

# Cache-Aided General Linear Function Retrieval

Kai Wan <sup>1,\*</sup>, Hua Sun <sup>2</sup>, Mingyue Ji <sup>3</sup>, Daniela Tuninetti <sup>4</sup> and Giuseppe Caire <sup>1</sup>

<sup>1</sup> Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany; caire@tu-berlin.de

<sup>2</sup> Department of Electrical Engineering, University of North Texas, Denton, TX 76203, USA; hua.sun@unt.edu

<sup>3</sup> Electrical and Computer Engineering Department, University of Utah, Salt Lake City, UT 84112, USA; mingyue.ji@utah.edu

<sup>4</sup> Electrical and Computer Engineering Department, University of Illinois Chicago, Chicago, IL 60607, USA; danielat@uic.edu

\* Correspondence: kai.wan@tu-berlin.de

**Abstract:** Coded Caching, proposed by Maddah-Ali and Niesen (MAN), has the potential to reduce network traffic by pre-storing content in the users' local memories when the network is underutilized and transmitting coded multicast messages that simultaneously benefit many users at once during peak-hour times. This paper considers the linear function retrieval version of the original coded caching setting, where users are interested in retrieving a number of linear combinations of the data points stored at the server, as opposed to a single file. This extends the scope of the authors' past work that only considered the class of linear functions that operate element-wise over the files. On observing that the existing cache-aided scalar linear function retrieval scheme does not work in the proposed setting, this paper designs a novel coded caching scheme that outperforms uncoded caching schemes that either use unicast transmissions or let each user recover all files in the library.

**Keywords:** coded caching; linear function retrieval; uncoded cache placement



**Citation:** Wan, K.; Sun, H.; Ji, M.; Tuninetti, D.; Caire, G. Cache-Aided General Linear Function Retrieval. *Entropy* **2021**, *23*, 25. <https://dx.doi.org/10.3390/e23010025>

Received: 7 November 2020

Accepted: 22 December 2020

Published: 26 December 2020

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Content caching is an efficient technique to handle the increase of requests for massive amounts of data and content over communication networks. By leveraging low-cost memory components at the user sides, caching reduces peak-time traffic by prefetching contents closer to users during off-peak time, thereby reducing the transmission delay or equivalently increasing the bandwidth in communication systems. Traditional caching techniques aim at prefetching popular content by predicting the user demands, thus realizing a “local caching gain” (i.e., that scales with the amount of local memory) [1]. Maddah-Ali and Niesen (MAN) showed that it is possible to actually attain a “global caching gain” (i.e., that scales with the global amount of memory in the network) by using codes [2]. The idea is that, if a single transmission can serve a number of users simultaneously, the network load can be reduced by the same factor thus speeding-up communications significantly.

In the MAN setting, a server has a library of  $N$  files and broadcasts to  $K$  users through an error-free shared-link. Each user has a cache of size of at most  $M$  files. The MAN scheme consists of two phases: placement phase, where the server pushes content from the library to the local caches without knowledge of user future demands, and delivery phase, where each user requests one file and the server broadcasts coded packets such that each user can correctly recover its desired file. The objective is to minimize the worst-case load over all possible user demands, that is, the number of files that must be communicated so that any demands can be satisfied. The MAN scheme is optimal under the constraint of uncoded cache placement (i.e., each user directly stores a collection of segments of the library files in its cache) when  $N \geq K$  [3,4]. By removing the redundant transmissions in the MAN scheme when a file is requested multiple times, Yu, Maddah-Ali, and Avestimehr

(YMA) derived a scheme that is optimal under the constraint of uncoded cache placement for  $N < K$  [5]. In general, the YMA scheme is order optimal to within a factor of 2 [6], that is, coded placement can at best half the load of the YMA scheme.

On the motivation that linear and multivariate polynomial queries naturally arise in modern engineering problems and deep learning algorithms such as matrix-vector, matrix-matrix multiplications, in [7] the authors posed the question of what is the optimal worst-case load when the cache-aided users are interested in retrieving a scalar linear function of the files rather than a single file. For the class of functions considered in [7], which are restricted to operate element-wise on the file entries, it was surprisingly shown that the YMA load can be achieved, that is, there is no penalty in terms of load in retrieving scalar linear functions under the constraint of uncoded cache placement. It was noted in [7] that the proposed scalar linear function scheme can be extended to all scenarios to which the original MAN scheme has been extended, such as for example demand-private retrieval [8] and Device-to-Device networks [9,10]. In addition, the scalar linear function scheme [7] can be used as a building block to provide demand-privacy and content-security against colluding users [11,12].

In this paper, we move to a more general case of cache-aided linear function retrieval than in [7], where users can request general linear combinations of all symbols in the library, and not necessarily restricted to operate element-wise on the file entries. For example, each user aims to compute some statistics of a bunch of data such as local weighted averages (which are general linear functions) of the data; these are very common tasks in many applications depending on the data and on the weights.

Instead, each user may want to compute some statistics of a bunch of data such as average, or compute local weighted averages (which are general linear functions) of the data. We think that it is a very common task in many applications depending on the data and on the weights. So in our paper, if the Academic Editor agrees, we will replace the application in deep neural networks by the application in computing local weighted averages.

Besides the novel and realistic problem formulation, our main contributions are as follows. We first introduce a baseline scheme that either lets each user recover all the symbols in the library or uses unicast transmissions to satisfy each user. The main challenge to implement a coded caching strategy in this problem is that each symbol in a user's demand is a linear combination of all the symbols in the library. Inspired by the grouping coded caching strategy in [13], which was used to reduce the sub-packetization level (The sub-packetization level is the smallest file length necessary to realize an achievable scheme.), we propose a scheme that treats the demand of each user as a matrix-vector multiplication and uses the grouping strategy to generate multicast messages after possibly performing invertible linear matrix operations. The proposed scheme outperforms the baseline scheme in all parameter regimes.

### 1.1. Paper Organization

The rest of this paper is organized as follows. Section 2 formulates the shared-link cache-aided general linear function retrieval problem. Section 3 provides the main result of this paper. Section 4 provides some numerical evaluations. Section 5 concludes the paper. Some proofs may be found in Appendices.

### 1.2. Notation Convention

Calligraphic symbols denote sets, bold symbols denote vectors and matrices, and sans-serif symbols denote system parameters. We use  $|\cdot|$  to represent the cardinality of a set or the length of a vector;  $[a : b] := \{a, a + 1, \dots, b\}$  and  $[n] := [1 : n]$ ;  $\oplus$  represents bit-wise XOR;  $[a]^+ := \max\{a, 0\}$ ;  $\mathbb{F}_q$  represents a finite field with order  $q$ ;  $\mathbf{A}^T$  and  $\mathbf{A}^{-1}$  represent the transpose and the inverse of matrix  $\mathbf{A}$ , respectively;  $\text{rank}_q(\mathbb{A})$  represents the rank of matrix  $\mathbb{A}$  on field  $\mathbb{F}_q$ ;  $\mathbf{I}_n$  represents the identity matrix with dimension  $n \times n$ ;

$(\mathbf{A})_{m \times n}$  represents the dimension of matrix  $\mathbf{A}$  is  $m \times n$ ; we let  $\binom{x}{y} = 0$  if  $x < 0$  or  $y < 0$  or  $x < y$ .

## 2. System Model

Different from [7], here we consider the case where the users' desired linear functions are no longer scalar or operating element-wise across the files entries, thus we consider the whole library as a single file.

The  $(K, F, L, q)$  shared-link cache-aided general linear function retrieval problem consists of a central server with access to a library of  $F$  independent and identically distributed (i.i.d.) symbols over a finite field  $\mathbb{F}_q$ , denoted by  $\mathbf{w} = (w_1, \dots, w_F)^T \in (\mathbb{F}_q)^F$ . We often treat  $\mathbf{w}$  as a column vector, which should be clear from the context. The server is connected to  $K$  cache-aided users through an error-free shared-link. The system has two phases.

- In the **placement phase**, the server pushes up to  $M$  symbols into the local cache of each user, where  $M \in [0 : F]$ , without knowing what the users will demand later. The cached content of user  $k \in [K]$  is denoted by

$$Z_k = \phi_k(\mathbf{w}), \quad (1)$$

where  $\phi_k$  is the placement function for user  $k$  defined as

$$\phi_k : (\mathbb{F}_q)^F \rightarrow (\mathbb{F}_q)^M, \quad k \in [K]. \quad (2)$$

$M$  is referred to as the cache (or memory) size. If each user directly copies  $M$  symbols from the library into its cache, the cache placement is said to be uncoded.

