

# Uncoded Placement Optimization for Coded Delivery

Sian Jin, *Student Member, IEEE*, Ying Cui, *Member, IEEE*,  
Hui Liu, *Fellow, IEEE*, Giuseppe Caire, *Fellow, IEEE*

## Abstract

We consider the classical coded caching problem as defined by Maddah-Ali and Niesen, where a server with a library of  $N$  files of equal size is connected to  $K$  users via a shared error-free link. Each user is equipped with a cache with capacity of  $M$  files. The goal is to design a static content placement and delivery scheme such that the average load over the shared link is minimized. Existing coded caching schemes fail to simultaneously achieve efficient content placement for non-uniform file popularity and efficient content delivery in the presence of common requests, and hence may not achieve desirable average load under a non-uniform, possibly very skewed, popularity distribution. In addition, existing coded caching schemes usually require the splitting of a file into a large number of subfiles, i.e., high subpacketization level. To address the above two challenges, we first present a class of centralized coded caching schemes consisting of a general content placement strategy specified by a file partition parameter, enabling efficient and flexible content placement, and a specific content delivery strategy, enabling load reduction by exploiting common requests of different users. For the proposed class of schemes, we consider two cases for the optimization of the file partition parameter, depending on whether a large subpacketization level is allowed or not. In the case of an unrestricted subpacketization level, we formulate the coded caching optimization in order to minimize the average load under an arbitrary file popularity. A direct formulation of the problem involves  $N2^K$  variables. By imposing some additional conditions, the problem is reduced to a linear program with  $N(K+1)$  variables under an arbitrary file popularity and with  $K+1$  variables under the uniform file popularity. We can recover Yu *et al.*'s optimal scheme for the uniform file popularity as an optimal solution of our problem. When a low subpacketization level is desired, we introduce a subpacketization level constraint involving the  $\ell_0$  norm for each file. Again, by imposing the same additional conditions, we can simplify the problem to a difference of two convex functions (DC) problem with  $N(K+1)$  variables that can be efficiently solved.

S. Jin, Y. Cui and H. Liu are with Shanghai Jiao Tong University, China. G. Caire is with Technical University of Berlin, Germany. This paper was presented in part at IEEE WiOpt 2018 [1].

## Index Terms

Coded caching, coded multicasting, content distribution, optimization, subpacketization level, the  $\ell_0$ -norm.

## I. INTRODUCTION

The rapid proliferation of smart mobile devices has triggered an unprecedented growth of the global mobile data traffic, with a predicted nearly seven-fold increase between 2016 and 2021 [2]. In order to support such dramatic growth of wireless data traffic, caching and multicasting have been recently proposed as two promising approaches for massive content delivery in wireless networks. Joint design of the two promising techniques is expected to achieve superior performance for massive content delivery in wireless networks. In [3]–[5], the authors consider joint design of traditional uncoded caching and multicasting, the gain of which mainly derives from making content available locally and serving multiple requests of the same contents concurrently.

Recently, a new class of caching schemes for content placement in user caches, referred to as *coded caching* [6], have received significant interest. In [6], Maddah-Ali and Niesen consider a system with one server connected through a shared error-free link to  $K$  users. The server has a database of  $N$  files (of  $F$  data units), and each user has an isolated cache memory containing up to  $M$  files. They formulate a caching problem consisting of two phases, namely, a content placement phase and a content delivery phase. The content placement is performed once, before operating the network, and independently of the user requests. Then, the users place requests in rounds, and at each round the server responds with a multicast message constructed by coded multicast XOR operations that satisfies all user requests simultaneously. The goal of [6] is to reduce the worst-case (over all possible requests) load of the shared link in the delivery phase. In [7], we consider a different class of centralized coded caching schemes specified by a general file partition parameter, and optimize the parameter to minimize the average (over random requests) load within the class under an arbitrary file popularity. In [8], the parameter-based coded caching design approach in [7] is generalized to minimize the average load in a heterogeneous setting with nonuniform cache size and file size under an arbitrary file popularity. In [9], Yu *et al.* propose a centralized coded caching scheme where the delivery strategy exploits the chance of load reduction in common requests of different users and prove its information theoretic optimality for the worst-case load and average load under the uniform file popularity.

Note that the delivery strategies in [6]–[8] do not capture the opportunity of load reduction in common requests of different users, and the placement strategies in [6] and [9] allocate the same fraction of memory to each file without reflecting popularity difference of files. Therefore, the coded caching schemes in [6]–[9] may not achieve desirable average load under a non-uniform, possibly very skewed, popularity distribution. At the moment, a general optimality result for random requests with an arbitrary file popularity is not known.

Another limitation of [6]–[9] is the issue of high subpacketization level, i.e., the number of non-overlapping subfiles for each file is large. In [10]–[12], the authors tackle the subpacketization level issue for centralized coded caching. Specifically, in [10], Tang *et al.* connect coded caching to resolvable combinatorial designs and propose a centralized coded caching scheme where the subpacketization level is exponential with respect to (w.r.t.) the number of users but with a smaller exponent constant than in the centralized coded caching scheme of [6] at the cost of a marginal increase in the worst-case load. In [11], Shanmugam *et al.* connect coded caching to Ruzsa-Szemerédi graphs and show the existence of a centralized coded caching scheme where the subpacketization level grows linearly with the number of users. However, such scheme exists only when the number of users is impractically large. In [12], the authors propose a centralized coded caching scheme with low subpacketization level based on Pareto-optimal placement delivery array (PDA). Note that the centralized coded caching schemes in [10]–[12] addressing the subpacketization level issue are applicable only for certain system parameters (e.g., the number of users, the number of files, cache size, etc.). Furthermore, the centralized coded caching schemes in [10]–[12] are based on combinatorial designs and do not explicitly solve any optimization problem under subpacketization level constraints.

In this paper, we would like to address the above challenges in the same centralized setting as in [6]–[12], with the focus on minimizing the average load under an arbitrary file popularity in two cases, namely, the case without considering the subpacketization level issue and the case considering the subpacketization level issue. We present a class of coded caching schemes consisting of a general content placement strategy specified by a file partition parameter, enabling efficient and flexible content placement, and a specific content delivery strategy, enabling load reduction by exploiting common requests of different users. Then, we focus on the average load minimization irrespectively of the subpacketization level issue. In this case, we formulate the coded caching optimization problem over the considered class of schemes to minimize the

average load under an arbitrary file popularity. The average load expression is not tractable due to the complex delivery strategy. Therefore, we impose some additional conditions on the parameter to simplify the average load expression under an arbitrary file popularity and the uniform file popularity respectively, by connecting the file request event to the “balls into bins” problem. Based on the simplified expressions, we transform the original optimization problem with  $N2^K$  variables into a linear program with  $N(K + 1)$  variables under an arbitrary file popularity and a linear program with  $K + 1$  variables under the uniform file popularity, which are much easier to solve than the original problem. We also show that Yu *et al.*’s centralized coded caching scheme corresponds to an optimal solution of our problem, thus implying that the imposed conditions incur no loss of optimality for the uniform file popularity. Next, we focus on the average load minimization considering the subpacketization level issue. In this case, we first formulate the coded caching optimization problem over the considered class of schemes to minimize the average load under an arbitrary file popularity subject to subpacketization level constraints in terms of the  $\ell_0$ -norm of the file partition parameter. To the best of our knowledge, this is the first work explicitly considering subpacketization level constraints in the optimization of coded caching design. By imposing the same additional conditions as before and using the exact difference of two convex functions (DC) reformulation method in [13], we convert the original problem with  $N2^K$  variables into a simplified DC problem with  $N(K + 1)$  variables. Then, we use a DC algorithm to solve the simplified DC problem. Numerical results reveal that the imposed conditions do not affect the optimality of the original problem under an arbitrary file popularity in both cases. Furthermore, our numerical results demonstrate that the optimized coded caching scheme without considering the subpacketization level constraints outperforms those in [6], [7], [9] in terms of the average load, and the optimized coded caching scheme considering the subpacketization level constraints outperforms those in [10] and [12] in terms of both the average load and application region.

## II. CENTRALIZED CODED CACHING

### A. Problem Setting

As in [6]–[12], we consider a system with one server connected through a shared error-free link to  $K \in \mathbb{N}_{>0}$  users (see Fig. 1), where  $\mathbb{N}_{>0}$  denotes the set of all positive integers. The server has access to a library of  $N \in \mathbb{N}_{>0}$  files, denoted by  $W_1, \dots, W_N$ , each consisting of  $F \in \mathbb{N}_{>0}$

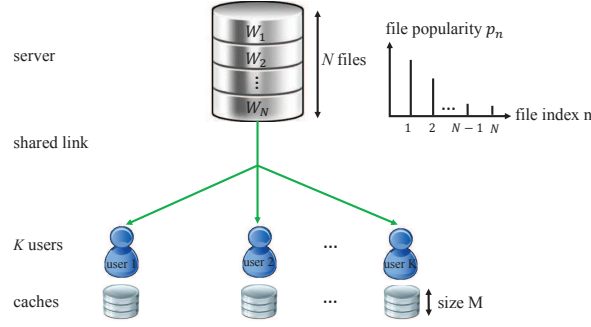


Fig. 1: Problem setup for coded caching [7].

indivisible data units. Let  $\mathcal{N} \triangleq \{1, 2, \dots, N\}$  and  $\mathcal{K} \triangleq \{1, 2, \dots, K\}$  denote the set of file indices and the set of user indices, respectively. Each user has an isolated cache memory of  $MF$  data units, for some real number  $M \in [0, N]$ . Let  $Z_k$  denote the cache content for user  $k$ . The system operates in two phases, i.e., a placement phase and a delivery phase [6]. In the placement phase, each user is able to fill the content of its cache using the library of  $N$  files. In the delivery phase, each user randomly and independently requests one file in  $\mathcal{N}$  according to file popularity distribution  $\mathbf{p} \triangleq (p_n)_{n=1}^N$ , where  $p_n$  denotes the probability of a user requesting file  $W_n$  and  $\sum_{n=1}^N p_n = 1$ . Without loss of generality, we assume  $p_1 \geq p_2 \geq \dots \geq p_N$ . Let  $D_k \in \mathcal{N}$  denote the index of the file requested by user  $k \in \mathcal{K}$ , and let  $\mathbf{D} \triangleq (D_1, \dots, D_K) \in \mathcal{N}^K$  denote the requests of all the  $K$  users. The server replies to these  $K$  requests by sending messages over the shared link, which are observed by all the  $K$  users. Each user should be able to recover its requested file from the messages received over the shared link and its cache content. Our goal is to minimize the average load of the shared link under an arbitrary file popularity.

### B. Centralized Coded Caching Scheme

In this part, we present a class of centralized coded caching schemes utilizing a general uncoded placement strategy and a specific coded delivery strategy, which are specified by a general file partition parameter, as summarized in Alg. 1. This uncoded placement strategy was introduced in our previous work [7] and it is repeated here for completeness. For all  $n \in \mathcal{N}$ , file  $W_n$  is partitioned into  $2^K$  nonoverlapping subfiles  $W_{n,\mathcal{S}}$ ,  $\mathcal{S} \subseteq \mathcal{K}$ , i.e.,  $W_n = \{W_{n,\mathcal{S}} : \mathcal{S} \subseteq \mathcal{K}\}$ .

If the number of data units in a subfile is zero, then there is no need to consider this subfile. Thus,  $2^K$  is the maximum number of non-overlapping subfiles of a file. We say subfile  $W_{n,\mathcal{S}}$  is of type  $s$  if  $|\mathcal{S}| = s$  [7]. User  $k$  stores  $W_{n,\mathcal{S}}$ ,  $n \in \mathcal{N}, k \in \mathcal{S}, \mathcal{S} \subseteq \mathcal{K}$  in its cache, i.e.,  $Z_k = \{W_{n,\mathcal{S}} : n \in \mathcal{N}, k \in \mathcal{S}, \mathcal{S} \subseteq \mathcal{K}\}$ . Let  $x_{n,\mathcal{S}}$  denote the size of subfile  $W_{n,\mathcal{S}}$ , normalized by the file size  $F$ . Denote  $\mathbf{x}_n \triangleq (x_{n,\mathcal{S}})_{\mathcal{S} \subseteq \mathcal{K}}$ . Let  $\mathbf{x} \triangleq (\mathbf{x}_n)_{n \in \mathcal{N}}$  denote the file partition parameter, which will be optimized to minimize the average load in Section III and Section IV. Thus,  $\mathbf{x}$  satisfies

$$0 \leq x_{n,\mathcal{S}} \leq 1, \quad \forall \mathcal{S} \subseteq \mathcal{K}, n \in \mathcal{N}, \quad (1)$$

$$\sum_{s=0}^K \sum_{\mathcal{S} \subseteq \mathcal{K}: |\mathcal{S}|=s} x_{n,\mathcal{S}} = 1, \quad \forall n \in \mathcal{N}, \quad (2)$$

$$\sum_{n=1}^N \sum_{s=1}^K \sum_{\mathcal{S} \subseteq \mathcal{K}: |\mathcal{S}|=s, k \in \mathcal{S}} x_{n,\mathcal{S}} \leq M, \quad \forall k \in \mathcal{K}, \quad (3)$$

where (1) and (2) represent the file partition constraints and (3) represents the cache memory constraint.

