Cache-Aided Multi-message Private Information Retrieval

Yang Li, Nan Liu(✉), and Wei Kang

National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
{younglee,nanliu,wkang}@seu.edu.cn

Abstract. We consider the problem of multi-message private information retrieval (MPIR) from N non-colluding and replicated servers when the user is equipped with a cache that holds an uncoded fraction r from each of the K stored messages in the servers. We assume that the servers are unaware of the cache content. We investigate $D_P^*(r)$, which is the optimal download cost normalized by the message size, as a function of K, N, r, P. For a fixed K, N, we develop an inner bound (converse bound) for the $D_P^*(r)$ curve. The inner bound is a piece-wise linear function in r. For the achievability, we propose specific schemes that exploit the cached as private side information to achieve some corner points. We obtain an outer bound (achievability) for any caching ratio by memory-sharing between these corner points. Thus, the outer bound is also a piece-wise linear function in r. The inner and the outer bounds match for the cases where the number of desired messages P is at least half of the number of the overall stored messages K. Furthermore, the bounds match in two specific regimes for the case $\frac{K}{P} > 2$ and $\frac{K}{P} \in \mathbb{N}$: the very high ratio regime, i.e., $r \geq \frac{1}{N+1}$ and the very low ratio regime, i.e., $r \leq \frac{(N-1)P\alpha_1}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right)+(N-1)P\alpha_1}$. Finally, the bounds meet in one specific regime for arbitrarily fixed K, P, N: the very high ratio regime, i.e., $r \geq \frac{1}{N+1}$.

Keywords: Multi-message · Cache · PIR

1 Introduction

With the upgrading of Internet technology and the invention of smart phones, people are using the Internet all the time, leaving more and more traces of each user's preferences. These traces reveal the privacy of users, and when exploited can threaten the safety and well-being of each user. Thus, the protection of user privacy is becoming more and more important.

Along the lines of privacy protection, an important problem called private information retrieval (PIR) has been proposed [2]. It allows each user to retrieve a single message from a database without revealing to the servers that store the database which message the user is interested in. For example, a user interested in a particular stock in the stock market may not want to reveal his/her interest,

H. Gao et al. (Eds.): ChinaCom 2020, LNICST 352, pp. 127–141, 2021.
https://doi.org/10.1007/978-3-030-67720-6_9

worried that this revelation will affect the price of the stock. Or, a user while retrieving some information on a particular disease, do not want to reveal his/her health issues to the outside world.

The private information retrieval (PIR) problem was first considered in [2], where a user wishes to privately download one bit out of a database of K bits. Here, privacy is in the information-theoretic sense, i.e., the queries of the user can not reveal anything about which bit it is interested in. When the user makes queries to one server that stores the database, the only information-theoretic private retrieval way is to download the entire database [2], i.e., all K bits. This incurs a huge download cost and a small download efficiency of $\frac{1}{K}$, as to retrieve 1 bit of desired information, K bits must be downloaded. Hence, it is proposed that the user queries N servers, $N > 1$, that stores the database, without revealing which bit it is interested in to any single server [2]. Reference [3] proposed a private retrieval scheme that can achieve the download efficiency of $\frac{N-1}{N}$, i.e., to download $N-1$ bits of desired information, N total bits must be downloaded from the N servers.

The recent significant work by Sun and Jafar [4] reformulated the PIR problem from an information-theoretic perspective, and rather than retrieving a bit, they considered retrieving a message from a database of K messages, where each message is arbitrarily long. They found that the maximum efficiency, which they termed *capacity*, of PIR retrieval from N servers is $(1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^{K-1}})^{-1}$. It can be seen that the capacity is a monotonically increasing function of the number of servers. The more servers the user retrieves from, the larger the PIR capacity, i.e., by increasing the number of queried servers, the user can effectively reduce the size of the download.

The PIR problem has recently attracted much attention in the information-theoretic community for its capacity characterizations under various constraints and due to limited space, we can not list all of them here. The work that is most related to our paper are references [5] and [1]. Reference [5] considered a cache-aided PIR problem with unknown and uncoded prefetc.hing, i.e., user can prefetc.h Lr number of bits from each message W_k, which is of length L, for all messages, $k = 1, 2, \cdots, K$. Furthermore, the servers are unaware of which bits are stored in the user's cache. The optimal tradeoff between the normalized download cost $\frac{D(r)}{L}$, which is the inverse of capacity, and the caching ratio r was the subject of focus in [5]. The outer and inner bounds are both piece-wise linear curves which consists of K line segments. The inner and the outer bounds meet in general for the cases of very low caching ratio ($r \leq \frac{1}{1+N+N^2+\cdots+N^{K-1}}$) and very high caching ratio ($r \geq \frac{K-2}{(N+1)K+N^2-2N-2}$).