- In the **delivery phase**, each user wants to retrieve  $L$  linear combinations of all the symbols in the library, where  $L \in [1 : F]$ . The demand of user  $k \in [K]$  is represented by the matrix  $\mathbf{D}_k \in (\mathbb{F}_q)^{L \times F}$ , meaning user  $k$  aims to retrieve

$$\mathbf{y}_k = \mathbf{D}_k \mathbf{w} \in (\mathbb{F}_q)^L, \quad (3)$$

Let the collection of all demand matrices be  $\mathbf{D} := [\mathbf{D}_1; \dots; \mathbf{D}_K] \in (\mathbb{F}_q)^{KL \times F}$ . We assume that the server and all users know  $\mathbf{D}$  which is communicated on a separate channel, thus not impacting the downlink load next—see also Remark 4. (Notice that differently from the cache-aided matrix multiplication problem in [14], where the matrix on the each side of the desired multiplication is one of the library files, in this paper each user  $k \in [K]$  desires  $\mathbf{D}_k \mathbf{w}$  where  $\mathbf{D}_k$  is known by all the users in the delivery phase and  $\mathbf{w}$  represents the vector of all symbols in the library.)

According to all the users' demand matrix  $\mathbf{D}$ , the server broadcasts the message

$$X = \psi(\mathbf{D}, \mathbf{w}), \quad (4)$$

where  $\psi$  is the encoding function

$$\psi : (\mathbb{F}_q)^{KL \times F} \times (\mathbb{F}_q)^F \rightarrow (\mathbb{F}_q)^R, \quad (5)$$

for some  $R \in [0 : F]$ .  $R$  is referred to as the load.

**Achievability:** For the  $(K, F, L, q)$  shared-link cache-aided general linear function retrieval problem, we say that the pair  $(M, R)$  is achievable if for any possible demand  $\mathbf{D}$  there exist placement functions in (2) and a delivery function in (5) such that

$$H(\mathbf{D}_k \mathbf{w} | \mathbf{D}, Z_k, X) = 0, \quad \forall k \in [K]. \quad (6)$$

*Optimal memory-load tradeoff:* For the  $(K, F, L, q)$  shared-link cache-aided general linear function retrieval problem, the objective is to determine the minimum worst-case downlink load (or load for simplicity) defined as

$$R^*(M) = \min_{\phi_1, \dots, \phi_K, \psi} \{R : (M, R) \text{ is achievable}\}. \quad (7)$$

*Optimal memory-load tradeoff in the limit for large file size:* Since solving the problem in (7) for any given  $(K, F, L, q)$  is challenging, in the following we shall consider the regime where the file size  $F$  is as large as desired and we thus let the system parameters scale with the file length as follows

$$M := \mu F, \mu \in [0, 1], \quad (8)$$

$$L := \lambda F, \lambda \in [0, 1], \quad (9)$$

$$R := \rho F, \rho \in [0, 1]. \quad (10)$$

For fixed  $(K, \lambda)$  we aim to characterize the minimum worst-case normalized downlink load (or normalized load for simplicity)

$$\rho^*(\mu) = \min_{\phi_1, \dots, \phi_K, \psi} \{\rho : (M, R) = (\mu F, \rho F) \text{ is achievable for some } (F, q)\}. \quad (11)$$

**Remark 1** (Relationship to [7]). The cache-aided scalar linear function retrieval problem in [7] is a special case of the formulation here. More precisely, let  $F = NL$  (i.e.,  $\frac{1}{N} = \lambda$ ), where  $N$  indicates the number of files and  $\lambda F$  is the file length. The demand of user  $k \in [K]$  is represented by the vector  $\mathbf{y}_k = (y_{k,1}, y_{k,2}, \dots, y_{k,N}) \in (\mathbb{F}_q)^N$  by which we mean that the user is requesting

$$\mathbf{D}_k = [y_{k,1}\mathbf{I}_L, y_{k,2}\mathbf{I}_L, \dots, y_{k,N}\mathbf{I}_L] \in (\mathbb{F}_q)^{L \times NL}, \quad (12)$$

where  $\mathbf{I}_n$  is the identity matrix with dimension  $n \times n$ . In the restricted setting where the demands are as in (12) the optimal load under the constraint of uncoded cache placement is the lower convex envelop of the points

$$\left(\frac{M}{L}, \frac{R_{\text{scalar}}}{L}\right) = \left(\frac{N}{K}, \frac{\binom{K}{t+1} - \binom{K - \min\{K, N\}}{t+1}}{\binom{K}{t}}\right), t \in [0 : K], \quad (13)$$

$$\iff (\mu, \rho_{\text{scalar}}) = \left(\frac{t}{K}, \lambda \frac{\binom{K}{t+1} - \binom{K - \min\{K, N\}}{t+1}}{\binom{K}{t}}\right), t \in [0 : K], \quad (14)$$

where for a given value of  $t$  in (13) the subpacketization level  $L$  must be an integer multiple of  $\binom{K}{t}$ .

**Remark 2** (A minrank solution). For the  $(K, F, L, q)$  shared-link cache-aided general linear function retrieval problem, the best linear scheme, inspired by [15,16], is as follows. Linear placement: user  $k \in [K]$  caches  $Z_k = \mathbf{P}_k \mathbf{w} \in (\mathbb{F}_q)^M$  for some  $\mathbf{P}_k \in (\mathbb{F}_q)^{M \times F}$ . Linear delivery: the server sends, in the worst case, a number of symbols given by

$$R_{\text{minrank}} = \min_{\mathbf{P}_1, \dots, \mathbf{P}_K} \max_{\mathbf{D}_1, \dots, \mathbf{D}_K} \min_{\mathbf{T}_1, \dots, \mathbf{T}_K} \text{rank} \begin{bmatrix} \mathbf{D}_1 + \mathbf{T}_1 \mathbf{P}_1 \\ \mathbf{D}_2 + \mathbf{T}_2 \mathbf{P}_2 \\ \vdots \\ \mathbf{D}_K + \mathbf{T}_K \mathbf{P}_K \end{bmatrix}, \quad (15)$$

where  $\mathbf{T}_k \in (\mathbb{F}_q)^{L \times M}$ ,  $k \in [K]$ .

Solving the minrank problem in (15) is hard [15,16], thus in the following we shall design a scheme with lower complexity.

**Remark 3** (A baseline scheme). For the  $(K, F, L, q)$  shared-link cache-aided general linear function retrieval problem, the load

$$\begin{aligned} R_{\text{baseline}} &= \min\{KL, F - M\} \\ \Longleftrightarrow \rho_{\text{baseline}} &= \min\{K\lambda, 1 - \mu\}, \end{aligned} \quad (16)$$

can be achieved by an uncoded caching strategy as follows.

- In order to achieve the load  $KL$ , we transmit one by one the elements of  $\mathbf{y}_k, k \in [K]$ , in (3). The main limitation of this unicast transmission scheme is the lack of multicast gain.
- In order to achieve  $F - M$  we let each user recover all the symbols in the library. In the placement phase, each user caches the first  $M$  symbols in the library. In the delivery phase, the server transmits all the remaining  $F - M$  symbols. The main limitation of this scheme is that, if  $L < F - M$ , the users do not need to recover all the symbols in the library in order to retrieve their desired function.

The main contribution of this paper is to find schemes that, despite the lack of structure on the demand matrices in general, achieve a smaller load than (16).

**Remark 4** (Uplink and downlink loads). Besides downlink load, uplink load is also considered in the distributed matrix-matrix multiplication problem [17–19]. In this work, the communication cost of uploading the demand matrix to the server is not a focus, i.e., we assume that each user communicates the whole demand matrix to the server and all other users on a separate channel that is not the bottleneck in the system. This assumption can be also justified as follows. Let  $\mathcal{D}(k)$  denotes the set of possible demand matrices of user  $k \in [K]$ , referred to as demand range, that is, user  $k$  chooses one matrix in  $\mathcal{D}(k)$  as its demand. We assume that  $\mathcal{D}(k)$  is known by the server and all users. The communication cost to let the server and the other users know the realization of the demand matrix is negligible compared to the number of transmissions from the server if  $\sum_{k \in [K]} \log_q(|\mathcal{D}(k)|) \ll F$ .

### 3. Main Result

Based on Remark 3, the main challenge is to design a coded caching strategy that (i) lets each user directly recover the desired linear combinations, instead of recovering all the library symbols, and (ii) attains coded caching gain, as opposed to serving the users one-by-one with unicast transmissions. The main contribution of this paper is the following theorem, which is proved in Appendix A.