The coded delivery strategy is an extension of that in [9]. For all  $\mathbf{D} \in \mathcal{N}^K$ , let  $\underline{\mathcal{D}}(\mathbf{D})$  denote the set of distinct files in  $\mathbf{D}$ . For all  $n \in \underline{\mathcal{D}}(\mathbf{D})$ , the server arbitrarily selects user  $k_n \in \mathcal{K}$  such that  $D_{k_n} = n$ . Let  $\underline{\mathcal{K}}(\mathbf{D}) \triangleq \{k_n : n \in \underline{\mathcal{D}}(\mathbf{D})\}$  denote the set of representative users that request  $|\underline{\mathcal{D}}(\mathbf{D})|$  different files. Each user  $k \in \mathcal{S}$  requests subfile  $W_{D_k, \mathcal{S} \setminus \{k\}}$ , for all subset  $\mathcal{S} \subseteq \mathcal{K}$ . The server broadcasts coded-multicast message  $\oplus_{k \in \mathcal{S}} W_{D_k, \mathcal{S} \setminus \{k\}}$ <sup>1</sup> for all subset  $\mathcal{S} \subseteq \mathcal{K}$  that satisfies  $\mathcal{S} \cap \underline{\mathcal{K}}(\mathbf{D}) \neq \emptyset$ , and all subfiles in the coded-multicast message are zero-padded to the length of the longest subfile. By Lemma 1 of [9], we can conclude that each user can decode the requested file based on the received coded-multicast messages and the contents stored in its cache.

*Remark 1 (Comparison with Existing Coded Caching Schemes):* The uncoded placement strategy in this paper is more general than that in [6], [9]–[12], and it can be optimized to minimize the average load under an arbitrary file popularity (see Section III and Section IV) [7]. The coded delivery strategy in this paper is more efficient than that in [6]–[8], [10]–[12], since it avoids transmitting the redundant coded-multicast messages  $\oplus_{k \in \mathcal{S}} W_{D_k, \mathcal{S} \setminus \{k\}}$ ,  $\mathcal{S} \subseteq \mathcal{K} \setminus \underline{\mathcal{K}}(\mathbf{D})$  in the presence of common requests [9].

<sup>1</sup>Note that in [9], since all files are partitioned into subfiles of type  $t \in \{0, 1, \dots, K\}$ , only coded-multicast messages  $\oplus_{k \in \mathcal{S}} W_{D_k, \mathcal{S} \setminus \{k\}}$  satisfying  $\mathcal{S} \subseteq \mathcal{K}$ ,  $\mathcal{S} \cap \underline{\mathcal{K}}(\mathbf{D}) \neq \emptyset$  and  $|\mathcal{S}| = t + 1$  are transmitted.

---

**Algorithm 1** Parameter-based Centralized Coded Caching
 

---

**placement strategy**

- 1: **for all**  $k \in \mathcal{K}$  **do**
- 2:    $Z_k \leftarrow \{W_{n,S} : n \in \mathcal{N}, k \in \mathcal{S}, \mathcal{S} \subseteq \mathcal{K}\}$
- 3: **end for**

**delivery strategy**

- 1: **for**  $s = K, K-1, \dots, 1$  **do**
  - 2:   **for**  $\mathcal{S} \subseteq \mathcal{K} : |\mathcal{S}| = s, \mathcal{S} \cap \underline{\mathcal{K}}(\mathbf{D}) \neq \emptyset$  **do**
  - 3:     server sends  $\oplus_{k \in \mathcal{S}} W_{D_k, \mathcal{S} \setminus \{k\}}$
  - 4:   **end for**
  - 5: **end for**
- 

*C. Average Load*

Let  $R_{\text{avg}}(K, N, M, \mathbf{x})$  denote the average load for serving the  $K$  users with cache size  $M$  under a given file partition parameter  $\mathbf{x}$ , where the average is taken over random requests  $\mathbf{D}$  for  $N$  files, according to an arbitrary file popularity distribution  $\mathbf{p}$ . By Alg. 1, we have

$$R_{\text{avg}}(K, N, M, \mathbf{x}) = \sum_{\mathbf{d} \in \mathcal{N}^K} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{\mathcal{S} \subseteq \mathcal{K} : \mathcal{S} \cap \underline{\mathcal{K}}(\mathbf{D}) \neq \emptyset} \max_{k \in \mathcal{S}} x_{d_k, \mathcal{S} \setminus \{k\}}, \quad (4)$$

where  $\mathbf{d} \triangleq (d_1, \dots, d_K) \in \mathcal{N}^K$  and  $\max_{k \in \mathcal{S}} x_{d_k, \mathcal{S} \setminus \{k\}}$  is the length of the coded message  $\oplus_{k \in \mathcal{S}} W_{d_k, \mathcal{S} \setminus \{k\}}$ , normalized by the file size  $F$ .

From (4), we can observe that the file partition parameter  $\mathbf{x}$  fundamentally affects the average load  $R_{\text{avg}}(K, N, M, \mathbf{x})$ . In Section III, we would like to find an optimal file partition parameter to minimize the average load in (4). The optimal file partition parameter may correspond to high subpacketization level. In fact, if for some  $n$ , all the elements  $x_{n,S}$  are non-zero, it means that file  $W_n$  is divided into  $2^K$  subfiles, which is exponential with the number of users  $K$ . For systems with even a moderate number of users and files of practical length, such partition becomes quickly impossible. For example, for  $K = 50$  and the size of each indivisible data unit equal to 1 Byte, we need files with size larger than 1 Petabyte. To avoid high subpacketization level, in Section IV, we would like to find an optimal file partition parameter to minimize the average load in (4) under subpacketization level constraints.

### III. AVERAGE LOAD MINIMIZATION WITHOUT SUBPACKETIZATION LEVEL CONSTRAINT

In this section, we consider the minimization of the average load without any restriction on the subpacketization level.

#### A. Problem Formulation

We would like to minimize the average load under the file partition constraints in (1) and (2) as well as the cache memory constraint in (3).

*Problem 1 (Optimization for Arbitrary File Popularity):*

$$\begin{aligned} R_{\text{avg}}^*(K, N, M) &\triangleq \min_{\mathbf{x}} R_{\text{avg}}(K, N, M, \mathbf{x}) \\ \text{s.t.} \quad &(1), (2), (3), \end{aligned}$$

where  $R_{\text{avg}}(K, N, M, \mathbf{x})$  is given by (4).

The objective function of Problem 1 is convex, as it is a positive weighted sum of convex piecewise linear functions. In addition, the constraints of Problem 1 are linear. Hence, Problem 1 is a convex optimization problem. The number of variables in Problem 1 is  $N2^K$ . Thus, the complexity of Problem 1 is huge, especially when  $K$  and  $N$  are large. In Section III-B and Section III-C, we shall focus on deriving simplified formulations for Problem 1 to facilitate low-complexity optimal solutions under an arbitrary file popularity distribution and the uniform file popularity distribution, respectively.

#### B. Optimization for Arbitrary File Popularity

First, we present two structural conditions on the file partition parameter  $\mathbf{x}$ . These conditions impose a restriction on the feasible region and enable a simplification of Problem 1. We cannot prove that the resulting solution is optimal w.r.t. Problem 1 due to its complex objective function, but we can numerically verify the optimality of the resulting solution.

*Condition 1 (Symmetry w.r.t. Type):* For all  $n \in \mathcal{N}$  and  $s \in \{0, 1, \dots, K\}$ , the values of  $x_{n,\mathcal{S}}$ ,  $\mathcal{S} \subseteq \{\widehat{\mathcal{S}} \subseteq \mathcal{K} : |\widehat{\mathcal{S}}| = s\}$  are the same.

Recall that all subfiles in one coded-multicast message are zero-padded to the length of the longest subfile in the coded-multicast message, causing the “bit waste” effect [7]. Thus, imposing



Condition 1 can reduce the variance of the lengths of messages involved in the coded-multicast XOR operations, hence addressing “bit waste” problem. By Condition 1, we can set

$$x_{n,\mathcal{S}} = y_{n,s}, \quad \forall \mathcal{S} \subseteq \mathcal{K}, n \in \mathcal{N}, \quad (5)$$

where  $s = |\mathcal{S}| \in \{0, 1, \dots, K\}$ . Here,  $y_{n,s}$  can be viewed as the size of each subfile of type  $s$  in each file  $W_n$ , normalized by the file size  $F$ . Let  $\mathbf{y}_n \triangleq (y_{n,s})_{s \in \{0,1,\dots,K\}}$  and  $\mathbf{y} \triangleq (\mathbf{y}_n)_{n \in \mathcal{N}}$ .

*Condition 2 (Monotonicity w.r.t. Popularity):* For all  $n \in \{1, 2, \dots, N-1\}$  and  $s \in \{1, 2, \dots, K\}$ , when  $p_n \geq p_{n+1}$ ,

$$y_{n,s} \geq y_{n+1,s}. \quad (6)$$

Condition 2 indicates that, for all  $n \in \{1, 2, \dots, N-1\}$  and  $s \in \{1, 2, \dots, K\}$ , when  $p_n \geq p_{n+1}$ , the size of subfiles  $W_{n,\mathcal{S}}$ ,  $\mathcal{S} \subseteq \{\widehat{\mathcal{S}} \subseteq \mathcal{K} : |\widehat{\mathcal{S}}| = s\}$  is no smaller than that of subfiles  $W_{n+1,\mathcal{S}}$ ,  $\mathcal{S} \subseteq \{\widehat{\mathcal{S}} \subseteq \mathcal{K} : |\widehat{\mathcal{S}}| = s\}$ . Intuitively, imposing Condition 2 can reduce the average load, by dedicating more memory to a more popular file. Unlike in our previous work [7], due to the complex objective function in (4), we cannot show that imposing Conditions 1 and 2 maintains the optimality of the solution w.r.t. the original Problem 1. Later, in Section V, we provide numerical evidence suggesting that indeed Conditions 1 and 2 do not involve any loss of optimality.

