},3} \\ \mathbf{Q}^{i}_{\{1,2,3\},1} & \mathbf{Q}^{i}_{\{1,2,3\},2} & \mathbf{Q}^{i}_{\{1,2,3\},3} \end{bmatrix}, \tag{40}$$

where $\mathbf{Q}^{i}_{\mathcal{T},n}$, for $\mathcal{T} \subseteq [\mathsf{N}]$ and $n \in \mathcal{T}$, can be thought of as a linear encoding matrix involving $\mathcal{T}$ files for the $n^{\mathsf{th}}$ file. For $i \in [3], \mathcal{T} \subseteq [\mathsf{N}], n \in \mathcal{T}$, we can set the rank of $\mathbf{Q}^{i}_{\mathcal{T},n}$ as,

$$\mathsf{rk}(\mathbf{Q}^{i}_{\mathcal{T},n}) = q_j\mathsf{B}, \quad j = |\mathcal{T}|. \tag{41}$$

We next solve with Mathematica [11] the LP we obtain from Theorem V.1 with similar tricks as in Section IV. In Mathematica, we transform these symbolic matrices $\mathbf{E}$'s to $\{0,1\}$ matrices, then we perform some linear algebra operations to compute the rank, as detailed in Appendix. We assume demand $d_k = k, k \in [3]$. Due to the symmetry, a number constraints are the same and we can let without loss of generality

$$\lambda_3 := \lambda_{123}, \tag{42}$$
$$\lambda_2 := \lambda_{12} = \lambda_{13} = \lambda_{23}, \tag{43}$$
$$\lambda_1 := \lambda. \tag{44}$$

The final LP, to be optimized over all non-negative $r, q, \lambda$, is

$$3(1 - q_1 - r_{1,1} - 2r_{2,1} - r_{3,1}) - (2\lambda_3 + 3\lambda_2 + \lambda_1), \tag{45a}$$

subject to

$$\lambda_3 \le r_{2,1}, \tag{45b}$$
$$\lambda_2 + \lambda_3 \le q_1 + 2q_2 + r_{1,1} + 2r_{1,2} + r_{2,1} + 3r_{2,2} + r_{3,2}, \tag{45c}$$
$$2\lambda_2 + \lambda_3 \le q_1 + 4q_2 + 2r_{1,1} + 4r_{1,2} + r_{2,1} + 6r_{2,2} + 2r_{3,2}, \tag{45d}$$
$$\lambda_1 + \lambda_2 + \lambda_3 \le q_1 + 4q_2 + 2q_3 + 2r_{1,1} + 6r_{1,2} + 3r_{1,3} + r_{2,1} + 6r_{2,2} + 3r_{2,3} + 2r_{3,2} + r_{3,3}, \tag{45e}$$

and additional cache and file constraints

$$\sum_{j=1}^{\mathsf{N}} \frac{1}{\mathsf{N}} \binom{\mathsf{N}}{j} (r_{1,j} + 2r_{2,j} + r_{3,j} + q_j) \le \frac{\mathsf{M}}{\mathsf{N}}, \tag{45f}$$

$$\sum_{j=1}^{\mathsf{N}} \binom{\mathsf{N}-1}{j-1} (3r_{1,j} + 3r_{2,j} + r_{3,j} + 2q_j) \le 1. \tag{45g}$$

We can further simplify the LP in (45). The constraint for $\lambda_3$ holds equality, i.e., $\lambda_3 = r_{2,1}$. The constraints for $\lambda_2$ and $\lambda_1$ reduces to

$$2\lambda_2 = 2r_{1,1} + 4r_{1,2} + 6r_{2,2} + 2r_{3,2} + q_1 + 4q_2, \tag{46}$$

$$\lambda_1 = 2r_{1,2} + 3r_{1,3} + 3r_{2,3} + r_{3,3}. \quad (47)$$

Thus, the LP objectives becomes

$$\min_{r,q\geq 0} \; 3 - 6r_{1,1} - 8r_{1,2} - 3r_{1,3} - 8r_{2,1} - 9r_{2,2} - 3r_{2,3} - 3r_{3,1} - 3r_{3,2} - r_{3,3} - 4.5q_1 - 6q_2. \quad (48)$$

By solving the LP in (48) subject to (45f) and (45g), a red dash-dotted memory-load tradeoff is attained as shows in Fig 3b. The optimal values of the LP variables and the optimal load for the LP in (48) are shown in Table III.