Reference [1] studied the multi-message PIR problem, where the user need to retrieve P messages, $P > 1$, without exposing the identities of the P messages to any one server. Instead of using a single file private information retrieval scheme [4] over and over again for P times, reference [1] showed that the P messages can be retrieved more efficiently by an MDS code.

In this paper, we consider a new PIR scenario, namely cache-aided multi-message PIR (MPIR). In this setup we assume that the user requires multiple

messages and at the same time has access to some cached bits, which is a subset of the bits evenly taken from each message. We propose lower and upper bounds on the normalized download cost, which are both piecewise-linear functions of P, K, N, r. We show that the upper and lower bounds match in the following three cases:

1. $P \geq \frac{K}{2}$, i.e., when the number of requested messages is more than half of the total number of messages.
2. $P < \frac{K}{2}$ and $r \geq \frac{1}{N+1}$, i.e., when the caching ratio is sufficiently large.
3. For very low caching ratio and $P < \frac{K}{2}$ and $\frac{K}{P} \in \mathbb{N}$, i.e., when the number of total messages is an integer multiple of the number of retrieved messages and when the caching ratio is sufficiently small.

The rest of the paper is organized as follows. In Sect. 2 we describe the system model. In Sect. 3 our main results for the download cost of the cache-aided MPIR problem is stated. Sections 4 and 5 provide the achievability and converse proofs, respectively. Finally, Sect. 6 concludes the paper.

2 System Model

We consider a PIR problem with K independent messages $W_1, \ldots, W_K$. The length of each message is L bits, i.e.,

$$H(W_1, \ldots, W_K) = H(W_1) + \cdots + H(W_K) = KL. \tag{1}$$

We use the random variable Z to represent the contents of the user's cache. Before sending out retrieval requests to the servers, the user caches Lr bits of each message in advance, where $0 \leq r \leq 1$, and r is called *caching ratio*. The Lr bits are taken randomly and evenly from each L-length message, Therefore,

$$H(Z) = KLr. \tag{2}$$

We use the random variable $\mathbb{H}$ to represent indices of the cached bits, and we assume that all servers know only the cache ratio r and the prefetching rule of the user, but not the value of $\mathbb{H}$.

After the prefetching phase, the user privately generates an index set $\mathcal{P} = \{i_1, \cdots, i_P\} \subseteq \{1, \cdots, K\}$, independent of $\mathbb{H}$, and wants to retrieve the content of the P messages $W_{\mathcal{P}} = (W_{i_1}, W_{i_2}, \cdots, W_{i_P})$ from the servers, while ensuring that the index set of these messages is not known to any single server. Therefore, we have

$$H(\mathcal{P}, \mathbb{H}, W_1, \ldots, W_K) = H(\mathcal{P}) + H(\mathbb{H}) + H(W_1) + \cdots + H(W_K). \tag{3}$$

To retrieve the contents of the message set of interest, i.e., $W_{\mathcal{P}}$, the user sends out request $Q_n^{[\mathcal{P}]}$ to Server n, $n = 1, 2, \cdots, N$. During the retrieval phase, the user does not know anything about the content of any messages in the

database. Therefore, the request sent by the user and the messages' contents of the database are independent of each other, i.e.,

$$I\left(W_1, \cdots, W_K; Q_1^{[\mathcal{P}]}, \cdots, Q_N^{[\mathcal{P}]}\right) = 0 \tag{4}$$

Upon receipt of the request $Q_n^{[\mathcal{P}]}$, the n-th server returns an answer string $A_n^{[\mathcal{P}]}$ to the user, determined by the files stored in the server and the request received, hence

$$H(A_n^{[\mathcal{P}]}|Q_n^{[\mathcal{P}]}, W_{1:K}) = 0 \tag{5}$$

User privacy requires that each individual server cannot know anything about the message set $\mathcal{P}$ the user wants to download based on the retrieval request sent by the user. Therefore, for any P messages, the retrieval request and the answer are indistinguishable from any individual server, i.e., $\forall n \in [N], \forall \mathcal{P}_1, \mathcal{P}_2 \subseteq \{1, \cdots, K\}$, $|\mathcal{P}_1| = |\mathcal{P}_2| = P$,

$$(Q_n^{[\mathcal{P}_1]}, A_n^{[\mathcal{P}_1]}, W_1, \ldots, W_K) \sim (Q_n^{[\mathcal{P}_2]}, A_n^{[\mathcal{P}_2]}, W_1, \ldots, W_K). \tag{6}$$

Here, "$\sim$" means with the same distribution.