Next, we simplify Problem 1 under Conditions 1 and 2. First, we introduce some notations. Consider the number of representative users  $|\underline{\mathcal{K}}(\mathbf{D})| = u$ . Let  $\tilde{D}_{u,\langle 1 \rangle} \leq \tilde{D}_{u,\langle 2 \rangle} \leq \dots \leq \tilde{D}_{u,\langle K-u \rangle}$  denote  $(D_k)_{k \in \underline{\mathcal{K}}(\mathbf{D})}$  arranged in ascending order, so that  $\tilde{D}_{u,\langle i \rangle}$  is the  $i$ -th smallest. Let  $P'_{i,u,n} \triangleq \Pr \left[ \tilde{D}_{u,\langle i \rangle} = n \right]$ , for all  $i = 1, \dots, K-u$ . Note that the file request event  $\tilde{D}_{u,\langle i \rangle} = n$  can be treated as the “balls into bins” problem, i.e.,  $K$  balls are placed in an i.i.d. manner into  $N$  bins, where bin  $n$  is selected with probability  $p_n$ . Let  $K_n$  denote the number of users requesting file  $n$ . Note that  $\sum_{n=1}^N K_n = K$  and  $\sum_{n=1}^N \mathbf{1}[K_n > 0] = u$ . Let  $A_n \triangleq \sum_{n'=1}^{n-1} \mathbf{1}[K_{n'} > 0]$ ,  $B_{n,1} \triangleq \sum_{n'=1}^{n-1} K_{n'}$  and  $B_{n,2} \triangleq \sum_{n'=n+1}^N K_{n'}$ . Let  $\mathcal{L}_{A_n,1} \triangleq \{\mathcal{L} \subseteq \{1, 2, \dots, n-1\} : |\mathcal{L}| = A_n\}$  and  $\mathcal{L}_{A_n,2} \triangleq \{\mathcal{L} \subseteq \{n+1, n+2, \dots, N\} : |\mathcal{L}| = u - A_n - 1\}$ . Let

$$\mathcal{A}_{B_{n,1},\mathcal{L}} \triangleq \left\{ (\alpha_{n'})_{n' \in \mathcal{L}} \in \mathbb{N}_{>0}^{|\mathcal{L}|} : \sum_{n' \in \mathcal{L}} \alpha_{n'} = B_{n,1} \right\},$$

for all  $\mathcal{L} \in \mathcal{L}_{A_n,1}$  and

$$\mathcal{A}_{B_{n,2},\mathcal{L}} \triangleq \left\{ (\alpha_{n'})_{n' \in \mathcal{L}} \in \mathbb{N}_{>0}^{|\mathcal{L}|} : \sum_{n' \in \mathcal{L}} \alpha_{n'} = B_{n,2} \right\},$$

for all  $\mathcal{L} \in \mathcal{L}_{A_n,2}$ . Let  $\mathbb{N}$  denote the set of all natural numbers. Define  $\binom{K}{m_1, m_2, K-m_1-m_2} \triangleq \frac{K!}{m_1! m_2! (K-m_1-m_2)!}$ , where  $m_1 \in \mathbb{N}$ ,  $m_2 \in \mathbb{N}$  and  $m_1 + m_2 \leq K$ . By using results for the “balls into bins” problem and considering Conditions 1 and 2, we have the following result.

*Lemma 1 (Simplification of Problem 1 for Arbitrary File Popularity):* Under Conditions 1 and 2, Problem 1 can be converted into:

*Problem 2 (Simplified Problem for Arbitrary File Popularity):*

$$\tilde{R}_{\text{avg}}^*(K, N, M, \mathbf{y}) \triangleq \min_{\mathbf{y}} \tilde{R}_{\text{avg}}(K, N, M, \mathbf{y})$$

$$s.t. \quad y_{n,s} \geq y_{n+1,s}, \quad \forall n \in \{1, 2, \dots, N-1\}, s \in \{1, 2, \dots, K\} \quad (7)$$

$$0 \leq y_{n,s} \leq 1, \quad \forall s \in \{0, 1, \dots, K\}, n \in \mathcal{N}, \quad (8)$$

$$\sum_{s=0}^K \binom{K}{s} y_{n,s} = 1, \quad \forall n \in \mathcal{N}, \quad (9)$$

$$\sum_{n=1}^N \sum_{s=1}^K \binom{K-1}{s-1} y_{n,s} \leq M, \quad (10)$$

where

$$\begin{aligned} \tilde{R}_{\text{avg}}(K, N, M, \mathbf{y}) \triangleq & \sum_{s=1}^K \binom{K}{s} \sum_{n=1}^N \left( \left( \sum_{n'=n}^N p_{n'} \right)^s - \left( \sum_{n'=n+1}^N p_{n'} \right)^s \right) y_{n,s-1} \\ & - \sum_{u=1}^{\min\{K,N\}} \sum_{s=1}^{K-u} \binom{K-u}{s} \sum_{i=1}^{K-u} \binom{K-u-i}{s-1} \sum_{n=1}^N P'_{i,u,n} y_{n,s-1}, \end{aligned} \quad (11)$$

and  $P'_{i,u,n}$  is given in (12)-(15) at the top of the next page.

*Proof:* Please refer to Appendix A. ■

Problem 2 is a linear program with  $N(K+1)$  variables and can be solved by using linear optimization techniques.

### C. Optimization for Uniform File Popularity

In this part, we consider a special case, i.e., the uniform file popularity ( $p_n = \frac{1}{N}$ , for all  $n \in \mathcal{N}$ ). First, we present another structural condition on the file partition parameter.

*Condition 3 (Symmetry w.r.t. File):* For all  $n \in \{1, 2, \dots, N-1\}$  and  $s \in \{1, 2, \dots, K\}$ , when  $p_n = p_{n+1}$ ,

$$y_{n,s} = y_{n+1,s}. \quad (16)$$

$$\begin{aligned}
P'_{i,u,n} = & \sum_{a \in \{1, \dots, u-2\}} \sum_{b_1 \in \{a, \dots, i+a-1\}} \sum_{b_2 \in \{u-a-1, \dots, K-i-a-1\}} \binom{K}{b_1, K-b_1-b_2, b_2} P_n^{K-b_1-b_2} \\
& \times \sum_{\mathcal{L}_1 \in \mathcal{L}_{a,1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}} \sum_{\mathcal{L}_2 \in \mathcal{L}_{a,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}} \\
& + \sum_{b_2 \in \{u-1, \dots, K-2\}} \binom{K}{b_2} P_n^{K-b_2} \sum_{\mathcal{L}_2 \in \mathcal{L}_{0,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}} \\
& + \sum_{b_1 \in \{u-1, \dots, K-2\}} \binom{K}{b_1} P_n^{K-b_1} \sum_{\mathcal{L}_1 \in \mathcal{L}_{u-1,1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}}, u \leq n, u+n \leq N+1
\end{aligned} \tag{12}$$

$$\begin{aligned}
P'_{i,u,n} = & \sum_{a \in \{u-1+n-N, \dots, u-2\}} \sum_{b_1 \in \{a, \dots, i+a-1\}} \sum_{b_2 \in \{u-a-1, \dots, K-i-a-1\}} \binom{K}{b_1, K-b_1-b_2, b_2} P_n^{K-b_1-b_2} \\
& \times \sum_{\mathcal{L}_1 \in \mathcal{L}_{a,1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}} \sum_{\mathcal{L}_2 \in \mathcal{L}_{a,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}} \\
& + \sum_{b_1 \in \{u-1, \dots, K-2\}} \binom{K}{b_1} P_n^{K-b_1} \sum_{\mathcal{L}_1 \in \mathcal{L}_{u-1,1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}}, u \leq n, u+n > N+1
\end{aligned} \tag{13}$$

$$\begin{aligned}
P'_{i,u,n} = & \sum_{a \in \{1, \dots, n-1\}} \sum_{b_1 \in \{a, \dots, i+a-1\}} \sum_{b_2 \in \{u-a-1, \dots, K-i-a-1\}} \binom{K}{b_1, K-b_1-b_2, b_2} P_n^{K-b_1-b_2} \\
& \times \sum_{\mathcal{L}_1 \in \mathcal{L}_{a,1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}} \sum_{\mathcal{L}_2 \in \mathcal{L}_{a,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}} \\
& + \sum_{b_2 \in \{u-1, \dots, K-2\}} \binom{K}{b_2} P_n^{K-b_2} \sum_{\mathcal{L}_2 \in \mathcal{L}_{0,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}}, u > n, u+n \leq N+1
\end{aligned} \tag{14}$$

$$\begin{aligned}
P'_{i,u,n} = & \sum_{a \in \{u-1+n-N, \dots, n-1\}} \sum_{b_1 \in \{a, \dots, i+a-1\}} \sum_{b_2 \in \{u-a-1, \dots, K-i-a-1\}} \binom{K}{b_1, K-b_1-b_2, b_2} P_n^{K-b_1-b_2} \\
& \times \sum_{\mathcal{L}_1 \in \mathcal{L}_{a,1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}} \\
& \times \sum_{\mathcal{L}_2 \in \mathcal{L}_{a,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}}, u > n, u+n > N+1
\end{aligned} \tag{15}$$

---

Condition 3 indicates that for all  $n \in \{1, 2, \dots, N-1\}$  and  $s \in \{1, 2, \dots, K\}$ , when

$p_n = p_{n+1}$ , the size of subfiles  $W_{n,\mathcal{S}}$ ,  $\mathcal{S} \subseteq \{\widehat{\mathcal{S}} \subseteq \mathcal{K} : |\widehat{\mathcal{S}}| = s\}$  is the same as that of subfiles  $W_{n+1,\mathcal{S}}$ ,  $\mathcal{S} \subseteq \{\widehat{\mathcal{S}} \subseteq \mathcal{K} : |\widehat{\mathcal{S}}| = s\}$ . Condition 3 ensures zero variance of the lengths of messages involved in the coded-multicast XOR operations, hence avoiding “bit waste” effect and further increasing coded-multicasting opportunities for the uniform file popularity. Later, we shall show that imposing this condition will not lose optimality of Problem 2 under the uniform file popularity.

By Condition 3, we can set

$$y_{n,s} = z_s, \quad \forall s \in \{0, 1, \dots, K\}, \quad n \in \mathcal{N}. \quad (17)$$

Here,  $z_s$  can be viewed as the size of each subfile of type  $s$ , normalized by the file size  $F$ . Let  $\mathbf{z} \triangleq (z_s)_{s \in \{0,1,\dots,K\}}$ .

Next, we simplify Problem 1 under Conditions 1 and 3. First, we introduce some notations. Let  $P''_u \triangleq \Pr[|\mathcal{K}(\mathbf{D})| = u]$ , for all  $u \in \{1, 2, \dots, \min\{K, N\}\}$ . Note that event  $|\mathcal{K}(\mathbf{D})| = u$  corresponds to the event in the “balls into bins” problem that there are  $u$  nonempty bins after placing  $K$  balls uniformly at random into  $N$  bins. By using results for the “balls into bins” problem in the uniform case [14] and considering Conditions 1 and 3, we have the following result.

*Lemma 2 (Simplification of Problem 1 for Uniform File Popularity):* Under Conditions 1 and 3, Problem 1 can be converted into:

*Problem 3 (Simplified Problem for Uniform File Popularity):*

$$\begin{aligned} \widehat{R}_{\text{avg}}^*(K, N, M) &\triangleq \min_{\mathbf{z}} \sum_{s=0}^{K-1} \binom{K}{s+1} z_s - \sum_{u=1}^{\min\{K,N\}} P''_u \sum_{s=0}^{K-u-1} \binom{K-u}{s+1} z_s \\ \text{s.t.} \quad &0 \leq z_s \leq 1, \quad s \in \{0, 1, \dots, K\}, \end{aligned} \quad (18)$$

$$\sum_{s=0}^K \binom{K}{s} z_s = 1, \quad (19)$$

$$\sum_{s=0}^K \binom{K}{s} s z_s \leq \frac{KM}{N}, \quad (20)$$

where  $P''_u = \left\{ \begin{matrix} K \\ u \end{matrix} \right\} \frac{(N)_u u!}{N^K}$ , and  $\left\{ \begin{matrix} K \\ u \end{matrix} \right\}$  is the Stirling number of the second kind.

*Proof:* Please refer to Appendix B. ■

Problem 3 is a linear program with  $K + 1$  variables and can be solved more efficiently than Problem 1.

Finally, we discuss the relation between an optimal solution of Problem 3 and Yu *et al.*'s centralized coded caching scheme [9].<sup>2</sup> Using KKT conditions, we have

*Lemma 3 (Optimal Solution to Problem 3):* For cache size  $M \in \{0, \frac{N}{K}, \frac{2N}{K}, \dots, N\}$ ,  $\mathbf{z}^* \triangleq (z_s^*)_{s \in \{0, 1, \dots, K\}}$  is an optimal solution to Problem 3, where

$$z_s^* = \begin{cases} \frac{1}{\binom{\frac{K}{N}}{\frac{KM}{N}}}, & s = \frac{KM}{N} \\ 0, & s \in \{0, 1, \dots, K\} \setminus \{\frac{KM}{N}\}, \end{cases} \quad (21)$$

and the optimal value of Problem 3 is given by<sup>3</sup>

$$\hat{R}_{\text{avg}}^*(K, N, M) = \frac{K(1 - M/N)}{1 + KM/N} - \sum_{u=1}^{\min\{K, N\}} P_u'' \left( \frac{K-u}{\frac{KM}{N} + 1} \right) / \left( \frac{K}{\frac{KM}{N}} \right). \quad (22)$$

*Proof:* Please refer to Appendix C. ■

Lemma 3 indicates that Yu *et al.*'s centralized coded caching scheme corresponds to an optimal solution of Problem 3. In addition, the optimal average load  $\hat{R}_{\text{avg}}^*(K, N, M)$  in (22) is equivalent to that in [9]. Specifically, the first term in (22) corresponds to the worst-case load in [6] and the second term in (22), which is given in explicit form as opposed to the implicit form containing an expectation w.r.t. the random requests given in [9], indicates the load reduction due to the ability of the delivery strategy to take advantage of common requests. Note that it has been shown that Yu *et al.*'s centralized coded caching scheme is optimal among all uncoded placement and all delivery under the uniform file popularity. Thus, for the uniform file popularity, Conditions 1 and 3 are actually optimal properties.