**Theorem 1.** For the  $(K, \lambda)$  shared-link cache-aided general linear function retrieval problem, we have:

- if  $\mu = \alpha \frac{g-1}{g} + (1-\alpha) \frac{g}{g+1}$  where  $g \in [K-1]$  and  $\alpha \in [0, 1]$ , the following normalized load is achievable

$$\rho_{\text{ach}} := \begin{cases} \left\lceil \frac{K}{g} \right\rceil \lambda, & \text{if } \frac{\alpha}{g} \geq \left\lceil \frac{K}{g} \right\rceil \lambda \\ \min\{\rho_1, \rho_2\} & \text{if } \frac{\alpha}{g} \leq \left\lceil \frac{K}{g} \right\rceil \lambda \end{cases}, \quad (17)$$

$$\rho_1 := \frac{\alpha}{g} + \min\left\{\left\lceil \frac{K}{g+1} \right\rceil \lambda, \frac{(1-\alpha)}{g+1}\right\}, \quad (18)$$

$$\rho_2 := \left\lceil \frac{K}{g} \right\rceil \min\left\{\frac{\alpha}{g}, \lambda\right\} + \min\left\{\left\lceil \frac{K}{g+1} \right\rceil \left[\lambda - \frac{\alpha}{g}\right]^+, \frac{(1-\alpha)}{g+1}\right\}; \quad (19)$$

- if  $\mu = \alpha \frac{K-1}{K} + (1-\alpha)$  where  $\alpha \in [0, 1]$ , the following normalized load is achievable

$$\rho_{\text{ach}} = \rho_3 = \min\left\{\frac{\alpha}{K}, \lambda\right\}. \quad (20)$$

□

Next, we provide the intuition behind the proposed scheme in Theorem 1, which is based on three ingredients:

1. We start by the achievable scheme for (20) with  $\alpha = 1$ . We aim to design the cache placement such that each user caches a fraction  $\frac{K-1}{K}$  of the file and the uncached part of file by this user is known by the remaining  $K - 1$  users. With this cache placement, the delivery consists of a single multicast message with multicasting gain  $K$ . More precisely, the construction of the proposed scheme is as follows. recalling that, in Remark 1 with  $t = K - 1$ , each user misses a fraction  $1/K$  of each file and that missing fraction is known by the remaining  $K - 1$  users; with  $t + 1 = K$ , the delivery consists of a single multicast message with multicasting gain  $K$  that is the sum of each user's missing fraction of the demanded file. In our context, this idea becomes the following scheme.

Assume  $K$  divides  $F$ . We use here a Matlab-like notation for submatrices. The library is partitioned into  $K$  equal length subfiles as follows

$$\mathcal{I}_k := \left[ (k-1)\frac{F}{K} + 1 : k\frac{F}{K} \right], k \in [K], \quad (21)$$

$$\mathbf{w}_k := \mathbf{w}(\mathcal{I}_k) \in (\mathbb{F}_q)^{\frac{F}{K}}, k \in [K], \quad (22)$$

$$\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K); \quad (23)$$

user  $k \in [K]$  caches  $Z_k = (\mathbf{w}_j : j \in [K] \setminus \{k\})$ ; the server delivers the multicast message

$$X = \begin{cases} \sum_{k \in [K]} \mathbf{D}_{k; : \mathcal{I}_k} \mathbf{w}_k \in (\mathbb{F}_q)^L, & \text{if } \frac{F}{K} > L \\ \sum_{k \in [K]} \mathbf{w}_k \in (\mathbb{F}_q)^{\frac{F}{K}}, & \text{if } \frac{F}{K} \leq L \end{cases} \quad (24)$$

where  $\mathbf{D}_{k; : \mathcal{I}_k}$  represents the sub-matrix of  $\mathbf{D}_k$  including the columns with indices in  $\mathcal{I}_k$ . In  $X$ , each user  $k \in [K]$  knows all but the requested vector

$$\begin{cases} \mathbf{D}_{k; : \mathcal{I}_k} \mathbf{w}_k, & \text{if } \frac{F}{K} > L; \\ \mathbf{w}_k, & \text{if } \frac{F}{K} \leq L, \end{cases}$$

such that user  $k$  can recover either of them. Thus an achieved normalized memory-load tradeoff is

$$(\mu, \rho) = \left( 1 - \frac{1}{K}, \min \left\{ \frac{1}{K}, \lambda \right\} \right). \quad (25)$$

2. We then introduce the achievable scheme for (17) with  $\alpha \in \{0, 1\}$ . Assume now the  $K$  users are portioned into  $g$  groups of  $\left\lceil \frac{K}{g} \right\rceil$  users each, where  $g \in [K - 1]$ . Let the users in the same group share the same cache content and recover all the linear combinations demanded by the users in the group. Then the normalized memory-load tradeoff is as in (25) but with  $K$  replaced by  $g$  and  $L$  replaced by  $\left\lceil \frac{K}{g} \right\rceil L$ . Therefore, we get that the following normalized memory-load points are achievable

$$(\mu, \rho) = \left( 1 - \frac{1}{g}, \min \left\{ \frac{1}{g}, \lambda \left\lceil \frac{K}{g} \right\rceil \right\} \right), g \in [K]. \quad (26)$$

3. The rest of the proof of Theorem 1 consists of a method to 'interpolate' among the points in (26), as explained in Appendix A. Unlike cache-aided scalar linear function retrieval in [7], the difficulty in the considered problem is that connecting two normalized memory-load points by a line segment is generally impossible. The main

reason is that if we partition  $\mathbf{w}$  as  $\mathbf{w} = [\mathbf{w}'; \mathbf{w}'']$  and use a different cache placement strategy on each part, each demanded function  $\mathbf{D}_k \mathbf{w}$  is in the form

$$\mathbf{D}_k \mathbf{w} = \mathbf{D}'_k \mathbf{w}' + \mathbf{D}''_k \mathbf{w}''; \quad (27)$$

thus it cannot be divided into two separate parts, where the first part only contains the linear combinations of  $\mathbf{w}'$  and the second part only contains the linear combinations of  $\mathbf{w}''$ . An example to highlight this limitation and our approach to overcome it is provided at the end of this section.

**Remark 5** (Comparison to the baseline scheme). We show here that the proposed scheme in Theorem 1 outperforms the baseline scheme in (3).

- Case  $\mu = \alpha \frac{g-1}{g} + (1-\alpha) \frac{g}{g+1}$  where  $g \in [K]$  and  $\alpha \in [0, 1]$ : From (17) and (19), it can be seen that

$$\rho_{ach} \leq \left\lceil \frac{K}{g} \right\rceil \lambda \leq K\lambda. \quad (28)$$

From (17) and (18), it can be seen that

$$\rho_{ach} \leq \frac{\alpha}{g} + \frac{1-\alpha}{g+1} = 1 - \mu. \quad (29)$$

Hence, from (28) and (29), we can prove  $\rho_{ach} \leq \rho_{baseline}$  in this case.

- Case  $\mu = \alpha \frac{K-1}{K} + (1-\alpha)$  where  $\alpha \in [0, 1]$ : Since in this case  $\frac{\alpha}{K} = 1 - \mu$ , from (20) we can prove  $\rho_{ach} \leq \rho_{baseline}$  in this case.

**Remark 6** (Connection to Remark 1). For the proposed scheme achieving (25), the cache placement is the same as the cache-aided scalar linear function retrieval scheme in Remark 1 with  $t = K - 1$ .

ecalling that, in Remark 1 with  $t = K - 1$ , each user misses a fraction  $1/K$  of each file and that missing fraction is known by the remaining  $K - 1$  users; with  $t + 1 = K$ , the delivery consists of a single multicast message with multicasting gain  $K$  that is the sum of each user's missing fraction of the demanded file. In our context, this idea becomes the following scheme.

Notice that, for the considered cache-aided general linear function retrieval problem where  $\mu = \frac{t}{K}$  and  $t \in [K]$ , we could use the cache-aided scalar linear function retrieval scheme in Remark 1 to deliver  $\binom{K}{t+1}$  pieces of the requested vectors. The scheme would achieve

$$(\mu, \rho) = \left( \frac{t}{K}, \lambda \binom{K}{t+1} \right), t \in [K], \quad (30)$$

which reduces to (25) for  $t = K - 1$ . By the grouping argument we would achieve

$$(\mu, \rho') = \left( \frac{t}{g}, \lambda \left\lceil \frac{K}{g} \right\rceil \binom{g}{t+1} \right), t \in [g], g \in [K]. \quad (31)$$

Let then fix one  $g \in [K]$  and one  $t \in [g - 2]$ , and analyse the achieved normalized load in (31). We will show that

$$\rho' = \lambda \left\lceil \frac{K}{g} \right\rceil \binom{g}{t+1} \geq \rho_{baseline}. \quad (32)$$

as follows. It can be seen that

$$\lambda \left\lceil \frac{K}{g} \right\rceil \binom{g}{t+1} \geq K\lambda \frac{\binom{g}{t+1}}{g} \quad (33)$$

$$\geq K\lambda \quad (34)$$

$$\geq \rho_{\text{baseline}} \quad (35)$$

where (34) follows since  $t \in [g-2]$  and thus  $\binom{g}{t+1} \geq g$ . This shows that, with the exception for the normalized memory-load points with  $t = g-1$ , the scheme in (31) is inferior to the baseline scheme in (16), and will thus not be pursued in the rest of the paper.

We finish this section with an example to illustrate the main ideas of the proposed scheme.