After receiving the replies from all the servers, i.e., $A_1^{[\mathcal{P}]}, \ldots, A_N^{[\mathcal{P}]}$, with the help of the cache bits, the user can reliably extract all the required messages $W_{\mathcal{P}}$. Hence, the following reliability constraint needs to be met:

$$H\left(W_{\mathcal{P}}|Z, \mathbb{H}, Q_1^{[\mathcal{P}]}, \ldots, Q_N^{[\mathcal{P}]}, A_1^{[\mathcal{P}]}, \ldots, A_N^{[\mathcal{P}]}\right) = o(L), \tag{7}$$

where $o(L)$ denotes a function such that $\frac{o(L)}{L} \to 0$ as $L \to \infty$.

For a fixed N, K, P, and caching ratio r, a pair $(D_P(r), PL)$ is achievable if there exists a PIR scheme for message size L with unknown and uncoded prefetching satisfying the privacy constraint (6) and the reliability constraint (7), where $D_P(r)$ represents the expected number of downloaded bits (over all the queries) from the N servers via the answering strings $A_{1:N}^{[\mathcal{P}]}$, i.e.,

$$D_P(r) = \sum_{n=1}^{N} H\left(A_n^{[\mathcal{P}]}\right). \tag{8}$$

In this work, we aim to characterize the optimal normalized download cost $D^*(r)$ corresponding to every caching ratio $0 \le r \le 1$, where

$$D_P^*(r) = \inf\left\{\frac{D_P(r)}{PL} : (D_P(r), PL) \text{ is achievable}\right\}, \tag{9}$$

which is a function of the caching ratio r and the number of interested messages P.

3 Main Results and Discussions

Our first result is the exact characterization of the normalized download cost for the case $P \geq \frac{K}{2}$, i.e., when the user needs to privately retrieve at least half of the messages stored in the servers.

Theorem 1. *For the cache-aided MPIR problem with non-colluding and replicated servers, if the number of desired messages P is at least half of the number of overall stored messages K, i.e., if $P \geq \frac{K}{2}$, then the optimal normalized download cost is given by,*

$$D_P^*(r) = \begin{cases} \left(1 + \frac{K-P}{NP}\right) - \left[1 + \frac{(N+1)(K-P)}{NP}\right] r, & 0 \leq r \leq \frac{1}{1+N} \\ 1 - r, & \frac{1}{1+N} \leq r \leq 1 \end{cases} \tag{10}$$

Remark 1. Our results can be reduced to known results of PIR for replicated servers:

1. When the cache ratio is zero, *i.e.,* r = 0, the above results degenerate into multi-message private information retrieval [1], where the optimal normalized downloads is $1 + \frac{K-P}{NP}$.
2. When the number of required messages is one, *i.e.,* $P = 1$, Theorem 1 becomes single-message cache-aided PIR with unknown and uncoded perfecting [5], where the optimal normalized downloads is shown below.

$$D_1^*(r) = \begin{cases} \left(1 + \frac{K-1}{N}\right) - \left[1 + \frac{(N+1)(K-1)}{N}\right] r, & 0 \leq r \leq \frac{1}{1+N} \\ 1 - r, & \frac{1}{1+N} \leq r \leq 1 \end{cases} \tag{11}$$

Our second main result is that for the case of $P \leq \frac{K}{2}$, we have the following theorem which characterizes the upper and lower bounds on the download cost.