#### IV. AVERAGE LOAD MINIMIZATION WITH SUBPACKETIZATION LEVEL CONSTRAINT

In this section, we minimize the average load by optimizing the file partition parameter under the subpacketization level constraint for each file, which is given by

$$\|\mathbf{x}_n\|_0 \leq \hat{F}, \quad n \in \mathcal{N}, \quad (23)$$

<sup>2</sup>Yu *et al.*'s centralized coded caching scheme focuses on cache size  $M \in \{0, \frac{N}{K}, \frac{2N}{K}, \dots, N\}$ , so that  $\frac{KM}{N}$  is an integer in  $\{0, 1, \dots, K\}$ . For general  $M \in [0, N]$ , the worst-case load can be achieved by memory sharing.

<sup>3</sup>In this paper, we define  $\binom{n}{k} = 0$  when  $k > n$  [9].

where  $\|\mathbf{x}_n\|_0 \triangleq \sum_{S \in \mathcal{K}} \mathbf{1}[x_{n,S} \neq 0] \in \{1, 2, \dots, 2^K\}$  denotes the  $\ell_0$ -norm of the vector  $\mathbf{x}_n$ , i.e., the total number of subfiles for file  $W_n$ , and  $\widehat{F} \in \{1, 2, \dots, 2^K\}$  represents the maximum admissible subpacketization level for all files. To the best of our knowledge, this is the first work explicitly considering the subpacketization level constraint in optimizing coded caching design.

#### A. Problem Formulation

In this part, we minimize the average load under the file partition constraints in (1) and (2), the cache memory constraint in (3), and the subpacketization level constraint in (23).

*Problem 4 (Optimization for Arbitrary File Popularity with Subpacketization Constraint):*

$$\begin{aligned} R_{\text{avg}}^\dagger(K, N, M) &\triangleq \min_{\mathbf{x}} R_{\text{avg}}(K, N, M, \mathbf{x}) \\ \text{s.t.} \quad &(1), (2), (3), (23), \end{aligned}$$

where  $R_{\text{avg}}(K, N, M, \mathbf{x})$  is given by (4).

Compared with Problem 1, Problem 4 has an extra constraint, i.e., the subpacketization level constraint in (23). There are two main challenges in solving Problem 4. First, Problem 4 is an NP-Hard problem due to the combinatorial constraint in (23) involving the  $\ell_0$ -norm [13]. Second, as in Problem 1, the number of variables in Problem 4 is  $N2^K$ , which is huge, especially when  $K$  and  $N$  are large.

#### B. Simplified Formulation

There are extensive research dealing with optimization problems involving the  $\ell_0$ -norm. Those works can be divided into three main categories according to the way of treating the  $\ell_0$ -norm, i.e., convex approximation, non-convex approximation, and non-convex exact reformulation [15]. For the category of convex approximation, one of the best known approaches is approximating the  $\ell_0$ -norm with the  $\ell_1$ -norm [15]. If the original optimization problem is convex except the constraint involving the  $\ell_0$ -norm, this convex approximation approach can transform the original NP-Hard problem into a convex problem. However, it has been shown that an optimal solution of the approximated convex problem is not always sparse (may not be a feasible solution of the original problem) [16].<sup>4</sup> For the category of non-convex approximation, a variety of sparsity-

<sup>4</sup>Given the constraint in (9), the  $\ell_1$ -norm of  $\mathbf{x}_n$  is equal to a constant, i.e.,  $\|\mathbf{x}_n\|_1 = \sum_{i=1}^{2^K} x_i = 1$ . Thus, replacing  $\|\mathbf{x}_n\|_0$  with  $\|\mathbf{x}_n\|_1$  in (23) results in the constraint  $1 \leq \widehat{F}$ , which always holds and cannot restrain  $\mathbf{x}_n$ .

inducing penalty functions, e.g., the  $\ell_p$  pseudo-norm with  $0 < p < 1$  [17], exponential concave function [18], and logarithmic function [19], have been proposed to approximate the  $\ell_0$ -norm. In general, non-convex approximation can provide better sparsity than convex approximation, but may still not provide a feasible solution of the original problem. Few works focus on non-convex exact reformulation, which is proposed to guarantee the equivalence between the reformulated problem and the original problem. Using exact penalty techniques, [15] and [20] show that the reformulated problems with suitable parameters are equivalent to the original problems (share the same feasible solutions with the original problems). However, the reformulated problems are quite convoluted as they rely on several parameters [13]. In the recent work [13], the authors propose an exact DC reformulation which is simpler than the reformulated problems proposed in [15] and [20], and then obtain a stationary point of the DC problem using a DC algorithm. In the following, we use the exact DC reformulation method in [13] in order to obtain a simple equivalent formulation of the original problem.

We first simplify Problem 4 to facilitate a low-complexity solution. Let  $a_{[i]}$  denote the element whose value is the  $i$ -th largest among the  $m$  elements of the vector  $\mathbf{a}$ , i.e.,  $a_{[1]} \geq a_{[2]} \geq \dots \geq a_{[m]}$ . Let  $\|\mathbf{a}\|_{l_{gst}, \hat{F}}$  denote the largest- $\hat{F}$  norm of the vector  $\mathbf{a}$ , i.e.,  $\|\mathbf{a}\|_{l_{gst}, \hat{F}} \triangleq |a_{[1]}| + |a_{[2]}| + \dots + |a_{[\hat{F}]}|$  [21]. Using the method for obtaining Problem 2 and Theorem 1 of [13] for simplifying the constraint in (23) under Conditions 1 and 2, we have the following result.

*Lemma 4 (Simplification for Problem 4 for Arbitrary File Popularity):* Under Conditions 1 and 2, Problem 4 can be converted into:

*Problem 5 (Simplified Problem for Arbitrary File Popularity):*

$$\begin{aligned} \tilde{R}_{\text{avg}}^\dagger(K, N, M, \mathbf{y}) &\triangleq \min_{\mathbf{y}} \quad \tilde{R}_{\text{avg}}(K, N, M, \mathbf{y}) \\ \text{s.t.} \quad &(7), (8), (9), (10), \\ &\|U_n \mathbf{y}\|_{l_{gst}, \hat{F}} \geq 1, \quad n \in \mathcal{N}, \end{aligned} \tag{24}$$

where  $\tilde{R}_{\text{avg}}(K, N, M, \mathbf{y})$  is given by (11),

$$U_n \triangleq [\mathbf{0}, \dots, \mathbf{0}, \underbrace{J}_{n\text{-th block}}, \mathbf{0}, \dots, \mathbf{0}] \tag{25}$$

denotes the block matrix of the dimension  $2^K \times (K+1)N$ , with  $2^K \times (K+1)$  matrix  $J \triangleq (j_{m,l})_{m \in \{1, \dots, 2^K\}, l \in \{1, \dots, K+1\}}$  as its  $n$ -th block and  $2^K \times (K+1)$  zero matrices as other blocks,

and the element of row  $m$  and column  $l$  of  $J$  is

$$j_{m,l} = \begin{cases} 1, & \sum_{i=1}^{l-1} \binom{K}{i-1} < m \leq \sum_{i=1}^l \binom{K}{i-1} \\ 0, & \text{otherwise} \end{cases}. \quad (26)$$

*Proof:* Please refer to Appendix D. ■

The number of variables in Problem 5 is  $N(K+1)$ , which is much smaller than that of Problem 4, i.e.,  $N2^K$ . In addition, compared with Problem 2, Problem 5 has an extra constraint in (24), which has two advantages over the subpacketization level constraint in (23): (i) the constraint in (24) is a DC constraint, making Problem 2 a DC problem, which can be solved by a DC algorithm in polynomial time; (ii) a subgradient of  $\|U_n \mathbf{y}\|_{l_{gst}, \hat{F}}$  can be efficiently computed, making the DC algorithm an efficient one.

Thus, in the following, we solve Problem 5 by using a DC algorithm. The main idea of the DC algorithm is to iteratively solve a sequence of convex problems, each of which is obtained by linearizing the second term of the objective function of the DC problem. A subgradient of the second term of the objective function is required in the linearization in each iteration. Thus, to solve Problem 5, we first obtain a subgradient of  $\|U_n \mathbf{y}\|_{l_{gst}, \hat{F}}$  by extending the closed-form expression of a subgradient of  $\|\mathbf{y}\|_{l_{gst}, \hat{F}}$  given in [13].

**Lemma 5 (Subgradient of  $\|U_n \mathbf{y}\|_{l_{gst}, \hat{F}}$ ):**  $\mathbf{g}_n(\mathbf{y}) \triangleq (g_{n,i}(\mathbf{y}))_{i \in \{1, 2, \dots, N(K+1)\}}$  is a subgradient of  $\|U_n \mathbf{y}\|_{l_{gst}, \hat{F}}$ , where

$$g_{n,(m-1)(K+1)+[i]}(\mathbf{y}) = \begin{cases} \binom{K}{[i]-1}, & m = n, i \in \{1, \dots, I-1\} \\ \hat{F} - \sum_{i=1}^{I-1} \binom{K}{[i]-1}, & m = n, i = I \\ 0, & \text{otherwise} \end{cases}, \quad (27)$$

$[i]$  represents the index of  $i$ -th largest element in  $\mathbf{y}_n$ , and  $I$  satisfies  $\sum_{i=1}^{I-1} \binom{K}{[i]-1} \leq \hat{F}$  and  $\sum_{i=1}^I \binom{K}{[i]-1} > \hat{F}$ .

*Proof:* Please refer to Appendix E. ■

Then, based on the subgradient  $\mathbf{g}_n(\mathbf{y})$  given in Lemma 5, we can obtain a stationary point of Problem 5 using the DC algorithm as summarized in Alg. 2 [22]. As in [15], to approach a globally optimal solution of Problem 5, we obtain multiple stationary points of Problem 5 by performing the DC algorithm multiple times, each with a random initial feasible point of Problem 5, and adopt the stationary point with the lowest average load among all the obtained stationary points of Problem 5.



---

**Algorithm 2** DC Algorithm for Solving Problem 5
 

---

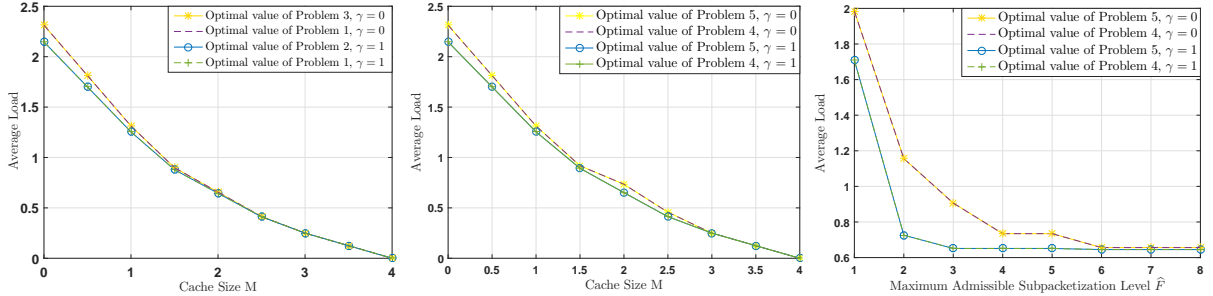
- 1: Find an initial feasible point  $\mathbf{y}^{(0)}$  of Problem 5 and set  $t = 0$
- 2: **repeat**
- 3: Set  $\mathbf{y}^{(t+1)}$  to be an optimal solution of the convex problem:

$$\min_{\mathbf{y}} \quad \tilde{R}_{\text{avg}}(K, N, M, \mathbf{y})$$

$$s.t. \quad (7), (8), (9), (10)$$

$$\|U_n \mathbf{y}^{(t)}\|_{l_{gst, \hat{F}}} + \mathbf{g}_n(\mathbf{y}^{(t)})^T (\mathbf{y} - \mathbf{y}^{(t)}) \geq 1, \quad n \in \mathcal{N} \quad (28)$$

- 4: Set  $t = t + 1$
  - 5: **until**  $\tilde{R}_{\text{avg}}(K, N, M, \mathbf{y}^{(t-1)}) - \tilde{R}_{\text{avg}}(K, N, M, \mathbf{y}^{(t)}) \leq \delta$
- 



(a) Case without considering the subpacketization level issue. (b) Case considering the subpacketization level issue at  $\hat{F} = 5$ . (c) Case considering the subpacketization level issue at  $M = 2$ .