**Example 1.** We consider a system with  $K = 6$  users, cache fraction  $\mu = \frac{47}{72}$ , and demand fraction  $\lambda = \frac{1}{12}$ . It can be seen that

$$\mu = \frac{47}{72} = \alpha \frac{g-1}{g} + (1-\alpha) \frac{g}{g+1} \Big|_{\alpha=\frac{1}{12}, g=2}. \quad (36)$$

*Placement Phase.* It can be seen that the memory size is between  $\mu_1 = \frac{g-1}{g} = \frac{1}{2}$  and  $\mu_2 = \frac{g}{g+1} = \frac{2}{3}$ . We partition  $\mathbf{w}$  into two parts as  $\mathbf{w} = [\mathbf{w}^1; \mathbf{w}^2]$  where  $\mathbf{w}^1 \in (\mathbb{F}_q)^{F/12}$  and  $\mathbf{w}^2 \in (\mathbb{F}_q)^{11F/12}$ . Furthermore,

- $\mathbf{w}^1$  is partitioned into two equal-length subfiles,  $\mathbf{w}^1 = [\mathbf{w}_{\{1\}}^1; \mathbf{w}_{\{2\}}^1]$ , each of which has  $\frac{F}{24}$  symbols. We divide the 6 users into 2 groups where  $\mathcal{G}_1^1 = \{1, 3, 5\}$  and  $\mathcal{G}_2^1 = \{2, 4, 6\}$ . We let users in  $\mathcal{G}_1^1$  cache  $\mathbf{w}_{\{1\}}^1$  and let users in  $\mathcal{G}_2^1$  cache  $\mathbf{w}_{\{2\}}^1$ .
- $\mathbf{w}^2$  is partitioned into three equal-length subfiles,  $\mathbf{w}^2 = [\mathbf{w}_{\{1,2\}}^2; \mathbf{w}_{\{1,3\}}^2; \mathbf{w}_{\{2,3\}}^2]$ , each of which has  $\frac{11F}{36}$  symbols. We divide the 6 users into 3 groups, where  $\mathcal{G}_1^2 = \{1, 4\}$ ,  $\mathcal{G}_2^2 = \{2, 5\}$ , and  $\mathcal{G}_3^2 = \{3, 6\}$ . We let users in  $\mathcal{G}_1^2$  cache  $\mathbf{w}_{\{1,2\}}^2$  and  $\mathbf{w}_{\{1,3\}}^2$ , let users in  $\mathcal{G}_2^2$  cache  $\mathbf{w}_{\{1,2\}}^2, \mathbf{w}_{\{2,3\}}^2$ , and let users in  $\mathcal{G}_3^2$  cache  $\mathbf{w}_{\{1,3\}}^2$  and  $\mathbf{w}_{\{2,3\}}^2$ .

Each user caches  $\frac{F}{24} + \frac{2 \times 11F}{36} = \frac{47F}{72}$  symbols, thus satisfying the memory size constraint.

*Delivery Phase.* With some permutation on the rows of  $\mathbf{w}$ , the demand of user 1 can be expressed as

$$\mathbf{D}_1 \mathbf{w} = \mathbf{D}_{1,\{1\}} \mathbf{w}_{\{1\}}^1 + \mathbf{D}_{1,\{1,2\}} \mathbf{w}_{\{1,2\}}^2 + \mathbf{D}_{1,\{1,3\}} \mathbf{w}_{\{1,3\}}^2 + \mathbf{D}_{1,\{2\}} \mathbf{w}_{\{2\}}^1 + \mathbf{D}_{1,\{2,3\}} \mathbf{w}_{\{2,3\}}^2. \quad (37)$$

User 1 can recover  $\mathbf{D}_{1,\{1\}} \mathbf{w}_{\{1\}}^1 + \mathbf{D}_{1,\{1,2\}} \mathbf{w}_{\{1,2\}}^2 + \mathbf{D}_{1,\{1,3\}} \mathbf{w}_{\{1,3\}}^2$  from its cache, and similarly for the other users. Thus in the delivery phase, the users need to recover

$$\mathbf{B}_1 := \mathbf{D}_{1,\{2\}} \mathbf{w}_{\{2\}}^1 + \mathbf{D}_{1,\{2,3\}} \mathbf{w}_{\{2,3\}}^2, \quad (38)$$

$$\mathbf{B}_2 := \mathbf{D}_{2,\{1\}} \mathbf{w}_{\{1\}}^1 + \mathbf{D}_{2,\{1,3\}} \mathbf{w}_{\{1,3\}}^2, \quad (39)$$

$$\mathbf{B}_3 := \mathbf{D}_{3,\{2\}} \mathbf{w}_{\{2\}}^1 + \mathbf{D}_{3,\{1,2\}} \mathbf{w}_{\{1,2\}}^2, \quad (40)$$

$$\mathbf{B}_4 := \mathbf{D}_{4,\{1\}} \mathbf{w}_{\{1\}}^1 + \mathbf{D}_{4,\{2,3\}} \mathbf{w}_{\{2,3\}}^2, \quad (41)$$

$$\mathbf{B}_5 := \mathbf{D}_{5,\{2\}} \mathbf{w}_{\{2\}}^1 + \mathbf{D}_{5,\{1,3\}} \mathbf{w}_{\{1,3\}}^2, \quad (42)$$

$$\mathbf{B}_6 := \mathbf{D}_{6,\{1\}} \mathbf{w}_{\{1\}}^1 + \mathbf{D}_{6,\{1,2\}} \mathbf{w}_{\{1,2\}}^2. \quad (43)$$

If we treat each sum in (38)–(43) as a block and use the MAN strategy to delivery these blocks, we would transmit  $\mathbf{B}_1 + \mathbf{B}_2, \mathbf{B}_3 + \mathbf{B}_4, \mathbf{B}_5 + \mathbf{B}_6$  for a total of  $\frac{F}{4}$  symbols. Hence, the scheme achieves the same normalized load as the proposed scheme in (26) with  $\mu_1 = \frac{1}{2}$ ;

in other words, a portion of the memory of size  $\mu - \mu_1 = \frac{47}{72} - \frac{1}{2} = \frac{11}{72}$  would be wasted. We next propose two novel schemes to let each user recover its desired sum in (38)–(43) while leveraging the whole memory.

The solution that achieves  $\rho_1$  in (18). Focus on the demanded sum of user 1 in (38). The key idea is to let user 1 recover  $\mathbf{D}_{1,\{2\}} \mathbf{w}_{\{2\}}^1$  and  $\mathbf{D}_{1,\{2,3\}} \mathbf{w}_{\{2,3\}}^2$  separately. In particular,

- For the first term in  $\mathbf{B}_1$  in (38), since the dimension of  $\mathbf{D}_{1,\{2\}}$  is  $\frac{F}{12} \times \frac{F}{24}$  and the sub-demand matrix  $\mathbf{D}_{1,\{2\}}$  is known by each user, we let user 1 directly recover  $\mathbf{w}_{\{2\}}^1$ , which contains  $\frac{F}{24}$  symbols, and then compute  $\mathbf{D}_{1,\{2\}} \mathbf{w}_{\{2\}}^1$ . Similarly, we let users 3, 5 recover  $\mathbf{w}_{\{2\}}^1$ , and users 2, 4, 6 recover  $\mathbf{w}_{\{1\}}^1$ . Thus in the delivery phase, the server transmits

$$\mathbf{w}_{\{1\}}^1 + \mathbf{w}_{\{2\}}^1, \quad (44)$$

for a total of  $\frac{F}{24}$  symbols, where users 1, 3, 5 know  $\mathbf{w}_{\{1\}}^1$  and users 2, 4, 6 know  $\mathbf{w}_{\{2\}}^1$ .

- For the second term in  $\mathbf{B}_1$  in (38), since the dimension of  $\mathbf{D}_{1,\{2,3\}}$  is  $\frac{F}{12} \times \frac{11F}{36}$  and the sub-demand matrix  $\mathbf{D}_{1,\{2,3\}}$  is known by each user, user 1 needs to recover all symbols in  $\mathbf{D}_{1,\{2,3\}} \mathbf{w}_{\{2,3\}}^2$ . We denote  $\mathbf{C}_{1,\{2,3,5,6\}}^2 := \mathbf{D}_{1,\{2,3\}} \mathbf{w}_{\{2,3\}}^2$  since it is known by users 2, 3, 5, 6. Hence, in order to let each user recover the first term in its desired sum, the server transmits

$$\mathbf{C}_{1,\{2,3,5,6\}}^2 + \mathbf{C}_{2,\{1,3,4,6\}}^2 + \mathbf{C}_{3,\{1,2,4,5\}}^2, \quad (45)$$

$$\mathbf{C}_{4,\{2,3,5,6\}}^2 + \mathbf{C}_{5,\{1,3,4,6\}}^2 + \mathbf{C}_{6,\{1,2,4,5\}}^2, \quad (46)$$

for a total of  $\frac{F}{6}$  symbols.

Hence, in the delivery phase the server transmits  $\frac{F}{24} + \frac{F}{6} = \frac{5F}{24}$  symbols, and the normalized load is  $\rho_1 = \frac{5}{24}$ , which coincides with (18).

The solution that achieves  $\rho_2$  in (19). The idea is to partition each user's demand into two parts after having removed its cached content, where the partition is the result of a clever invertible linear transformation; we then have two steps, one for each of the two parts.