Theorem 2. *For the cache-aided MPIR problem with non-colluding and replicated servers, when $P \leq \frac{K}{2}$, the normalized download cost is lower and upper bounded as,*

$$\underline{D}_P(r) \leq D_P^*(r) \leq \overline{D}_P(r) \tag{12}$$

where the upper bound $\overline{D}_P(r)$ is given by,

$$\overline{D}_P(r) = \max\left\{ \frac{D_2 - D_1}{r_2 - r_1} r + D_1, \frac{D_3 - D_2}{r_3 - r_2}(r - r_3) + D_3, 1 - r \right\} \tag{13}$$

where

$$D_1 = \frac{1 - \left(\frac{1}{N}\right)^{\frac{K}{P}}}{1 - \frac{1}{N}}, \tag{14}$$

$$D_2 = \frac{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k\right)}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1}, \tag{15}$$

$$D_3 = \frac{N}{N+1}, \tag{16}$$

$$r_1 = 0, \tag{17}$$

$$r_2 = \frac{(N-1)P\alpha_1}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1}, \tag{18}$$

$$r_3 = \frac{1}{N+1}. \tag{19}$$

and $\underline{D}(r)$ is given by,

$$\underline{D}_P(r) = \max\left\{\sum_{i=0}^{\lfloor\frac{K}{P}\rfloor-1}\frac{1}{N^i} - r\sum_{i=0}^{\lfloor\frac{K}{P}\rfloor-1}\left(\frac{\lfloor\frac{K}{P}\rfloor - i}{N^i}\right), 1-r\right\} \tag{20}$$

The achievability proofs for Theorems 1 and 2 are given in Sect. 4, and the converse proofs are given in Sect. 5.

Remark 2. Regarding Theorem 2, we have the following remarks:

1. When $\frac{K}{P} \in \mathbb{N}$, the lower bound of Theorem 2 is reduced to the known result of [5] for $\frac{K}{P}$ messages.
2. For all cases of K and P, we have the result

$$D_P^*(r) \geq 1 - r,$$

for $r \geq \frac{1}{1+N}$. Recall that without the privacy constraint, the download cost is $1 - r$, which means that at $r \geq \frac{1}{N+1}$, requiring privacy does not incur any additional download cost, i.e., there is a smart way to retrieve the desired bits without revealing the privacy of the user. This result was shown to be true for the single-message cache-aided PIR problem [5], and in this paper, we show that this is also true for its multi-message counterpart.

Finally, our third main result is that in the case of $P \leq \frac{K}{2}$ *and* $\frac{P}{K} \in \mathbb{N}$ *and*

$$r \leq \frac{(N-1)P\alpha_1}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1} \tag{21}$$

where $\alpha_k = \frac{1}{N-1}\sum_{i=1}^{P}\binom{P}{i}\alpha_{k+i}$ and the initial conditions of the sequence recursion formula are as follows,

$$\alpha_K = (N-1)^{K-P} \tag{22}$$

$$\alpha_{K-1} = \cdots = \alpha_{K-P+1} = 0 \tag{23}$$

the upper and lower bounds of Theorem 2 match and we therefore have the following corollary.

Corollary 1. *For $P \leq \frac{K}{2}$, $\frac{K}{P} \in \mathbb{N}$ and r satisfying (31), the optimal normalized download cost is given by,*

$$D_P^*(r) = \sum_{i=0}^{\frac{K}{P}-1} \frac{1}{N^i}\left[1 - r\left(\frac{K}{P} - i\right)\right] \tag{24}$$

The proof of the corollary will be provided in the Appendix.

4 Achievability Proofs of Theorems 1 and 2

The achievability proof will be based on the original multi-message PIR [1] achievability scheme, except the corresponding first round of single bits download will now be cached by the user in advance.

4.1 Achievability Proof of Theorem 1

Note that in the case of $P \geq \frac{K}{2}$, we just need to show that when $r = \frac{1}{N+1}$, the retrieval rate matches the upper bound. The rate with the non-vertex can be obtained by memory-sharing at adjacent vertices.

The scheme requires $L = N + 1$, and is completed in two phases. In the first phase, the user cached a portion of the same size from each message. The user then downloads the new required message bits using the cached private bits. The details of the scheme are as follows.

1. *Index preparation:* Let $[u_i(1), u_i(2), \ldots, u_i(L)]$ denote a random permutation of the L bits of messages $W_i = [w_i(1), \cdots, w_i(L)]^T$ using a random interleaver $\pi_i(.)$ which is known to the user only, i.e.,

$$u_i(m) = w_i(\pi_i(m)), \quad m \in \{1, \cdots, L\} \tag{25}$$

2. *Phase one:* The user caches $(u_1(1), u_2(1), \cdots, u_K(1))$ from external servers which do not collude with the servers in the retrieval phase, i.e., we have cached the first round of the original MPIR [1] scheme.
3. *Phase two:* The user downloads new desired messages mixed with undesired symbols from the cache.
   (a) The user chooses an MDS generator matrix $G \in \mathbb{F}_q^{P\times K}$, where every P columns from G are linearly independent. This implies that the user can decode all the symbols using P linear combinations.