*Achievablility:* From Table III, when $\mathsf{M} = 1/2$, we find the optimal $r_{1,2} = 1/6, \lambda_2 = \lambda_1 = 1/3$, and the rest of variables are zeros. This result shows first of all that all cache encoding matrices are disjoint and LinP coding involves exactly two files; and secondly, that we can find a coded strategy that can save $3\lambda_2 + \lambda_1$ load compared to uncoded transmission. We use this result to guide our design for an achievable scheme with LinP.

We partition each file into 6 equal-length subfiles as $F_1 = (A_1, A_2, \ldots, A_6)$ and similarly for $F_2$ (whose subfiles are denoted by $B$) and $F_3$ (whose subfiles are denoted by $C$). The LinP cache placement is as follows

$$Z_1 = \begin{bmatrix} A_1 + B_1 \\ A_2 + C_1 \\ B_2 + C_2 \end{bmatrix}, \; Z_2 = \begin{bmatrix} A_3 + B_3 \\ A_4 + C_3 \\ B_4 + C_4 \end{bmatrix}, \; Z_3 = \begin{bmatrix} A_5 + B_5 \\ A_6 + C_5 \\ B_6 + C_6 \end{bmatrix}.$$

Assume demands $d_k = k, k \in [3]$, (and similarly for other demands). The server transmits

$$X = (A_3, A_6, B_1, B_6, C_1, C_4, C_2 - C_3, B_2 - B_5, A_4 - A_5, B_2 + C_2 + A_4 + C_3 + A_5 + B_5).$$

We explain how users decode their desired file. Take user 1 as an example who demands file 1. First, user 1 directly attains $\{A_3, A_6\}$, and use $\{B_1, C_1\}$ to decode $\{A_1, A_2\}$. Next, we know that it has $B_2 + C_2$, then it extracts $B_5 + C_3$ as $(C_2 - C_3 + B_2 - B_5) - (B_2 + C_2)$, and $A_4 + A_5$ as $(B_2 + C_2 + A_4 + C_3 + A_5 + B_5) - (B_5 + C_3 + B_2 + C_2)$. Finally, with the last message $A_4 - A_5$, user 1 can decode out $\{A_4, A_5\}$. Similar for the other users.

As we have 10 multicast messages with 6 sub-packetization, the point $(1/2, 5/3)$ is achievable. The size of finite field $\mathsf{q}$ can be any prime-power number except 2, as the last step requires $A_4 + A_5$ and $A_4 - A_5$ are independent.

### B. Proof of Theorem III.2

*Converse:* We solve again the LP in (48) with constraints in (45g) and (45f) for $\mathsf{N} > 3$. As $\mathsf{N}$ increases, the coefficients multiplying $r_{\cdot,j}$ and $q_j$, for $j \geq 2$, in (45g) and (45f) increase rapidly (i.e., $\frac{1}{\mathsf{N}}\binom{\mathsf{N}}{j} = \binom{\mathsf{N}-1}{j-1}\frac{1}{j}$ is a polynomial function of $\mathsf{N}$ for $j \geq 2$), i.e., non-zero values for those variables consume a lot "resource" within the constraints but contribute little to the objective function, which suggests that they should be set to zero. With this, the LP becomes

$$\min_{r,q} \; 3 - 6r_{1,1} - 8r_{2,1} - 3r_{3,1} - 4.5q_1 \quad (49a)$$

TABLE III: Optimal LP values for $\mathsf{K} = 3 = \mathsf{N}$.