Fig. 2: Verification of Conditions 1, 2 and 3 in both cases at  $K = 3$  and  $N = 4$ .

## V. NUMERICAL RESULTS

In the simulation, we assume that the file popularity follows Zipf distribution, i.e.,  $p_n = \frac{n^{-\gamma}}{\sum_{n \in \mathcal{N}} n^{-\gamma}}$  for all  $n \in \mathcal{N}$ , where  $\gamma$  is the Zipf exponent. Fig. 2 (a) shows the optimal values of Problems 1, 2 and 3, verifying that Conditions 1, 2 and 3 are optimal conditions in the case without considering the subpacketization level issue. Fig. 2 (b) and Fig. 2 (c) show the optimal values of Problems 4 and 5, verifying that Conditions 1 and 2 are optimal conditions in the case considering the subpacketization level issue.

Fig. 3 compares the average load of our optimized parameter-based scheme, the average loads of Maddah-Ali-Niesen's centralized scheme [6], Jin *et al.*'s centralized scheme [7], Yu

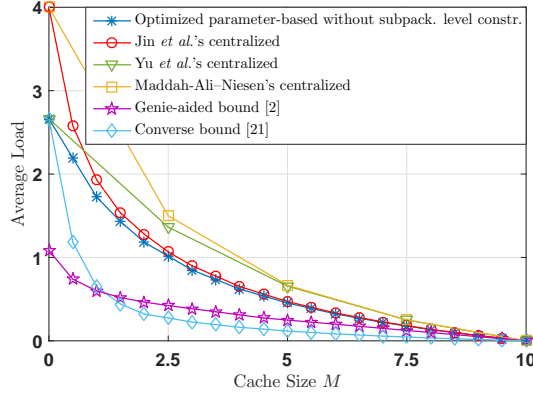


Fig. 3: Average load versus  $M$  in the case without considering the subpacketization level issue at  $K = 4$ ,  $N = 10$  and  $\gamma = 1.5$ . Note that Maddah-Ali-Niesen's and Yu *et al.*'s centralized coded caching schemes mainly focus on the cache size  $M \in \{0, \frac{N}{K}, \frac{2N}{K}, \dots, N\}$ . For other  $M \in [0, N]$ , the average loads of Maddah-Ali-Niesen's and Yu *et al.*'s centralized coded caching schemes are achieved by memory sharing [6], [9].

*et al.*'s centralized scheme [9], the genie-aided converse bound in [7] and the converse bound in [23], all without considering the subpacketization level issue. From Fig. 3, we can see that the optimized parameter-based scheme outperforms the three baseline schemes. The gain over Jin *et al.*'s optimized centralized coded caching scheme follows by using an extended version of the improved delivery strategy of Yu *et al.*, that takes advantage of common requests (which occur with positive probability in the case of random requests). The gain over Yu *et al.*'s centralized coded caching scheme is due to exploiting the explicit knowledge of the file popularity in the optimization of content placement.<sup>5</sup> In addition, the optimized average load is close to the converse bounds, implying that the optimal value obtained by solving Problem 2 is close to optimal.

Fig. 4 compares the average load of our optimized parameter-based scheme considering the subpacketization level constraint, the average loads of Tang *et al.*'s scheme [10] and Pareto-optimal PDA [12] both considering the subpacketization level issue. From Fig. 4, we see that the

<sup>5</sup>It has been proved in [7] that the optimized parameter-based scheme in [7] outperforms Maddah-Ali-Niesen's centralized coded caching scheme [6].

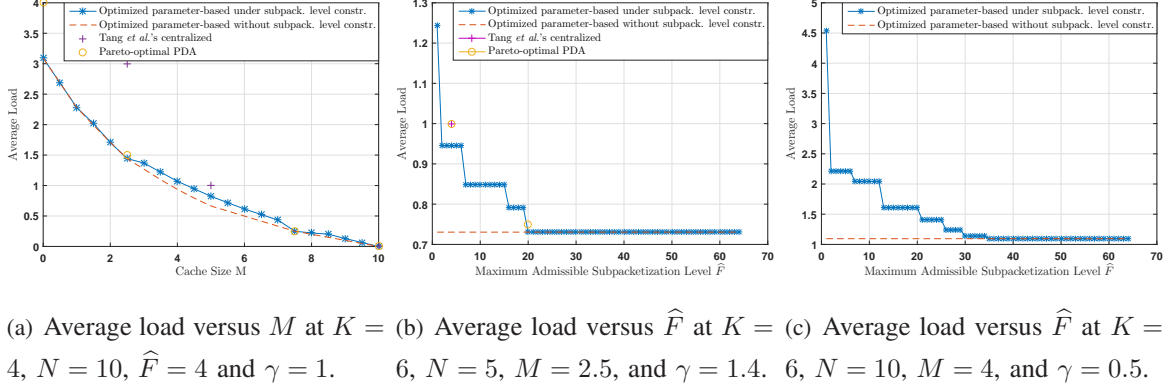


Fig. 4: Average load versus  $M$  and  $\hat{F}$  in the case considering the subpacketization level issue. To obtain the average load of our optimized parameter-based scheme under the subpacketization level constraint, we solve Problem 5 by performing Alg. 2 with  $\delta = 0.0001$  100 times each with a random initial feasible point and adopt the locally optimal solution with the lowest average load among all the obtained locally optimal solutions of Problem 5 that are also feasible solutions of Problem 5.

average load of our optimized parameter-based scheme without considering the subpacketization level constraint serves as a lower bound of the proposed one considering the subpacketization level constraint. Note that in Fig. 4 (a), for the considered  $K, N, \hat{F}$ , Tang *et al.*'s scheme is applicable only at  $M \in \{0, 2.5, 5, 7.5, 10\}$  and Pareto-optimal PDA is applicable only at  $M \in \{2.5, 5, 7.5, 10\}$ ; in Fig. 4 (b), for the considered  $K, N, M$ , Tang *et al.*'s scheme is applicable only at  $\hat{F} = 4$  and Pareto-optimal PDA is applicable only at  $\hat{F} \in \{4, 20\}$ ; in Fig. 4 (c), for the considered  $K, N, M$ , Tang *et al.*'s scheme and Pareto-optimal PDA are not applicable at any  $\hat{F}$ . In addition, from Fig. 4, we can see that our optimized parameter-based scheme outperforms Tang *et al.*'s scheme and Pareto-optimal PDA in terms of both the average load and application region. From Fig. 4, we can see that our optimized parameter-based scheme considering the subpacketization level constraint can achieve significantly lower subpacketization level than the one without considering the subpacketization level constraint, at the cost of a small increase of the average load. This means that sacrificing a small average load gain can achieve a huge subpacketization level reduction under the considered setting.

## VI. CONCLUSION

In this paper, we first presented a class of centralized coded caching schemes consisting of a general content placement strategy specified by a file partition parameter, enabling efficient and flexible content placement, and a specific content delivery strategy, enabling load reduction by exploiting common requests of different users. Then we considered two cases: the case without considering the subpacketization level issue and the case considering the subpacketization level issue. In the first case, we formulated the coded caching optimization problem over the considered class of schemes to minimize the average load under an arbitrary file popularity. Imposing some conditions on the file partition parameter, we transformed the original optimization problem with  $N2^K$  variables into a linear program with  $N(K + 1)$  variables under an arbitrary file popularity and a linear program with  $K + 1$  variables under the uniform file popularity. In the second case, we formulated the coded caching optimization problem over the considered class of schemes to minimize the average load under an arbitrary file popularity subject to subpacketization level constraints involving the  $\ell_0$ -norm. Imposing the same conditions and using the exact DC reformulation method, we converted the original problem with  $N2^K$  variables into a simplified DC problem with  $N(K + 1)$  variables, which is solved using DC algorithm. Finally, numerical results verify the optimality of the imposed conditions and demonstrate the advantages of the optimized scheme over existing schemes in both cases.

## APPENDIX A: PROOF OF LEMMA 1

By (5), it can be easily shown that the constraints in (1), (2) and (3) of Problem 1 can be converted into (8), (9) and (10). Now, we show that the objective function of Problem 1 in (4) can be converted into the objective function of Problem 2 in (11). First, by (5), we have

$$\begin{aligned}
 R_{\text{avg}}(K, N, M, \mathbf{x}) &\stackrel{(a)}{=} \sum_{\mathbf{d} \in \mathcal{N}^K} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{s=1}^K \left( \sum_{\mathcal{S} \subseteq \mathcal{K}; |\mathcal{S}|=s} \max_{k \in \mathcal{S}} y_{d_k, s-1} - \sum_{\mathcal{S} \subseteq \mathcal{K} \setminus \underline{\mathcal{K}}(\mathbf{d}); |\mathcal{S}|=s} \max_{k \in \mathcal{S}} y_{d_k, s-1} \right) \\
 &\stackrel{(b)}{=} \sum_{s=1}^K \binom{K}{s} \sum_{n=1}^N \left( \left( \sum_{n'=n}^N p_{n'} \right)^s - \left( \sum_{n'=n+1}^N p_{n'} \right)^s \right) y_{n, s-1} - \sum_{\mathbf{d} \in \mathcal{N}^K} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{s=1}^K \sum_{\mathcal{S} \subseteq \mathcal{K} \setminus \underline{\mathcal{K}}(\mathbf{d}); |\mathcal{S}|=s} \max_{k \in \mathcal{S}} y_{d_k, s-1},
 \end{aligned} \tag{29}$$

where (a) is due to (5), and (b) is due to Lemma 3 in [7]. Then, by using the same simplification method in the proof of Proposition 1 in [8], we further simplify the second term in (29) into:

$$\begin{aligned}
& \sum_{\mathbf{d} \in \mathcal{N}^K} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{s=1}^{K-|\underline{\mathcal{K}}(\mathbf{d})|} \sum_{i=1}^{K-|\underline{\mathcal{K}}(\mathbf{d})|} \binom{K-|\underline{\mathcal{K}}(\mathbf{d})|-i}{s-1} y_{d_{k_i}, s-1} \\
&= \sum_{u=1}^{\min\{K, N\}} \sum_{\mathbf{d} \in \mathcal{D}_u} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{s=1}^{K-u} \sum_{i=1}^{K-u-s+1} \binom{K-u-i}{s-1} y_{d_{k_i}, s-1} \\
&= \sum_{u=1}^{\min\{K, N\}} \sum_{s=1}^{K-u} \sum_{i=1}^{K-u-s+1} \binom{K-u-i}{s-1} \sum_{\mathbf{d} \in \mathcal{D}_u} \left( \prod_{k=1}^K p_{d_k} \right) y_{d_{k_i}, s-1} \\
&= \sum_{u=1}^{\min\{K, N\}} \sum_{s=1}^{K-u} \sum_{i=1}^{K-u-s+1} \binom{K-u-i}{s-1} \sum_{n=1}^N P'_{i,u,n} y_{n, s-1},
\end{aligned}$$

where  $d_{k_i}$  denotes the  $i$ -th most popular file in  $(D_k)_{k \in \mathcal{K} \setminus \underline{\mathcal{K}}(\mathbf{d})}$ , and  $\mathcal{D}_u \triangleq \left\{ \mathbf{d} \in \mathcal{N}^K : \sum_{n=1}^N \mathbf{1}[d_n > 0] = u \right\}$ .