(b) The user picks uniformly and independently at random the permutation matrices $\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_{N-1}$ of size $K \times K$. These matrices shuffle the order of columns of $\mathbf{G}$ to be independent of $\mathcal{P}$.

(c) Specially, the user could download P linear combination from first server as follows.

Server1
$u_1(2) + u_2(2) + \cdots + u_P(2) + u_{P+1}(1) + u_{P+2}(1) + \cdots + u_K(1)$
$\lambda_1 u_1(2) + \lambda_2 u_2(2) + \cdots + \lambda_P u_P(2) + \lambda_{P+1} u_{P+1}(1) + \cdots + \lambda_K u_K(1)$
$\vdots$
$\lambda_1^{P-1} u_1(2) + \lambda_2^{P-1} u_2(2) + \cdots + \lambda_P^{P-1} u_P(2) + \lambda_{P+1}^{P-1} u_{P+1}(1) + \cdots + \lambda_K^{P-1} u_K(1)$
$Z = (u_1(1), u_2(1), \ldots, u_K(1))$

where $\lambda_i, \quad i \in \{1, 2, \cdots, K\}$ are non-zero constants that are not equal to each other.

(d) Similarly, the user could download P linear combination from other servers.

4.2 Decodability, Privacy, and Calculation of the Achievable Rate

Next, we verify that the above achievability scheme satisfies cache size constraint, decodability and privacy.

For the cache size constraint: Above scheme cache a bit from each $L = N+1$ bits message, and the cache ratio is $r = \frac{1}{N+1}$, meeting the cache size constraint.

For the reliability: The user can subtract out all the undesired message symbols using the undesired symbols from cached bits. Consequently, the user is left with a $P \times P$ system of equations which is guaranteed to be invertible by the MDS property, hence all symbols that belong to W_P are decodable.

For the privacy: Since the permutation matrices are chosen uniformly and independently from each other, the probability distribution is uniform irrespective to $\mathcal{P}$. Furthermore, the symbols are chosen randomly and uniformly by applying the random interleaver. Hence, the retrieval scheme is private.

To calculate the achievable rate: The user exploits the side information from the cached bits to generate P equations for each side information set. Each set of P equations in turn generates P desired symbols. Hence, the achievable normalized download cost is calculated as,

$$\underline{D}_P(r_3) = \frac{\text{total downloaded equations}}{\text{total number of desired symbols}} \tag{26}$$

$$= \frac{NP}{(N+1)P} \tag{27}$$

$$= \frac{N}{N+1} \tag{28}$$

Therefore, $\underline{D}_P(r_3) = \frac{N}{N+1}$, where $r_3 = \frac{1}{N+1}$. The achievability scheme for non-corner points can be obtained by memory-sharing between the most adjacent interesting caching ratios. Since both $\overline{D}_P(r)$ and $\underline{D}_P(r)$ are piecewise linear

function of r, and since $\underline{D}_P(r_3) = \overline{D}_P(r_3)$, $\underline{D}_P(r_1) = \overline{D}_P(r_1)$, and $\underline{D}_P(1) = \overline{D}_P(1)$, we have $\underline{D}_P(r) = \overline{D}_P(r) = D_P^*(r)$ for $r_1 \leq r \leq 1$. Thus we complete the proof of Theorem 1.

4.3 Achievability Proof of Theorem 2

The achievability scheme we propose is the following: consider the scheme proposed in [1] where in round k, the user downloads sums of k terms from different symbols from the database. In our problem, in the prefetching phase, the user caches the bits, which are downloaded in all the stages in round one for every server in the scheme of [1]. In the retrieval phase, the user downloads the same bits as in the other rounds of the scheme proposed in [1]. Obviously, the reliability and the privacy is guaranteed. We only need to verify that when r satisfies (31), the retrieval rate matches the lower bound.