We first focus on the demand of user 1 in (38), i.e.,

$$\mathbf{B}_1 = \mathbf{D}_{1,\{2\}} \mathbf{w}_{\{2\}}^1 + \mathbf{D}_{1,\{2,3\}} \mathbf{w}_{\{2,3\}}^2 = \begin{bmatrix} \mathbf{D}_{1,\{2\}} & \mathbf{D}_{1,\{2,3\}} \end{bmatrix} \begin{bmatrix} \mathbf{w}_{\{2\}}^1 \\ \mathbf{w}_{\{2,3\}}^2 \end{bmatrix}. \quad (47)$$

The main strategy here is to take a linear transformations of (47) as follows

$$\mathbf{B}'_1 = (\mathbf{T}_1)_{\frac{F}{12} \times \frac{F}{12}} \begin{bmatrix} (\mathbf{D}_{1,\{2\}})_{\frac{F}{12} \times \frac{F}{24}} & (\mathbf{D}_{1,\{2,3\}})_{\frac{F}{12} \times \frac{11F}{36}} \end{bmatrix} \begin{bmatrix} \left( \mathbf{w}_{\{2\}}^1 \right)_{\frac{F}{24} \times 1} \\ \left( \mathbf{w}_{\{2,3\}}^2 \right)_{\frac{11F}{36} \times 1} \end{bmatrix}, \quad (48)$$

where  $\mathbf{T}_1$  is full-rank and the bottom  $\frac{F}{12} - \frac{F}{24} = \frac{F}{24}$  symbols in  $\mathbf{B}'_1$  are linear combinations of  $\mathbf{w}_{\{2,3\}}^2$  only (i.e., these linear combinations do not contain any term in  $\mathbf{w}_{\{2\}}^1$ ). This is possible because  $\mathbf{B}_1$  contains  $\frac{F}{12}$  linear combinations of all symbols in  $[\mathbf{w}_{\{2\}}^1; \mathbf{w}_{\{2,3\}}^2]$ , while  $\mathbf{w}_{\{2\}}^1$  contains  $\frac{F}{24}$  symbols. Hence, we can re-express  $\mathbf{B}'_1$  as

$$\mathbf{B}'_1 = \begin{bmatrix} \left( \mathbf{B}'_{1,\{2,6\}} \right)_{\frac{F}{24} \times 1} \\ \left( \mathbf{B}'_{1,\{2,3,5,6\}} \right)_{\frac{F}{24} \times 1} \end{bmatrix}, \quad (49)$$

where  $\mathbf{B}'_{1,\{2,6\}}$  contains  $\frac{F}{24}$  linear combinations of the symbols in  $\mathbf{w}_{\{2\}}^1$  and  $\mathbf{w}_{\{2,3\}}^2$ , which are both known by users 2 and 6, while  $\mathbf{B}'_{1,\{2,3,5,6\}}$  contains  $\frac{F}{24}$  linear combinations of the symbols in  $\mathbf{w}_{\{2,3\}}^2$  which are known by users in  $\{2, 3, 5, 6\}$ . It will be clarified later that the server transmits  $\mathbf{B}'_{1,\{2,6\}}$  with coded caching gain equal to  $g = 2$  (i.e., the multicast message satisfies two users simultaneously), and  $\mathbf{B}'_{1,\{2,3,5,6\}}$  with coded caching gain equal to  $g + 1 = 3$ .

Following the same line or reasoning, we can express the demands of the other users as

$$\mathbf{B}'_2 = [\mathbf{B}'_{2,\{1,3\}}; \mathbf{B}'_{2,\{1,3,4,6\}}]; \quad (50)$$

$$\mathbf{B}'_3 = [\mathbf{B}'_{3,\{2,4\}}; \mathbf{B}'_{3,\{1,2,4,5\}}]; \quad (51)$$

$$\mathbf{B}'_4 = [\mathbf{B}'_{4,\{3,5\}}; \mathbf{B}'_{4,\{2,3,5,6\}}]; \quad (52)$$

$$\mathbf{B}'_5 = [\mathbf{B}'_{5,\{4,6\}}; \mathbf{B}'_{5,\{1,3,4,6\}}]; \quad (53)$$

$$\mathbf{B}'_6 = [\mathbf{B}'_{6,\{1,5\}}; \mathbf{B}'_{6,\{1,2,4,5\}}]. \quad (54)$$

The transmission contains two steps.

- In the first step, we let each user  $k \in [6]$  recover the first term of its demand  $\mathbf{B}'_k$ . In this step, the server transmits

$$\mathbf{B}'_{1,\{2,6\}} + \mathbf{B}'_{2,\{1,3\}}, \quad (55)$$

$$\mathbf{B}'_{3,\{2,4\}} + \mathbf{B}'_{4,\{3,5\}}, \quad (56)$$

$$\mathbf{B}'_{5,\{4,6\}} + \mathbf{B}'_{6,\{1,5\}}, \quad (57)$$

which contains  $\frac{F}{8}$  symbols.

- In the second step, we let each user  $k \in [6]$  recover the second term of its demand  $\mathbf{B}'_k$ . In this step, the server transmits

$$\mathbf{B}'_{1,\{2,3,5,6\}} + \mathbf{B}'_{2,\{1,3,4,6\}} + \mathbf{B}'_{3,\{1,2,4,5\}}, \quad (58)$$

$$\mathbf{B}'_{4,\{2,3,5,6\}} + \mathbf{B}'_{5,\{1,3,4,6\}} + \mathbf{B}'_{6,\{1,2,4,5\}}, \quad (59)$$

for a total of  $\frac{F}{12}$  symbols. From the received multicast messages and its cache content, each user  $k \in [K]$  can recover  $\mathbf{B}'_k$ , and then compute  $\mathbf{B}_k$  from  $\mathbf{T}_k^{-1} \mathbf{B}'_k$ .

The normalized load is  $\rho_2 = \frac{1}{8} + \frac{1}{12} = \frac{5}{24}$ , which coincides with (19).

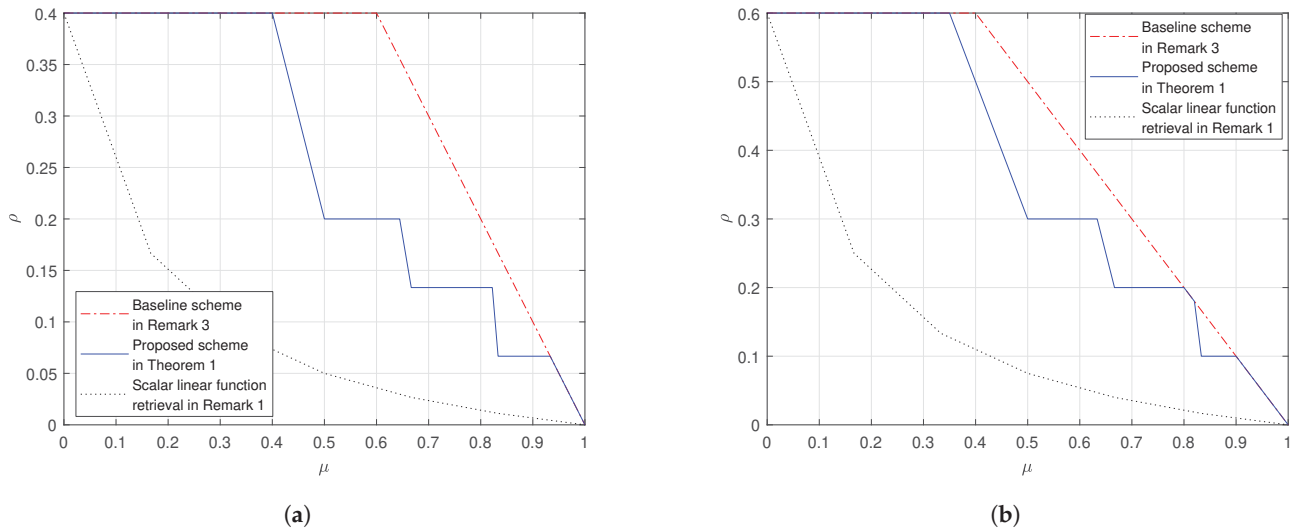
In conclusion, the normalized load of the proposed scheme is  $\rho_{\text{ach}} = \min\{\rho_1, \rho_2\} = \frac{5}{24}$ , while the baseline scheme in (16) achieves the normalized load equals  $\frac{25}{72}$ .