| $\mathsf{M}$ | $[0, \frac{1}{3}]$ | $[\frac{1}{3}, \frac{1}{2}]$ | $[\frac{1}{2}, 1]$ | $[1, 2]$ | $[2, 3]$ |
|---|---|---|---|---|---|
| $r_{1,1}$ | 0 | 0 | $\frac{2\mathsf{M}-1}{3}$ | $\frac{2-\mathsf{M}}{3}$ | 0 |
| $r_{1,2}$ | 0 | $\mathsf{M} - \frac{1}{3}$ | $\frac{1-\mathsf{M}}{3}$ | 0 | 0 |
| $r_{1,3}$ | $\mathsf{M}$ | $1 - 2\mathsf{M}$ | 0 | 0 | 0 |
| $r_{2,1}$ | 0 | 0 | 0 | $\frac{\mathsf{M}-1}{3}$ | $1 - \frac{\mathsf{M}}{3}$ |
| $r_{3,1}$ | 0 | 0 | 0 | 0 | $\mathsf{M} - 2$ |
| $\mathsf{R}$ | $3 - 3\mathsf{M}$ | $\frac{8-6\mathsf{M}}{3}$ | $\frac{7-4\mathsf{M}}{3}$ | $\frac{5-2\mathsf{M}}{3}$ | $\frac{3-\mathsf{M}}{3}$ |

TABLE IV: Optimal LP values for $\mathsf{K} = 3 < \mathsf{N}$.

| $\frac{\mathsf{M}}{\mathsf{N}}$ | $[0, \frac{1}{3}]$ | $[\frac{1}{3}, \frac{2}{3}]$ | $[\frac{2}{3}, 1]$ |
|---|---|---|---|
| $r_{1,1}$ | $\frac{\mathsf{M}}{\mathsf{N}}$ | $\frac{2}{3} - \frac{\mathsf{M}}{\mathsf{N}}$ | 0 |
| $r_{2,1}$ | 0 | $\frac{\mathsf{M}}{\mathsf{N}} - \frac{1}{3}$ | $1 - \frac{\mathsf{M}}{\mathsf{N}}$ |
| $r_{3,1}$ | 0 | 0 | $3\frac{\mathsf{M}}{\mathsf{N}} - 2$ |
| $\mathsf{R}$ | $3 - 6\frac{\mathsf{M}}{\mathsf{N}}$ | $\frac{5}{3} - 2\frac{\mathsf{M}}{\mathsf{N}}$ | $1 - \frac{\mathsf{M}}{\mathsf{N}}$ |

$$\text{s.t } r_{1,1} + 2r_{2,1} + r_{3,1} + q_1 \leq \mathsf{M}/\mathsf{N} \quad (49b)$$

$$3r_{1,1} + 3r_{2,1} + r_{3,1} + 2q_1 \leq 1. \quad (49c)$$

The optimal values of the LP variables and the load are reported in Table IV. We note that $q_1$ is always zero.

*Achievability:* From Table IV we find $q_1 = 0$ and $r_{\cdot,1}$ are non-zeros, which is equivalent to the LP with uncoded placement. Thus, MAN is optimal under LinP when $\mathsf{N} > \mathsf{K} = 3$. The memory-tradeoff is shown in Fig. 3c.

## VI. Conclusions

For the coded caching system with three users, the exact memory-load tradeoff under linear coding placement is found by leveraging a result in [7]. When $\mathsf{N} = \mathsf{K} = 3$, a novel corner point is discovered in this paper. The MAN scheme with uncoded placement is found to be optimal under linear coding placement when $\mathsf{N} > \mathsf{K} = 3$. Open questions include further investigating the small memory regime for $3 = \mathsf{K} \leq \mathsf{N} \leq 5$, and deriving the optimal tradeoff under linear coding placement for arbitrary $(\mathsf{N}, \mathsf{K})$.