It remains to derive  $P'_{i,u,n}$  by connecting the event  $\tilde{D}_{u,\langle i \rangle} = n$  to the “balls into bins” problem. Note that the event that  $K$  balls are placed in an i.i.d. manner into  $N$  bins where each of the  $K$  balls is placed into bin  $n'$  (the bin with index  $n'$ ) with probability  $p_{n'}$  corresponds to the event that each user randomly and independently requests file  $n' \in \mathcal{N}$  with probability  $p_{n'}$  (represented by random variable  $D_k, k \in \mathcal{K}$ ). Let  $\mathcal{E}_u$  denote the event that there are exactly  $u$  nonempty bins, which corresponds to the event that there are  $u$  representative users, i.e.,  $|\underline{\mathcal{K}}(\mathbf{D})| = u$ . Let  $\mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}}$  denote the event that  $b_1$  balls fall into  $a$  different bins with indices smaller than or equal to  $n-1$ , which corresponds to the event that there are  $b_1$  users requesting  $a$  different files with file indices smaller than or equal to  $n-1$ . Let  $\mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}}$  denote the event that  $b_2$  balls fall into  $u-a-1$  different bins with indices larger than or equal to  $n+1$ , which corresponds to the event that there are  $b_2$  users requesting  $u-a-1$  different files with file indices larger than or equal to  $n+1$ . Let  $\mathcal{E}_{b_3, 1, u}^{3, \{n\}}$  denote the event that  $b_3$  balls fall into bin  $n$ , which corresponds to the event that there are  $b_3$  users requesting file  $n$ . Let  $\Theta_a(u, n)$  denote the range of  $a$  in  $\mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}}$  and  $\mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}}$  and let  $\Theta_b(n, u, a)$  denote the range of  $(b_1, b_2, b_3)$

in  $\mathcal{E}_{b_1,a,u}^{1,\{1,2,\dots,n-1\}}$ ,  $\mathcal{E}_{b_2,u-a-1,u}^{2,\{n+1,n+2,\dots,N\}}$  and  $\mathcal{E}_{b_3,1,u}^{3,\{n\}}$  for a given  $a$ , where

$$\Theta_a(u, n) = \begin{cases} \{0, 1, \dots, u-1\}, & u \leq n, u+n \leq N+1 \\ \{u-1+n-N, \dots, u-1\}, & u \leq n, u+n > N+1 \\ \{0, \dots, n-1\}, & u > n, u+n \leq N+1 \\ \{u-1+n-N, \dots, n-1\}, & u > n, u+n > N+1 \end{cases}, \quad (30)$$

and

$$\Theta_b(n, u, a) = \begin{cases} \{(b_1, b_2, b_3) : b_1 \in \{a, \dots, K\}, b_2 \in \{u-a-1, \dots, K\}, b_1+b_2+b_3=K\}, & a \in \{1, \dots, u-2\} \cap \Theta_a(u, n) \\ \{(b_1, b_2, b_3) : b_1 = 0, b_2 \in \{u-1, \dots, K\}, b_1+b_2+b_3=K\}, & a \in \{0\} \cap \Theta_a(u, n) \\ \{(b_1, b_2, b_3) : b_1 \in \{u-1, \dots, K\}, b_2 = 0, b_1+b_2+b_3=K\}, & a \in \{u-1\} \cap \Theta_a(u, n). \end{cases} \quad (31)$$

We know that

$$\mathcal{E}_u = \bigcup_{a \in \Theta_a(u, n)} \bigcup_{(b_1, b_2, b_3) \in \Theta_b(a)} \left( \mathcal{E}_{b_1,a,u}^{1,\{1,2,\dots,n-1\}} \cap \mathcal{E}_{b_2,u-a-1,u}^{2,\{n+1,n+2,\dots,N\}} \cap \mathcal{E}_{b_3,1,u}^{3,\{n\}} \right). \quad (32)$$

Then, remove one ball from each of the  $|\underline{\mathcal{K}}(\mathbf{D})|$  nonempty bins and consider the remaining balls, which corresponds to consider the requests in  $(D_k)_{k \in \mathcal{K} \setminus \underline{\mathcal{K}}(\mathbf{D})}$ . Let  $\xi_{i-1}^{1,\{1,2,\dots,n-1\}}$  denote the event that there are at most  $i-1$  balls placed in the bins with indices smaller than or equal to  $n-1$  and let  $\xi_i^{2,\{1,2,\dots,n\}}$  denote the event that there are at least  $i$  balls placed in the bins with indices smaller than or equal to  $n$ . Note that  $\xi_{i-1}^{1,\{1,2,\dots,n-1\}} \cap \xi_i^{2,\{1,2,\dots,n\}} \cap \mathcal{E}_u$  is equivalent to the event  $\tilde{D}_{u,\langle i \rangle} = n$ , implying that  $P'_{i,u,n} = \Pr[\tilde{D}_{u,\langle i \rangle} = n] = \Pr[\xi_{i-1}^{1,\{1,2,\dots,n-1\}} \cap \xi_i^{2,\{1,2,\dots,n\}} \cap \mathcal{E}_u]$ . Similar to (32),  $\xi_{i-1}^{1,\{1,2,\dots,n-1\}} \cap \xi_i^{2,\{1,2,\dots,n\}} \cap \mathcal{E}_u$  can be represented as

$$\begin{aligned} & \xi_{i-1}^{1,\{1,2,\dots,n-1\}} \cap \xi_i^{2,\{1,2,\dots,n\}} \cap \mathcal{E}_u \\ &= \bigcup_{a \in \Theta_a(u, n)} \bigcup_{(b_1, b_2, b_3) \in \tilde{\Theta}_b(i, u, n, a)} \left( \mathcal{E}_{b_1,a,u}^{1,\{1,2,\dots,n-1\}} \cap \mathcal{E}_{b_2,u-a-1,u}^{2,\{n+1,n+2,\dots,N\}} \cap \mathcal{E}_{b_3,1,u}^{3,\{n\}} \right), \end{aligned} \quad (33)$$

where  $\tilde{\Theta}_b(i, u, n, a)$  denotes the range of  $(b_1, b_2, b_3)$ . In the following, we determine  $\tilde{\Theta}_b(i, u, n, a)$ :

- For any  $a \in \{1, \dots, u-2\} \cap \Theta_a(u, n)$ , by  $\mathcal{E}_{b_1,a,u}^{1,\{1,2,\dots,n-1\}}$  and  $\xi_{i-1}^{1,\{1,2,\dots,n-1\}}$ , we have

$$b_1 \leq i + a - 1; \quad (34)$$

by  $\mathcal{E}_{b_2,u-a-1,u}^{2,\{n+1,n+2,\dots,N\}}$  and  $\xi_i^{2,\{1,2,\dots,n\}}$ , we have

$$b_2 \leq K - i - a - 1. \quad (35)$$

By (34), (35) and (31), for all  $a \in \{1, \dots, u-2\} \cap \Theta_a(u, n)$ , we have

$$\begin{aligned} & \tilde{\Theta}_b(i, u, n, a) \\ &= \{(b_1, b_2, b_3) : b_1 \in \{a, \dots, i+a-1\}, b_2 \in \{u-a-1, \dots, K-i-a-1\}, b_1+b_2+b_3=K\}. \end{aligned} \quad (36)$$

- For  $a \in \{0\} \cap \Theta_a(u, n)$ , by  $\mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}}$  and  $\xi_{i-1}^{1, \{1, 2, \dots, n-1\}}$ , we have

$$i = 1, b_1 = 0, \quad (37)$$

implying that  $\mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}}$  does not happen; by  $\mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}}$  and  $\xi_i^{2, \{1, 2, \dots, n\}}$ , we have

$$b_2 \leq K - i - a - 1. \quad (38)$$

By (37), (38) and (31), for all  $a \in \{0\} \cap \Theta_a(u, n)$ , we have

$$\tilde{\Theta}_b(i, u, n, a) = \{(b_1, b_2, b_3) : b_1 = 0, b_2 \in \{u-1, \dots, K-a-2\}, b_1+b_2+b_3=K\}. \quad (39)$$

- For  $a \in \{u-1\} \cap \Theta_a(u, n)$ , by  $\mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}}$  and  $\xi_{i-1}^{1, \{1, 2, \dots, n-1\}}$ , we have

$$i = K - u, b_2 = 0, \quad (40)$$

implying that  $\mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}}$  does not happen; by  $\mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}}$  and  $\xi_i^{2, \{1, 2, \dots, n\}}$ , we have

$$b_1 \leq a + i - 1. \quad (41)$$

By (40), (41) and (31), for all  $a \in \{u-1\} \cap \Theta_a(u, n)$ , we have

$$\tilde{\Theta}_b(i, u, n, a) = \{(b_1, b_2, b_3) : b_1 \in \{u-1, \dots, K-u+a-1\}, b_2 = 0, b_1+b_2+b_3=K\}. \quad (42)$$

By (33), (36), (39) and (42), we have

$$\begin{aligned} & \xi_{i-1}^{1, \{1, 2, \dots, n-1\}} \cap \xi_i^{2, \{1, 2, \dots, n\}} \cap \mathcal{E}_u \\ &= \left( \bigcup_{a \in \{1, \dots, u-2\} \cap \Theta_a(u, n)} \bigcup_{(b_1, b_2, b_3) \in \tilde{\Theta}_b(i, u, n, a)} \left( \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_1-b_2, 1, u}^{3, \{n\}} \right) \right) \\ & \cup \left( \bigcup_{a \in \{0\} \cap \Theta_a(u, n)} \bigcup_{(b_1, b_2, b_3) \in \tilde{\Theta}_b(i, u, n, a)} \left( \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_2, 1, u}^{3, \{n\}} \right) \right) \end{aligned}$$

$$\bigcup \left( \bigcup_{a \in \{u-1\} \cap \Theta_a(u,n)} \bigcup_{(b_1, b_2, b_3) \in \tilde{\Theta}_b(i, u, n, a)} \left( \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{K-b_1, 1, u}^{3, \{n\}} \right) \right). \quad (43)$$

Based on (43), we have

$$\begin{aligned} P'_{i, u, n} &= \sum_{a \in \{1, \dots, u-2\} \cap \Theta_a(u, n)} \sum_{b_1 \in \{a, \dots, i+a-1\}} \sum_{b_2 \in \{u-a-1, \dots, K-i-a-1\}} \binom{K}{b_1, K-b_1-b_2, b_2} \\ &\times \Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_1-b_2, 1, u}^{3, \{n\}} \right] \\ &+ \sum_{a \in \{0\} \cap \Theta_a(u, n)} \sum_{b_2 \in \{u-1, \dots, K-a-2\}} \binom{K}{b_2} \Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_2, 1, u}^{3, \{n\}} \right] \\ &+ \sum_{a \in \{u-1\} \cap \Theta_a(u, n)} \sum_{b_1 \in \{u-1, \dots, K-u-a-1\}} \binom{K}{b_1} \Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{K-b_1, 1, u}^{3, \{n\}} \right], \end{aligned} \quad (44)$$

where  $\binom{K}{b_1, K-b_1-b_2, b_2}$  is the total number of partitions of  $K$  balls into three parts with the numbers of balls  $b_1$ ,  $b_2$  and  $K-b_1-b_2$ ,  $\binom{K}{b_2}$  is the total number of partitions of  $K$  balls into two parts with the numbers of balls  $b_2$  and  $K-b_2$ , and  $\binom{K}{b_1}$  is the total number of partitions of  $K$  balls into two parts with the numbers of balls  $b_1$  and  $K-b_1$ . To calculate (44), we first calculate  $\Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_1-b_2, 1, u}^{3, \{n\}} \right]$  for all  $a \in \{1, \dots, u-2\} \cap \Theta_a(u, n)$ ,  $b_1 \in \{a, a+1, \dots, i+a-1\}$  and  $b_2 \in \{u-a-1, u-a, \dots, K-i-a-1\}$ . By using results from “balls into bins” problem, we have

$$\begin{aligned} &\Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_1-b_2, 1, u}^{3, \{n\}} \right] \\ &= \Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] \Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \mid \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] \\ &\times \Pr \left[ \mathcal{E}_{K-b_1-b_2, 1, u}^{3, \{n\}} \mid \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \right], \end{aligned} \quad (45)$$

where  $\Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] = \sum_{\mathcal{L}_1 \in \mathcal{L}_{a, 1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}}$ ,

$$\Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \mid \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] = \sum_{\mathcal{L}_2 \in \mathcal{L}_{a, 2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}}$$