#### 4. Numerical Evaluations

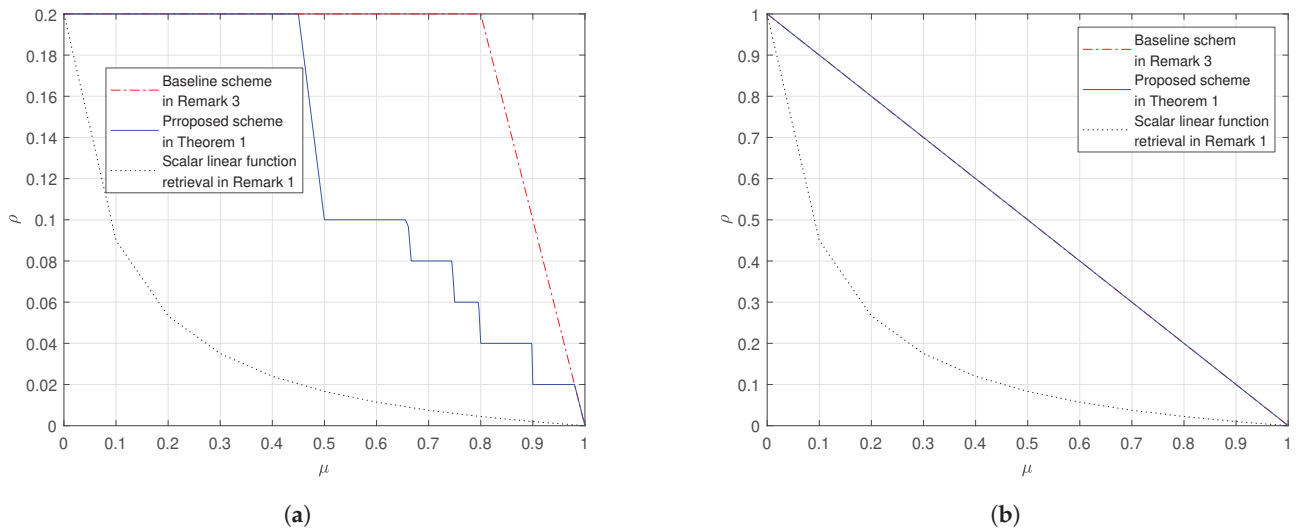
We provide here some numerical evaluations on the performance of the proposed scheme in (17). In Figure 1a we consider the case  $(K, \lambda) = (6, 1/15)$  and in Figure 1b the case  $(K, \lambda) = (6, 1/10)$ . In Figure 2a we consider the case  $(K, \lambda) = (10, 1/50)$  and in Figure 2b the case  $(K, \lambda) = (10, 1/10)$ . From the figures, we observe that:

- In both settings our proposed scheme in Theorem 1 outperforms the baseline scheme, as proved in Remark 5.
- Fix  $K$  and  $\mu$ . When  $\lambda$  grows, the gap between the proposed scheme and the baseline scheme reduces. When  $\lambda = \frac{1}{K}$ , the proposed scheme and the baseline scheme have the same load; this is because at the corner points of the proposed scheme in (26) we achieve the load  $1 - \mu$  which is the same as the baseline scheme.
- In addition, we also plot the cache-aided scalar linear function retrieval scheme in (14), which only works for the case where the demand matrices are with the form in (12). This comparison shows that, if the demand matrices are structured, we can design

caching schemes that leverage the special structure of the demands to achieve a load that is no larger than the load for the worst-case demands. Moreover, the more the structure the more the gain compared to in (17).



**Figure 1.** The shared-link cache-aided general linear function retrieval problem  $K = 6$ . (a)  $\lambda = \frac{1}{15}$ , (b)  $\lambda = \frac{1}{10}$ .



**Figure 2.** The shared-link cache-aided general linear function retrieval problem  $K = 10$ . (a)  $\lambda = \frac{1}{50}$ , (b)  $\lambda = \frac{1}{10}$ .

## 5. Conclusions

In this paper, we formulated the cache-aided general linear function retrieval problem, where each user requests some linear combinations of all the symbols in the library. The formulated problem generalizes the cache-aided scalar linear function retrieval problem. We proposed a novel scheme that strictly improves on an uncoded caching baseline scheme. Further directions include designing improved coded caching schemes for arbitrary users' demand ranges (the setting considered here), as well as for given specific users' demand ranges. In addition, the derivation of a converse bound is also part of on-going work.

**Author Contributions:** Conceptualization, K.W., H.S., M.J., D.T., and G.C.; methodology, K.W., H.S., M.J., D.T., and G.C.; writing—original draft preparation, K.W.; writing—review and editing, K.W., H.S., M.J., D.T., and G.C.; funding acquisition, H.S., M.J., D.T., and G.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work of K. Wan and G. Caire was partially funded by the European Research Council under the ERC Advanced Grant N. 789190, CARENET. The work of H. Sun was supported in part by NSF Award 2007108. The work of M. Ji was supported in part by NSF Awards 1817154 and 1824558. The work of D. Tuninetti was supported in part by NSF Award 1910309.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proof of Theorem 1

By a grouping strategy, we can achieve the normalized memory-load points in (26). In the following, inspired by Example 1, we introduce a general interpolation method among the points in (26).

We let  $\text{Mod}(b, a)$  represent the modulo operation on  $b$  with integer divisor  $a$  and we let  $\text{Mod}(b, a) \in \{1, \dots, a\}$  (i.e., we let  $\text{Mod}(b, a) = a$  if  $a$  divides  $b$ ).

*Appendix A.1.  $g \in [K - 1]$  and  $\frac{\alpha}{g} \geq \left\lceil \frac{K}{g} \right\rceil \lambda$*

We first consider the case where  $g \in [K - 1]$  and  $\frac{\alpha}{g} \geq \left\lceil \frac{K}{g} \right\rceil \lambda$ . Recall that  $\mu = \alpha \frac{g-1}{g} + (1 - \alpha) \frac{g}{g+1} > \frac{g-1}{g}$ . In this case, we directly use the caching scheme in (26) for the memory size  $\frac{g-1}{g}$  with achieved normalized load

$$\min \left\{ \left\lceil \frac{K}{g} \right\rceil \lambda, \frac{1}{g} \right\} = \left\lceil \frac{K}{g} \right\rceil \lambda, \quad (\text{A1})$$

which coincides with (17).

*Appendix A.2.  $g \in [K - 1]$  and  $\frac{\alpha}{g} \leq \left\lceil \frac{K}{g} \right\rceil \lambda$*

We then focus on the case where  $g \in [K - 1]$  and  $\frac{\alpha}{g} \leq \left\lceil \frac{K}{g} \right\rceil \lambda$ .

*Placement Phase.* The placement is done by the memory-sharing between the proposed placements in (26) for  $M_1 = \frac{g-1}{g}$  and  $M_2 = \frac{g}{g+1}$ . We divide  $\mathbf{w}$  into two parts,  $\mathbf{w} = [\mathbf{w}^1; \mathbf{w}^2]$  where the dimension of  $\mathbf{w}^1$  is  $\alpha F \times 1$  and the dimension of  $\mathbf{w}^2$  is  $(1 - \alpha)F \times 1$ .

For the first part, we further partition  $\mathbf{w}^1$  into  $g$  non-overlapping and equal-length subfiles,  $\mathbf{w}^1 = [\mathbf{w}_{\mathcal{T}}^1 : \mathcal{T} \subseteq [g], |\mathcal{T}| = g - 1]$ , where the dimension of each subfile  $\mathbf{w}_{\mathcal{T}}^1$  is  $\frac{\alpha F}{g} \times 1$ . Each user  $k \in [K]$  caches  $\mathbf{w}_{\mathcal{T}}^1$  where  $\mathcal{T} \subseteq [g], |\mathcal{T}| = g - 1$ , and  $\text{Mod}(k, g) \in \mathcal{T}$ .

For the second part of each file, we further partition  $\mathbf{w}^2$  into  $g + 1$  non-overlapping and equal-length subfiles,  $\mathbf{w}^2 = [\mathbf{w}_{\mathcal{T}}^2 : \mathcal{T} \subseteq [g + 1], |\mathcal{T}| = g]$ , where the dimension of each subfile  $\mathbf{w}_{\mathcal{T}}^2$  is  $\frac{(1 - \alpha)F}{g + 1} \times 1$ . Each user  $k \in [K]$  caches  $\mathbf{w}_{\mathcal{T}}^2$  where  $\mathcal{T} \subseteq [g + 1], |\mathcal{T}| = g$ , and  $\text{Mod}(k, g + 1) \in \mathcal{T}$ .

In total, each user caches

$$(g - 1) \frac{\alpha F}{g} + \frac{g(1 - \alpha)F}{g + 1} = \mu F \quad (\text{A2})$$

symbols, satisfying the memory size constraint.

*Delivery Phase.* For each  $\mathcal{T}_1 \in [g]$  where  $|\mathcal{T}_1| = g - 1$ , we define  $\mathbf{D}_{k, \mathcal{T}_1}$  as the sub-matrix of  $\mathbf{D}_k$  which contains the columns corresponding to the symbols in  $\mathbf{w}_{\mathcal{T}_1}^1$ . In addition, for each  $\mathcal{T}_2 \in [g + 1]$  where  $|\mathcal{T}_2| = g$ , we define  $\mathbf{D}_{k, \mathcal{T}_2}$  as the sub-matrix of  $\mathbf{D}_k$  which contains the columns corresponding to the symbols in  $\mathbf{w}_{\mathcal{T}_2}^2$ .