## Appendix

## Automatically Compute Matrix Rank

We map a symbolic matrix to another $\{0, 1\}$ matrix; then we transform union and intersection operations over symbolic matrices to linear algebra operations over the $\{0, 1\}$ matrices. With this, we transform the symbolic linear program in Theorem V.1 to a linear program we can solved in Mathematica.

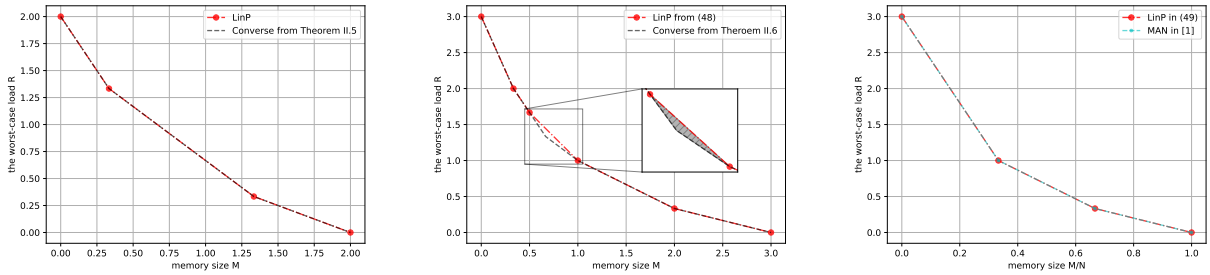
All symbolic independent variables we have so far are

$$\{\mathbf{P}^{\mathcal{S}}_{\mathcal{T},i} : \mathcal{S} \subseteq [3], \mathcal{T} \subseteq [\mathsf{N}], i \in \mathcal{T}\}, \quad (51)$$

$$\{\mathbf{Q}^{\ell}_{\mathcal{T},i} : \ell \in [2], \mathcal{T} \subseteq [\mathsf{N}], i \in \mathcal{T}\}. \quad (52)$$

Other implicit variables are $\{F_v : v \in [3]\}$. Recall that we have file constraint, i.e., fix $v \in [\mathsf{N}]$,

$$\sum_{\mathcal{T}\subseteq[\mathsf{N}], v\in\mathcal{T}} \Big( \sum_{\mathcal{S}\subseteq[3]} \mathsf{rk}(\mathbf{P}^{\mathcal{S}}_{\mathcal{T},v}) + \sum_{i\in[2]} \mathsf{rk}(\mathbf{Q}^{i}_{\mathcal{T},v}) \Big) \leq \mathsf{B}. \quad (53)$$




(a) $\mathsf{N} = 2$. (b) $\mathsf{N} = 3$. (c) $\mathsf{N} \geq 4$.

Fig. 3: The memory-load tradeoff under LinP for $\mathsf{K} = 3$ and various $\mathsf{N}$. Figure 3a shows LinP is optimal in general when $\mathsf{N} = 2$. Figure 3b shows a non-linear coding placement may be needed to close the gap in the grey region when $\mathsf{N} = 3$. Figure 3c shows MAN is optimal under LinP when $\mathsf{N} \geq 4$.

| $\mathbf{P}^1_{\{1\},1}$ | $\cdots$ | $\mathbf{P}^1_{\{1,2\},1}$ | $\mathbf{P}^1_{\{1,2\},2}$ | $\cdots$ | $\mathbf{P}^1_{\{1,2,3\},1}$ | $\mathbf{P}^1_{\{1,2,3\},2}$ | $\mathbf{P}^1_{\{1,2,3\},3}$ | $\cdots$ | $\mathbf{Q}^1_{\{1\},1}$ | $\mathbf{Q}^1_{\{1\},1}$ | $\cdots$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

(50)