$\Pr \left[ \mathcal{E}_{K-b_1-b_2, 1, u}^{3, \{n\}} \mid \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \right] = P_n^{K-b_1-b_2}$ . Next, we calculate

$$\Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}}, \mathcal{E}_{K-b_2, 1, u}^{3, \{n\}} \right]$$

for all  $a \in \{0\} \cap \Theta_a(u, n)$ ,  $b_1 = 0$ , and  $b_2 \in \{u-1, \dots, K-a-i-1\}$ . By using results from “balls into bins” problem, we have

$$\Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \cap \mathcal{E}_{K-b_2, 1, u}^{3, \{n\}} \right] = \Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \right] \Pr \left[ \mathcal{E}_{K-b_2, 1, u}^{3, \{n\}} \mid \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \right], \quad (46)$$



where  $\Pr \left[ \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \right] = \sum_{\mathcal{L}_2 \in \mathcal{L}_{0,2}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_2} \in \mathcal{A}_{b_2, \mathcal{L}_2}} \frac{b_2!}{\prod_{n' \in \mathcal{L}_2} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_2} P_{n'}^{\alpha_{n'}}$ , and

$$\Pr \left[ \mathcal{E}_{K-b_2, 1, u}^{3, \{n\}} \mid \mathcal{E}_{b_2, u-a-1, u}^{2, \{n+1, n+2, \dots, N\}} \right] = P_n^{K-b_2}.$$

Finally, we calculate  $\Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{K-b_1, 1, u}^{3, \{n\}} \right]$  for all  $a \in \{u-1\} \cap \Theta_a(u, n)$ ,  $b_2 = 0$  and  $b_1 \in \{u-1, \dots, a+i-1\}$ . By using results from “balls into bins” problem, we have

$$\Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \cap \mathcal{E}_{K-b_1, 1, u}^{3, \{n\}} \right] = \Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] \Pr \left[ \mathcal{E}_{K-b_1, 1, u}^{3, \{n\}} \mid \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right], \quad (47)$$

where  $\Pr \left[ \mathcal{E}_{K-b_1, 1, u}^{3, \{n\}} \mid \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] = P_n^{K-b_1}$ ,

$$\Pr \left[ \mathcal{E}_{b_1, a, u}^{1, \{1, 2, \dots, n-1\}} \right] = \sum_{\mathcal{L}_1 \in \mathcal{L}_{u-1, 1}} \sum_{(\alpha_{n'})_{n' \in \mathcal{L}_1} \in \mathcal{A}_{b_1, \mathcal{L}_1}} \frac{b_1!}{\prod_{n' \in \mathcal{L}_1} \alpha_{n'}!} \prod_{n' \in \mathcal{L}_1} P_{n'}^{\alpha_{n'}}.$$

Substituting (45), (46) and (47) into (44), we can obtain  $P'_{i, u, n}$  given in (12)-(15). Therefore, we complete the proof of Lemma 1.

## APPENDIX B: PROOF OF LEMMA 2

By (5) and (17), it can be easily shown that the constraints in (1), (2) and (3) of Problem 1 can be converted into (18), (19) and (20). Now, we show that the objective function of Problem 1 in (4) can be converted into the objective function of Problem 3. By (5) and (17), we have

$$\begin{aligned} R_{\text{avg}}(K, N, M, \mathbf{x}) &\stackrel{(a)}{=} \sum_{\mathbf{d} \in \mathcal{N}^K} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{s=1}^K \sum_{\mathcal{S} \subseteq \mathcal{K}: |\mathcal{S}|=s} z_{s-1} - \sum_{\mathbf{d} \in \mathcal{N}^K} \left( \prod_{k=1}^K p_{d_k} \right) \sum_{s=1}^K \sum_{\mathcal{S} \subseteq \mathcal{K} \setminus \underline{\mathcal{K}}(\mathbf{d}): |\mathcal{S}|=s} z_{s-1} \\ &= \sum_{s=1}^K \binom{K}{s} z_{s-1} - \sum_{u=1}^{\min\{K, N\}} \Pr[|\underline{\mathcal{K}}(\mathbf{D})| = u] \sum_{s=1}^K \binom{K-u}{s} z_{s-1} \end{aligned} \quad (48)$$

where (a) is due to (5) and (17). Therefore, we complete the proof of Lemma 2.

## APPENDIX C: PROOF OF LEMMA 3

The Lagrangian of Problem 3 is given by  $L(\mathbf{z}, \boldsymbol{\eta}, \theta, \nu) = \sum_{s=0}^{K-1} \binom{K}{s+1} z_s + \sum_{s=0}^K \eta_s (-z_s) - \sum_{u=1}^{\min\{K, N\}} P''_u \sum_{s=0}^{K-u-1} \binom{K-u}{s+1} z_s + \theta \left( \sum_{s=0}^K \binom{K}{s} s z_s - \frac{KM}{N} \right) + \nu \left( 1 - \sum_{s=0}^K \binom{K}{s} z_s \right)$ , where  $\eta_s \geq 0$  is the Lagrange multiplier associated with (18),  $\nu$  is the Lagrange multiplier associated with (19),  $\theta$  is the Lagrange multiplier associated with (20) and  $\boldsymbol{\eta} \triangleq (\eta_s)_{s \in \{0, 1, \dots, K\}}$ . Thus, we have

$$\frac{\partial L}{\partial z_s}(\mathbf{z}, \boldsymbol{\eta}, \theta, \nu) = \binom{K}{s+1} - \sum_{u=1}^{\min\{K, N\}} P''_u \binom{K-u}{s+1} - \eta_s + \theta s \binom{K}{s} - \nu \binom{K}{s}. \quad (49)$$

Since Problem 3 is a linear programming,  $\mathbf{z}^*$  is an optimal solution of Problem 3 if  $\mathbf{z}^*, \boldsymbol{\eta}^*, \nu^*, \theta^*$  satisfy KKT conditions, i.e., (i) primal constraints: (18), (19), (20), (ii) dual constraints: (a)  $\theta \geq 0$  and (b)  $\eta_s \geq 0$  for all  $s \in \{0, 1, \dots, K\}$ , (iii) complementary slackness: (a)  $\eta_s(-z_s) = 0$  for all  $s \in \{0, 1, \dots, K\}$  and (b)  $\theta \left( \sum_{s=0}^K \binom{K}{s} s z_s - \frac{KM}{N} \right) = 0$ , and (iv)  $\frac{\partial L}{\partial z_s}(\mathbf{z}, \boldsymbol{\eta}, \theta, \nu) = 0$  for all  $s \in \{0, 1, \dots, K\}$ . In the following, we obtain  $\mathbf{z}^*, \boldsymbol{\eta}^*, \nu^*$ , and  $\theta^*$  by considering three cases.

**Case 1:** When  $M = 0$ , it can be easily verified that  $\mathbf{z}^*$  given in (21), any  $\nu^* = \sum_{u=1}^{\min\{K, N\}} P_u'' u$ ,  $\eta_s^* = \sum_{u=1}^{\min\{K, N\}} P_u'' (-u \binom{K}{s} - \binom{K-u}{s+1} + \binom{K}{s+1}) + \theta^* s \binom{K}{s}$ ,

$$\theta^* \in \left[ \max \left\{ \max_{s \in \{1, 2, \dots, K\}} \left\{ \sum_{u=1}^{\min\{K, N\}} P_u'' \frac{u \binom{K}{s} + \binom{K-u}{s+1} - \binom{K}{s+1}}{s \binom{K}{s}} \right\}, 0 \right\}, +\infty \right),$$

satisfy the KKT conditions in (i)-(iv). Thus, we know that when  $M = 0$ ,  $\mathbf{z}^*$  given in (21) is an optimal solution of Problem 3.

**Case 2:** When  $M = N$ , it can be easily verified that  $\mathbf{z}^*$  given in (21), any  $\nu^* = K\theta^*$ ,  $\theta^* \in \left[ 0, \min_{s \in \{0, 1, \dots, K-1\}} \left\{ \sum_{u=1}^{\min\{K, N\}} P_u'' \frac{u \binom{K}{s} + \binom{K-u}{s+1} - \binom{K}{s+1}}{s \binom{K}{s}} \right\} \right]$ ,  $\eta_s^* = \binom{K}{s+1} - \sum_{u=1}^{\min\{K, N\}} P_u'' \binom{K-u}{s+1} + \theta^* (s - K) \binom{K}{s}$  satisfy the KKT conditions in (i)-(iv). Thus, we know that when  $M = N$ ,  $\mathbf{z}^*$  given in (21) is an optimal solution of Problem 3.

**Case 3:** When any  $M \in \left\{ \frac{N}{K}, \frac{2N}{K}, \dots, \frac{(K-1)N}{K} \right\}$ , we prove that  $\mathbf{z}^*$  given in (21) is an optimal solution of Problem 3 by proving that  $\mathbf{z}^*$  given in (21),

$$\theta^* = - \sum_{u=1}^{\min\{K, N\}} P_u'' \sum_{k=K-u+1}^K g'_k \left( \frac{KM}{N} \right), \quad (50)$$

$$\nu^* = \sum_{u=1}^{\min\{K, N\}} P_u'' \sum_{k=K-u+1}^K \left( g_k \left( \frac{KM}{N} \right) - g'_k \left( \frac{KM}{N} \right) \frac{KM}{N} \right), \quad (51)$$

$$\eta_s^* = \binom{K}{s} \sum_{u=1}^{\min\{K, N\}} P_u'' \sum_{k=K-u+1}^K \left( g_k(s) - \left( g_k \left( \frac{KM}{N} \right) + g'_k \left( \frac{KM}{N} \right) \left( s - \frac{KM}{N} \right) \right) \right), \quad s \in \{0, 1, \dots, K\}, \quad (52)$$

satisfy the KKT conditions in (i)-(iv), where

$$g_k(s) \triangleq \begin{cases} h_k(s) \triangleq \frac{\prod_{i=0}^{K-k} (K-s-i)}{\prod_{j=0}^{K-k} (K-j)}, & s \in [0, k) \\ 0, & s \in [k, K] \end{cases}, \quad k \in \{K-u+1, \dots, K\}. \quad (53)$$

In the following, we show that  $\mathbf{z}^*$  in (21),  $\theta^*$  in (50),  $\nu^*$  in (51), and  $\boldsymbol{\eta}^*$  in (52) satisfy KKT conditions (i), (ii), (iii), and (iv), respectively.

- Prove that  $\mathbf{z}^*$  satisfies (i). By substituting  $\mathbf{z}^*$  into the primal constraints in (18), (19) and (20), we can easily verify that  $\mathbf{z}^*$  satisfies (i). Thus, we complete proving that  $\mathbf{z}^*$  satisfies (i).
- Prove that  $\theta^*$  satisfies (ii.a) and  $\boldsymbol{\eta}^*$  satisfies (ii.b). First, we show that  $\theta^*$  satisfies (ii.a) by proving  $g'_k\left(\frac{KM}{N}\right) \leq 0$ . Since

$$g'_k(s) = \begin{cases} h'_k(s) = -\frac{\prod_{i=0}^{K-k}(K-s-i)}{\prod_{j=0}^{K-k}(K-j)} \sum_{i=0}^{K-k} \frac{1}{K-s-i}, & s \in (0, k) \\ 0, & s \in [k, K] \end{cases}, \quad (54)$$

we have  $g'_k(s) \leq 0$  for any  $s \in (0, K)$ . Since  $M \in \left\{\frac{N}{K}, \frac{2N}{K}, \dots, \frac{(K-1)N}{K}\right\}$ , we have  $\frac{KM}{N} \in \{1, 2, \dots, K-1\}$ . Therefore, we have  $g'_k\left(\frac{KM}{N}\right) \leq 0$ . Thus, we complete proving that  $\theta^*$  satisfies (ii.a). Next, we show that  $\boldsymbol{\eta}^*$  satisfies (ii.b) by proving

$$g_k(s) - \left( g_k\left(\frac{KM}{N}\right) + g'_k\left(\frac{KM}{N}\right) \left(s - \frac{KM}{N}\right) \right) \geq 0. \quad (55)$$

Consider the following four cases.