We can express the demand of user  $k \in [K]$  as

$$\mathbf{D}_k \mathbf{w} = \sum_{\mathcal{T}_1 \in [g]: |\mathcal{T}_1|=g-1} \mathbf{D}_{k, \mathcal{T}_1} \mathbf{w}_{\mathcal{T}_1}^1 + \sum_{\mathcal{T}_2 \in [g+1]: |\mathcal{T}_2|=g} \mathbf{D}_{k, \mathcal{T}_2} \mathbf{w}_{\mathcal{T}_2}^2 \quad (\text{A3})$$

$$= \sum_{\substack{\mathcal{T}_1 \in [g]: |\mathcal{T}_1|=g-1, \\ \text{Mod}(k, g) \in \mathcal{T}_1}} \mathbf{D}_{k, \mathcal{T}_1} \mathbf{w}_{\mathcal{T}_1}^1 + \sum_{\substack{\mathcal{T}_2 \in [g+1]: |\mathcal{T}_2|=g, \\ \text{Mod}(k, g+1) \in \mathcal{T}_2}} \mathbf{D}_{k, \mathcal{T}_2} \mathbf{w}_{\mathcal{T}_2}^2 \\ + \mathbf{D}_{k, [g] \setminus \{\text{Mod}(k, g)\}} \mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1 + \mathbf{D}_{k, [g+1] \setminus \{\text{Mod}(k, g+1)\}} \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k, g+1)\}}^2. \quad (\text{A4})$$

It can be seen that user  $k$  knows all the terms in (A4) except

$$\mathbf{B}_k := \mathbf{D}_{k, [g] \setminus \{\text{Mod}(k, g)\}} \mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1 + \mathbf{D}_{k, [g+1] \setminus \{\text{Mod}(k, g+1)\}} \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k, g+1)\}}^2. \quad (\text{A5})$$

Hence, in the delivery phase user  $k$  should recover  $\mathbf{B}_k$ . We then propose two solutions for this objective.

The solution that achieves  $\rho_1$  in (18). We let user  $k$  recover

$$\mathbf{B}_{k,1} := \mathbf{D}_{k, [g] \setminus \{\text{Mod}(k, g)\}} \mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1, \quad (\text{A6})$$

$$\mathbf{B}_{k,2} := \mathbf{D}_{k, [g+1] \setminus \{\text{Mod}(k, g+1)\}} \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k, g+1)\}}^2. \quad (\text{A7})$$

For the first term  $\mathbf{B}_{k,1}$ , the dimension of  $\mathbf{D}_{k, [g] \setminus \{\text{Mod}(k, g)\}}$  is  $\lambda F \times \frac{\alpha F}{g}$  and  $\mathbf{D}_{k, [g] \setminus \{\text{Mod}(k, g)\}}$  is known by each user. Recall that in this case we have  $\frac{\alpha F}{g} \leq \lambda F$ . Hence, we let user  $k$  directly recover  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1$ . Thus in the delivery phase, we let the server transmit

$$\sum_{i \in [g]} \mathbf{w}_{[g] \setminus \{i\}}^1, \quad (\text{A8})$$

with  $\frac{\alpha F}{g}$  symbols. It can be seen that each user  $k \in [K]$  desires  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1$  and caches all the other terms in (A8), such that user  $k$  can recover  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1$ .

For the second term  $\mathbf{B}_{k,2}$ , the dimension of  $\mathbf{D}_{k, [g+1] \setminus \{\text{Mod}(k, g+1)\}}$  is  $\lambda F \times \frac{(1-\alpha)F}{g+1}$ . Notice that  $\mathbf{B}_{k,2}$  only contains linear combinations of the second parts of files in the library. For the second part of each file, the users in

$$\mathcal{G}_i^2 = \{k \in [K] : \text{Mod}(k, g+1) = i\}$$

cache the same content, where  $i \in [g+1]$ . Thus we can use the proposed delivery scheme in (26). More precisely, for each  $i \in [g+1]$ , we generate a virtual user  $v_i$  with the demand

$$\mathbf{D}'_i \mathbf{w}_{[g+1] \setminus \{i\}}^2 = \begin{bmatrix} \mathbf{D}_{\mathcal{G}_i^2(1), [g+1] \setminus \{i\}} \\ \mathbf{D}_{\mathcal{G}_i^2(2), [g+1] \setminus \{i\}} \\ \vdots \\ \mathbf{D}_{\mathcal{G}_i^2(|\mathcal{G}_i^2|), [g+1] \setminus \{i\}} \end{bmatrix} \mathbf{w}_{[g+1] \setminus \{i\}}^2. \quad (\text{A9})$$

Notice that the dimension of  $\mathbf{D}'_i$  is  $|\mathcal{G}_i^2| \lambda F \times \frac{(1-\alpha)F}{g+1}$ . So virtual user  $v_i$  only needs to recover at most  $\min \left\{ \left\lceil \frac{K}{g+1} \right\rceil \lambda, \frac{1-\alpha}{g+1} \right\} F$  symbols in (A9). We denote the set of these symbols by  $\mathbf{P}'_{i, [g+1] \setminus \{i\}}$ , which is known by all the other virtual users. We then let the server transmit

$$\sum_{i \in [g+1]} \mathbf{P}'_{i, [g+1] \setminus \{i\}}, \quad (\text{A10})$$

with  $\min \left\{ \left\lceil \frac{K}{g+1} \right\rceil \lambda, \frac{1-\alpha}{g+1} \right\} F$  symbols, such that each virtual user can recover its demand.

In total, the server transmits

$$\frac{\alpha F}{g} + \min \left\{ \left\lceil \frac{K}{g+1} \right\rceil \lambda, \frac{1-\alpha}{g+1} \right\} F = \rho_1 F \quad (\text{A11})$$

symbols, which coincides with (18).

The solution that achieves  $\rho_2$  in (19). Recall that the demanded sum of user  $k$  is

$$\mathbf{B}_k = \mathbf{D}_{k,[g] \setminus \{\text{Mod}(k,g)\}} \mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1 + \mathbf{D}_{k,[g+1] \setminus \{\text{Mod}(k,g+1)\}} \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2 \quad (\text{A12})$$

$$= \begin{bmatrix} \mathbf{D}_{k,[g] \setminus \{\text{Mod}(k,g)\}} & \mathbf{D}_{k,[g+1] \setminus \{\text{Mod}(k,g+1)\}} \end{bmatrix} \begin{bmatrix} \mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1 \\ \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2 \end{bmatrix}. \quad (\text{A13})$$

We can take a linear transformations on  $\mathbf{B}_k$  as follows

$$\mathbf{B}'_k = \mathbf{T}_k \begin{bmatrix} \mathbf{D}_{k,[g] \setminus \{\text{Mod}(k,g)\}} & \mathbf{D}_{k,[g+1] \setminus \{\text{Mod}(k,g+1)\}} \end{bmatrix} \begin{bmatrix} \mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1 \\ \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2 \end{bmatrix}, \quad (\text{A14})$$

where  $\mathbf{T}_k$  is full-rank with dimension  $\lambda F \times \lambda F$ , and the bottom  $\left[\lambda - \frac{\alpha}{g}\right]^+ F$  symbols in  $\mathbf{B}'_k$  are some linear combinations of  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2$  (i.e., these linear combinations do not contain any term in  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1$ ). This is possible because  $\mathbf{B}_k$  contains  $\lambda F$  linear combinations of all symbols in  $[\mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1; \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2]$ , while  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1$  contains  $\frac{\alpha F}{g}$  symbols. Hence, we can re-express  $\mathbf{B}'_k$  as

$$\mathbf{B}'_k = \begin{bmatrix} \left( \mathbf{B}'_{k,1} \right)_{\min \left\{ \frac{\alpha F}{g}, \lambda F \right\} \times 1} \\ \left( \mathbf{B}'_{k,2} \right)_{\left[ \lambda - \frac{\alpha}{g} \right]^+ F \times 1} \end{bmatrix}. \quad (\text{A15})$$

The delivery phase is divided into two steps. In the first step, we first let each user  $k \in [K]$  recover  $\mathbf{B}'_{k,1}$ . Notice that  $\mathbf{B}'_{k,1}$  is the set of some linear combinations of the symbols in  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1 \mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2 \cdot \mathbf{w}_{[g] \setminus \{\text{Mod}(k,g)\}}^1$  is known by any user  $j_1 \in [K]$  where  $\text{Mod}(j_1, g) \neq k$ ;  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2$  is known by any user  $j_2 \in [K]$  where  $\text{Mod}(j_2, g+1) \neq k$ . Assume that  $k = a_k g + \text{Mod}(k, g)$ , where  $a_k = \left\lceil \frac{k}{g} \right\rceil - 1$  and  $\text{Mod}(k, g) \in [g]$ . In Appendix B, we prove the following lemma.

**Lemma A1.** Each user  $k_1 = a_k g + j$  where  $j \in [g] \setminus \{\text{Mod}(k, g)\}$  and  $k_1 \in [K]$ , caches both  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1$  and  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k, g+1)\}}^2$ .

For each  $i \in \left\lceil \frac{K}{g} \right\rceil$ , we let the server transmit

$$\sum_{j \in [g]: (i-1)g+j \leq K} \mathbf{B}'_{(i-1)g+j,1}. \quad (\text{A16})$$

From Lemma A1, each user  $(i-1)g+j$  knows all except  $\mathbf{B}'_{(i-1)g+j,1}$  such that it can recover  $\mathbf{B}'_{(i-1)g+j,1}$ . In this step, the server transmits  $\left\lceil \frac{K}{g} \right\rceil \min \left\{ \frac{\alpha}{g}, \lambda \right\} F$  symbols.