For $v \in [\mathsf{N}]$, we can add a complement and make this as a equality,

$$\sum_{\mathcal{T}\subseteq[\mathsf{N}], v\in\mathcal{T}} \Big( \sum_{\mathcal{S}\subseteq[3]} \mathsf{rk}(\mathbf{P}^{\mathcal{S}}_{\mathcal{T},v}) + \sum_{i\in[2]} \mathsf{rk}(\mathbf{Q}^{i}_{\mathcal{T},v}) \Big) + F^c_v = \mathsf{B}. \quad (54)$$

Then, we can express every $F_v$ as a form of $\{\mathbf{P}^{\mathcal{S}}_{\mathcal{T},v}\} \cup \{\mathbf{Q}^{(i)}_{\mathcal{T},v}\} \cup F^c_v$. The number of variables is

$$\mathsf{N} + (2^3 - 1)\Big(\sum_{t\in[\mathsf{N}]} \binom{\mathsf{N}}{t} t\Big) + 2\Big(\sum_{t\in[\mathsf{N}]} \binom{\mathsf{N}}{t} t\Big) = \mathsf{N} + 9\mathsf{N}2^{\mathsf{N}-1}. \quad (55)$$

Plug $\mathsf{N} = 3$ in this equation, we have $3 + 9(3 + 6 + 3) = 111$ independent variables.

Fix $\mathcal{T}, i$, we can express the dependent relationship among $\{\mathbf{Q}^{\ell}_{\mathcal{T},i} : \ell \in [3]\}$ as follows,

$$\mathbf{Q}^3_{\mathcal{T},i} = \mathbf{Q}^1_{\mathcal{T},i} + \mathbf{Q}^2_{\mathcal{T},i}. \quad (56)$$

We now have all symbolic variables we need. Next we can express all independent variables as a unit vector, then transform a symbolic matrix as a $\{0, 1\}$ matrix. For example, suppose we have a symbolic matrix $\mathbf{G}$ given by

$$\mathbf{G} = \begin{bmatrix} \mathbf{P}^1_{\{1\},1} & 0 & 0 \\ \mathbf{P}^1_{\{1,2\},1} & \mathbf{P}^1_{\{1,2\},2} & 0 \\ \mathbf{P}^1_{\{1,2,3\},1} & \mathbf{P}^1_{\{1,2,3\},2} & \mathbf{P}^1_{\{1,2,3\},3} \\ \mathbf{Q}^3_{\{1\},1} & 0 & 0 \end{bmatrix}$$

then it can be transformed to a $\{0, 1\}$ matrix as shows in (50). Naturally, this $\{0, 1\}$ representation supports the union and intersection operations over symbolic matrices. When we union two symbolic matrices, we can concatenate both $\{0, 1\}$ matrices, then eliminate those redundant rows that are linearly dependent on others. When we intersect two symbolic matrices, equivalently we can compute the nullspace of between these $\{0, 1\}$ matrices. Note that these linear algebra operations are based on the real domain.

## REFERENCES


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

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

[3] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1281–1296, 2017.

[4] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 647–663, 2018.

[5] C. Tian, "Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching," *Entropy*, vol. 20, no. 8, p. 603, 2018.

[6] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for users with small buffers," *IET Communications*, vol. 10, no. 17, pp. 2315–2318, 2016.

[7] Y. Yao and S. A. Jafar, "The capacity of 3 user linear computation broadcast," *arXiv preprint arXiv:2206.10049*, 2022.

[8] Y. Ma and D. Tuninetti, "A general coded caching scheme for scalar linear function retrieval," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 2, pp. 321–336, 2022.

[9] K. Wan, H. Sun, M. Ji, D. Tuninetti, and G. Caire, "On the optimal load-memory tradeoff of cache-aided scalar linear function retrieval," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4001–4018, 2021.

[10] H. Sun and S. A. Jafar, "On the capacity of computation broadcast," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3417–3434, 2019.

[11] Wolfram Research, Inc., "Mathematica." Champaign, IL, 2023.