- 1) When  $s \in [0, k]$  and  $\frac{KM}{N} \in (0, k)$ , (55) is equivalent to

$$h_k(s) \geq h_k\left(\frac{KM}{N}\right) + h'_k\left(\frac{KM}{N}\right) \left(s - \frac{KM}{N}\right). \quad (56)$$

Since for any convex set  $\mathcal{X}$  and any two points  $x, y \in \mathcal{X}$ ,  $f(y) \geq f(x) + f'(x)(y-x)$  if and only if  $f(x)$  is convex [24], we prove (56) by proving that  $h_k(s)$  is convex over  $s \in [0, k]$ . Since  $h''_k(s) = h_k(s) \left( \left( \sum_{i=0}^{K-k} \frac{1}{K-s-i} \right)^2 - \sum_{i=0}^{K-k} \frac{1}{(K-s-i)^2} \right) \geq 0$  for all  $s \in [0, k]$ , by second-order condition for convexity [24], we know that  $h_k(s)$  is convex over  $s \in [0, k]$ .

- 2) When  $s \in [0, k]$  and  $\frac{KM}{N} \in [k, K]$ , (55) is equivalent to  $h_k(s) \geq 0$ , which holds for all  $s \in [0, k]$ .

- 3) When  $s \in [k, K]$  and  $\frac{KM}{N} \in (0, k)$ , (55) is equivalent to  $\sum_{i=0}^{K-k} \frac{s - \frac{KM}{N}}{K - \frac{KM}{N} - i} \geq 1$ , which always holds since  $\sum_{i=0}^{K-k} \frac{s - \frac{KM}{N}}{K - \frac{KM}{N} - i} \geq \sum_{i=0}^{K-k} \frac{k - \frac{KM}{N}}{K - \frac{KM}{N} - i} = 1 + \sum_{i=0}^{K-k-1} \frac{k - \frac{KM}{N}}{K - \frac{KM}{N} - i} \geq 1$ .

- 4) When  $s \in [k, K]$  and  $\frac{KM}{N} \in [k, K]$ , (55) is equivalent to  $0 \geq 0$ , which always holds.

Thus, we complete proving that  $\boldsymbol{\eta}^*$  satisfies (ii.b).

- Prove that  $\mathbf{z}^*, \boldsymbol{\eta}^*$  satisfies (iii.a) and  $\mathbf{z}^*, \theta^*$  satisfies (iii.b). First, we show that  $\mathbf{z}^*, \boldsymbol{\eta}^*$  satisfies (iii.a). Since  $\eta_{\frac{KM}{N}}^* = 0$  and  $z_s^* = 0$  for all  $s \in \{0, 1, \dots, K\} \setminus \left\{\frac{KM}{N}\right\}$ , we know that

$\eta_s^*(-z_s) = 0$  for all  $s \in \{0, 1, \dots, K\}$ . Thus, we complete proving that  $\mathbf{z}^*, \boldsymbol{\eta}^*$  satisfies (iii.a). Next, we show that  $\mathbf{z}^*, \theta^*$  satisfies (iii.b). Since  $\sum_{s=0}^K \binom{K}{s} s z_s^* = \frac{KM}{N}$ , we have  $\theta^* \left( \sum_{s=0}^K \binom{K}{s} s z_s^* - \frac{KM}{N} \right) = 0$ . Thus, we complete proving that  $\mathbf{z}^*, \theta^*$  satisfies (iii.b).

- Prove that  $\mathbf{z}^*, \boldsymbol{\eta}^*, \nu^*$  and  $\theta^*$  satisfies (iv). For any  $s \in \{0, 1, \dots, K\}$ , we have

$$\frac{\partial L}{\partial z_s}(\mathbf{z}^*, \boldsymbol{\eta}^*, \theta^*, \nu^*) = \sum_{u=1}^{\min\{K, N\}} P_u'' \left( \binom{K}{s+1} - \binom{K-u}{s+1} - \sum_{k=K-u+1}^K g_k(s) \binom{K}{s} \right).$$

By Pascal's identity, i.e.,  $\binom{k+1}{t} = \binom{k}{t} + \binom{k}{t-1}$ , we have  $\binom{K}{s+1} - \binom{K-u}{s+1} = \sum_{k=K-u+1}^K \binom{k-1}{s}$ . Furthermore, for any  $s \in \{0, 1, \dots, K\}$ , we have  $g_k(s) = \frac{\binom{k-1}{s}}{\binom{K}{s}}$ . Therefore, for any  $s \in \{0, 1, \dots, K\}$ , we have  $\frac{\partial L}{\partial z_s}(\mathbf{z}^*, \boldsymbol{\eta}^*, \theta^*, \nu^*) = 0$ . Thus, we complete proving that  $\mathbf{z}^*, \boldsymbol{\eta}^*, \nu^*$  and  $\theta^*$  satisfies (iv).

Combining the above three cases, we know that  $\mathbf{z}^*, \boldsymbol{\eta}^*, \theta^*$  and  $\nu^*$  satisfy the KKT conditions in (i)-(iv). Therefore, we complete the proof of Lemma 3.

#### APPENDIX D: PROOF OF LEMMA 4

Under Conditions 1 and 2, for the simplified problem, the average load, the file partition constraints, and the cache memory constraint are the same as those of Problem 2. It remains to transform the subpacketization level constraint in (23) in terms of the vector  $\mathbf{x}$  to (24) in terms of the vector  $\mathbf{y}$ . First, by Theorem 1 of [13], the subpacketization level constraint in (23) is equivalent to

$$\|\mathbf{x}_n\|_{l_{gst}, \hat{F}} \geq \|\mathbf{x}_n\|_1, \quad n \in \mathcal{N}. \quad (57)$$

By the file partition constraint in (2), we have

$$\|\mathbf{x}_n\|_1 = \sum_{s=0}^K \sum_{\mathcal{S} \in \{\hat{\mathcal{S}} \subseteq \mathcal{K}: |\hat{\mathcal{S}}|=s\}} x_{n,\mathcal{S}} = 1, \quad n \in \mathcal{N}. \quad (58)$$

By (57) and (58), we have

$$\|\mathbf{x}_n\|_{l_{gst}, \hat{F}} \geq 1, \quad n \in \mathcal{N}. \quad (59)$$

Next, under Condition 1, it is clear that  $\mathbf{x}_n = U_n \mathbf{y}$ ,  $n \in \mathcal{N}$ . Thus, by (59), we can obtain (24). Therefore, we complete the proof of Lemma 4.

## APPENDIX E: PROOF OF LEMMA 5

Let  $\mathbf{v}$  be an  $L$ -dimensional vector. Denote  $\mathbf{f}^m(\mathbf{v}) \triangleq (f_j^{m,\mathbf{v}})_{j \in \{1,2,\dots,L\}}$ , where

$$f_{[j]}^{m,\mathbf{v}} = \begin{cases} 1, & j \in \{1, \dots, m\} \\ 0, & \text{otherwise} \end{cases}, \quad m \in \{1, 2, \dots, L\}, \quad (60)$$

and  $[j]$  represents the index of the  $j$ -th largest element in  $\mathbf{v}$ . By [13], we know that  $\mathbf{g}(\mathbf{y}) = U_n^T \mathbf{f}^{\hat{F}}(U_n \mathbf{y})$  is a subgradient of  $\|U_n \mathbf{y}\|_{l_{gst}, \hat{F}}$ . In the following, we first calculate  $\mathbf{f}^{\hat{F}}(U_n \mathbf{y})$ . Let  $y_{n,s[i]}$  denote the  $i$ -th largest element in  $\mathbf{y}_n$ , where  $s[i] \triangleq [i] - 1$ . Since  $U_n \mathbf{y}$  is equivalent to  $\mathbf{x}_n$  under Condition 1, the  $(\sum_{i'=1}^{i-1} \binom{K}{s[i']} + 1)$ -th to the  $(\sum_{i'=1}^i \binom{K}{s[i']})$ -th largest elements in  $U_n \mathbf{y}$  all equal to  $y_{n,s[i]}$ . Thus, the set of the indices of the  $\hat{F}$ -th largest element in  $U_n \mathbf{y}$  is

$$\begin{aligned} \Omega^{\hat{F}, U_n \mathbf{y}} = & \bigcup_{i \in \{1, \dots, I-1\}} \left\{ \sum_{s'=0}^{s[i]-1} \binom{K}{s'} + 1, \dots, \sum_{s'=0}^{s[i]} \binom{K}{s'} \right\} \\ & \bigcup \left\{ \sum_{s'=0}^{s[I]-1} \binom{K}{s'} + 1, \dots, \sum_{s'=0}^{s[I]-1} \binom{K}{s'} + \hat{F} - \sum_{i=1}^{I-1} \binom{K}{s[i]} \right\}. \end{aligned} \quad (61)$$

Thus, we have  $\mathbf{f}^{\hat{F}}(U_n \mathbf{y}) \triangleq (f_j^{\hat{F}, U_n \mathbf{y}})_{j \in \{1,2,\dots,2K\}}$ , where

$$f_j^{\hat{F}, U_n \mathbf{y}} = \begin{cases} 1, & j \in \Omega^{\hat{F}, U_n \mathbf{y}} \\ 0, & \text{otherwise} \end{cases}. \quad (62)$$

Next, we calculate  $U_n^T$ . Let  $U_n^T \triangleq (u_{m,h})_{m \in \{1,2,\dots,(K+1)N\}, h \in \{1,2,\dots,2K\}}$ , and  $l_m \triangleq (m-1) \bmod (K+1)$ . From the definition of  $U_n$ , we know that for any  $m \in \{1, 2, \dots, N(K+1)\}$ ,

$$u_{m,h} = \begin{cases} 1, & m = n, \quad h \in \left\{ \sum_{l=0}^{l_m-1} \binom{K}{l} + 1, \dots, \sum_{l=0}^{l_m} \binom{K}{l} \right\} \\ 0, & \text{otherwise} \end{cases}. \quad (63)$$

By (61), (62) and (63), we can obtain  $\mathbf{g}_n(\mathbf{y}) = U_n^T \mathbf{f}^{\hat{F}}(U_n \mathbf{y})$ , indicating (27). Therefore, we complete the proof of Lemma 5.

## REFERENCES

- [1] S. Jin, Y. Cui, H. Liu, and G. Caire, "Uncoded placement optimization for coded delivery," in *IEEE WiOpt*, May 2018, pp. 1–8.
- [2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016 - 2021," March 2017.

- [3] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *IEEE WCNC*, April 2014, pp. 2300–2305.
- [4] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, July 2016.
- [5] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, Jan 2017.
- [6] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [7] S. Jin, Y. Cui, H. Liu, and G. Caire, "Structural properties of uncoded placement optimization for coded delivery," *CoRR*, vol. abs/1707.07146, 2017.
- [8] A. M. Daniel and W. Yu, "Optimization of heterogeneous coded caching," *CoRR*, vol. abs/1708.04322, 2017.
- [9] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1281–1296, Feb 2018.
- [10] L. Tang and A. Ramamoorthy, "Coded caching with low subpacketization levels," in *IEEE Globecom Workshops*, Dec. 2016, pp. 1–6.
- [11] K. Shanmugam, A. M. Tulino, and A. G. Dimakis, "Coded caching with linear subpacketization is possible using ruzsa-szeméredi graphs," in *IEEE ISIT*, Jun. 2017, pp. 1237–1241.
- [12] M. Cheng, Q. Yan, X. Tang, and J. Jiang, "Coded caching schemes with low rate and subpacketizations," *CoRR*, vol. abs/1708.06650, 2017.
- [13] J.-y. Gotoh, A. Takeda, and K. Tono, "Dc formulations and algorithms for sparse optimization problems," *Math Programming*, pp. 1–36, 2017.
- [14] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, 2009.
- [15] H. A. Le Thi, T. P. Dinh, H. M. Le, and X. T. Vo, "Dc approximation approaches for sparse optimization," *European Journal of Operational Research*, vol. 244, no. 1, pp. 26–46, 2015.
- [16] S. Shalev-Shwartz, N. Srebro, and T. Zhang, "Trading accuracy for sparsity in optimization problems with sparsity constraints," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2807–2832, 2010.
- [17] W. Fu, "Penalized regressions: The bridge versus the lasso," *Journal of Computational and Graphical Statistics*, pp. 397–416, 1998.
- [18] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *ICML*, vol. 98, 1998.
- [19] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.
- [20] T. Pham Dinh and H. A. Le Thi, "Recent advances in dc programming and dca," in *Trans on Computational Intelligence*, 2014, pp. 1–37.
- [21] I. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0," <http://cvxr.com/cvx>, Aug. 2012.
- [22] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, 2016.
- [23] C. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," *CoRR*, vol. abs/1601.05690, 2016.
- [24] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.