In the second step, we then let each user  $k \in [K]$  recover  $\mathbf{B}'_{k,2}$ , which contains linear combinations of  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2$ . We can use the same delivery scheme as we used to delivery the second term in the first solution (i.e.,  $\mathbf{B}_{k,2}$  in (A7) which contains  $\lambda F$  linear combinations of  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k,g+1)\}}^2$ ). Here we do not repeat the scheme. Notice that  $\mathbf{B}'_{k,2}$

contains  $\left[\lambda - \frac{\alpha}{g}\right]^+ F$  linear combinations of  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k, g+1)\}}^2$ . Hence, in this step the server transmits  $\min\left\{\left\lceil \frac{K}{g+1} \right\rceil \left[\lambda - \frac{\alpha}{g}\right]^+, \frac{1-\alpha}{g+1}\right\} F$  symbols.

After recovering  $\mathbf{B}'_k$ , each user  $k \in [K]$  reconstructs

$$\mathbf{B}_k = \mathbf{T}_k^{-1} \mathbf{B}'_k,$$

and then recovers its demand.

In total, the achieved normalized load is

$$\left\lceil \frac{K}{g} \right\rceil \min\left\{\frac{\alpha}{g}, \lambda\right\} + \min\left\{\left\lceil \frac{K}{g+1} \right\rceil \left[\lambda - \frac{\alpha}{g}\right]^+, \frac{1-\alpha}{g+1}\right\} = \rho_2,$$

coinciding with (19).

### Appendix A.3. Proof of (20)

Finally, we focus on the case  $\mu = \alpha \frac{K-1}{K} + (1-\alpha)$  where  $\alpha \in (0, 1)$ . In this case, the proposed scheme is a direct extension from the proposed scheme in (25). More precisely,

- we directly use the caching scheme in (25) for the memory size  $\frac{K-1}{K}$  with the achieved normalized load equal to  $\lambda$ .
- In this case, the number of symbols which are not cached by user is  $\frac{\alpha F}{K}$ . Hence, we can let each user directly recover the uncached symbols with the achieved normalized load equal to  $\frac{\alpha}{K}$ .

This concludes the proof.

### Appendix B. Proof of Lemma A1

Recall that  $k = a_k g + \text{Mod}(k, g)$ , where  $a_k = \left\lceil \frac{k}{g} \right\rceil - 1$  and  $\text{Mod}(k, g) \in [g]$ . We focus on one user  $k_1 = a_k g + j$  where  $j \in [g] \setminus \{\text{Mod}(k, g)\}$ . Since  $j = \text{Mod}(k_1, g) \neq \text{Mod}(k, g)$ , it can be easily seen that  $\mathbf{w}_{[g] \setminus \{\text{Mod}(k, g)\}}^1$  is cached by user  $j$ . In the rest of this proof, we show that user  $j$  also caches  $\mathbf{w}_{[g+1] \setminus \{\text{Mod}(k, g+1)\}}^2$ ; or equivalently,  $\text{Mod}(k_1, g+1) \neq \text{Mod}(k, g+1)$ .

We prove it by contradiction. Assume that  $\text{Mod}(k_1, g+1) = \text{Mod}(k, g+1) = j'$ . Hence, we can re-express  $k$  as  $k = a'_k(g+1) + j'$  and re-express  $k_1$  as  $k_1 = a'_{k_1}(g+1) + j'$ , where  $a'_k = \left\lceil \frac{k}{g+1} \right\rceil - 1$  and  $a'_{k_1} = a'_k = \left\lceil \frac{k_1}{g+1} \right\rceil - 1$ .

Since  $k = a_k g + \text{Mod}(k, g) = a'_k(g+1) + j'$ , we have

$$a_k g = a'_k(g+1) + j' - \text{Mod}(k, g). \quad (\text{A17})$$

In addition, we have

$$k_1 = a_k g + j = a'_{k_1}(g+1) + j' \quad (\text{A18})$$

By taking (A17) into (A18), we have

$$a_k g + j = a'_{k_1}(g+1) + j' \quad (\text{A19})$$

$$\stackrel{(\text{A17})}{\iff} a'_k(g+1) + j' - \text{Mod}(k, g) + j = a'_{k_1}(g+1) + j' \quad (\text{A20})$$

$$\iff (a'_k - a'_{k_1})(g+1) = \text{Mod}(k, g) - j. \quad (\text{A21})$$

Since  $\text{Mod}(k, g) \in [g]$  and  $j \in [g]$ , it can be seen that (A21) holds if and only if  $a'_k - a'_{k_1} = 0$ , which leads to  $k = k_1$  and contradicts with  $\text{Mod}(k, g) \neq \text{Mod}(k_1, g)$ . Hence, we proved that  $\text{Mod}(k_1, g+1) \neq \text{Mod}(k, g+1)$  and proved Lemma A1.

## References

1. Borst, S.; Gupta, V.; Walid, A. Distributed caching algorithms for content distribution networks. In Proceedings of the 2010 Proceedings IEEE INFOCOM, San Diego, CA, USA, 14–19 March 2010; pp. 1478–1486.
2. Maddah-Ali, M.A.; Niesen, U. Fundamental Limits of Caching. *IEEE Trans. Inf. Theory* **2014**, *60*, 2856–2867. [\[CrossRef\]](#)
3. Wan, K.; Tuninetti, D.; Piantanida, P. On the Optimality of Uncoded Cache Placement. In Proceedings of the IEEE Information Theory Workshop (ITW), Cambridge, UK, 11–14 September 2016.
4. Wan, K.; Tuninetti, D.; Piantanida, P. An Index Coding Approach to Caching With Uncoded Cache Placement. *IEEE Trans. Inf. Theory* **2020**, *66*, 1318–1332. [\[CrossRef\]](#)
5. Yu, Q.; Maddah-Ali, M.A.; Avestimehr, S. The Exact Rate-Memory Tradeoff for Caching with Uncoded Prefetching. *IEEE Trans. Inf. Theory* **2018**, *64*, 1281–1296. [\[CrossRef\]](#)
6. Yu, Q.; Maddah-Ali, M.A.; Avestimehr, S. Characterizing the Rate-Memory Tradeoff in Cache Networks Within a Factor of 2. *IEEE Trans. Inf. Theory* **2019**, *65*, 647–663. [\[CrossRef\]](#)
7. Wan, K.; Sun, H.; Ji, M.; Tuninetti, D.; Caire, G. On Optimal Load-Memory Tradeoff of Cache-Aided Scalar Linear Function Retrieval. *arXiv* **2020**, arXiv:2001.03577.
8. Wan, K.; Caire, G. On Coded Caching with Private Demands. *IEEE Trans. Inf. Theory* **2020**. [\[CrossRef\]](#)
9. Ji, M.; Caire, G.; Molisch, A.F. Fundamental Limits of Caching in Wireless D2D Networks. *IEEE Trans. Inf. Theory* **2016**, *62*, 849–869. [\[CrossRef\]](#)
10. Yapar, C.; Wan, K.; Schaefer, R.F.; Caire, G. On the Optimality of D2D Coded Caching With Uncoded Cache Placement and One-Shot Delivery. *IEEE Trans. Commun.* **2019**, *67*, 8179–8192. [\[CrossRef\]](#)
11. Yan, Q.; Tuninetti, D. Fundamental Limits of Caching for Demand Privacy against Colluding Users. *arXiv* **2020**, arXiv:2008.03642.
12. Yan, Q.; Tuninetti, D. Key Superposition Simultaneously Achieves Security and Privacy in Cache-Aided Linear Function Retrieval. *arXiv* **2020**, arXiv:2009.06000.
13. Shanmugam, K.; Ji, M.; Tulino, A.M.; Llorca, J.; Dimakis, A.G. Finite-Length Analysis of Caching-Aided Coded Multicasting. *IEEE Trans. Inf. Theory* **2016**, *62*, 5524–5537. [\[CrossRef\]](#)
14. Wan, K.; Sun, H.; Ji, M.; Tuninetti, D.; Caire, G. Cache-Aided Matrix Multiplication Retrieval. *arXiv* **2020**, arXiv:2007.00856.
15. Bar-Yossef, Z.; Birk, Y.; Jayram, T. S.; Kol, T. Index Coding with Side Information. *IEEE Trans. Inf. Theory* **2011**, *57*, 1479–1494. [\[CrossRef\]](#)
16. Lee, N.; Dimakis, A.G.; Heath, R.W. Index Coding with Coded Side-Information. *IEEE Commun. Lett.* **2015**, *19*, 319–322. [\[CrossRef\]](#)
17. Jia, Z.; Jafar, S.A. Cross Subspace Alignment Codes for Coded Distributed Batch Computation. *arXiv* **2019**, arXiv:1909.13873.
18. Chang, W.; Tandon, R. On the Upload versus Download Cost for Secure and Private Matrix Multiplication. *arXiv* **2019**, arXiv:1906.10684.
19. Kakar, J.; Khristoforov, A.; Ebadifar, S.; Sezgin, A. Uplink-downlink tradeoff in secure distributed matrix multiplication. *arXiv* **2019**, arXiv:1910.13849.