1. *Index preparation:* Let $[u_i(1), u_i(2), \ldots, u_i(L)]$ denote a random permutation of the L bits of messages $W_i = [w_i(1), \cdots, w_i(L)]^T$ using a random interleaver $\pi_i(.)$ which is known to the user only, i.e.,

$$u_i(m) = w_i(\pi_i(m)), \quad m \in \{1, \cdots, L\} \tag{29}$$

2. *Number of stages:* The number of stages needed in each round can be obtained by the following difference equation,

$$y[n] = \frac{1}{N-1} \sum_{i=1}^{P} \binom{P}{i} y[n-i] \tag{30}$$

   where the initial conditions of the equation is $y[-P] = (N-1)^{M-P}$, $y[-P+1] = \cdots = y[-1] = 0$. The number of stages in round k is $\alpha_i = y[(K-P)-k]$.
3. *Caching:* The user cache all the stages needed in round one for every server studied by [1].
4. *Retrieval:* The user downloads sums of k terms from different symbols as in the scheme of [1].

4.4 Calculation of the Achievable Rate

First, we verify that the above scheme satisfies cache size constraint. The user cache all the stages needed in round one for every server studied by [1]. In this case, the cache ratio is r_2, which satisfies cache size constraint.

$$r_2 = \frac{(N-1)P\alpha_1}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1} \tag{31}$$

In round k, the user downloads the sums of k symbols. The user repeats this round for α_k stages. Each stage contains all the possible combinations of any k symbols. There are $\binom{K}{k}$ such combinations.

$$\overline{D}_P(r_2) = \frac{\text{total downloaded equations}}{\text{total number of desired symbols}} \tag{32}$$

$$= \frac{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k\right)}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1} \tag{33}$$

$$= \frac{1}{(1-\frac{1}{N})U}\sum_{i=1}^{P}\gamma_i r_i^{K-P}\left[(N-1)(N^{\frac{K}{P}}-1) - \frac{PK(N-1)}{\gamma_i}\right] \tag{34}$$

where $U = N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1$, $\alpha_k = \sum_{i=1}^{P}\gamma_i r_i^{K-P-k}$, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_P)^T$ is the solution to the system of linear equations,

$$\begin{bmatrix} r_1^{-P} & r_2^{-P} & \cdots & r_P^{-P} \\ r_1^{-P+1} & r_2^{-P+1} & \cdots & r_P^{-P+1} \\ \vdots & \vdots & \cdots & \vdots \\ r_1^{-1} & r_2^{-1} & \cdots & r_P^{-1} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_P \end{bmatrix} = \begin{bmatrix} (N-1)^{M-P} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{35}$$

5 Converse Proofs of Theorems 1 and 2

The converse proof will consists of two major steps:

1) the *initialization* step where we connect the normalized download cost $D_P^*(r)$ to $H\left(A_{1:N}^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right)$ for some P-size set $\mathcal{P}$ where $\mathcal{P} \subset \{1, 2, \cdots, K\}$, and

2) the *induction* step where we connect $H\left(A_{1:N}^{[\mathcal{P}_i]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}, \mathbb{H}, Z\right)$ to $H\left(A_{1:N}^{[\mathcal{P}_0]}|W_{[\mathcal{P}'\bigcup\mathcal{P}_0]}, Q_{1:N}^{[\mathcal{P}_0]}, \mathbb{H}, Z\right)$ where $\mathcal{P}' = \bigcup_{i=1}^{I}\mathcal{P}_i$, and $\mathcal{P}_i$, $i = 0, 1, \cdots, I$ are all P-size subsets of $\{1, 2, \cdots, K\}$ with no intersections.

The Initialization Step

For any private retrieval scheme with normalized download cost $D_p(r)$, for any P-sized subset of $\{1, 2, \cdots, K\}$, denoted as $\mathcal{P}$, we have

$$D_P(r) = \sum_{n=1}^{N} H\left(A_n^{[\mathcal{P}]}\right) \tag{36}$$

$$\geq H\left(A_{1:N}^{[\mathcal{P}]}|Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) \tag{37}$$

$$= H\left(A_{1:N}^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) - H\left(W_{\mathcal{P}}|A_{1:N}^{[\mathcal{P}]}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) \tag{38}$$

$$+ H\left(W_{\mathcal{P}}|Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right)$$

$$= H\left(A_{1:N}^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) - o(L) + H\left(W_{\mathcal{P}}|Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) \tag{39}$$

$$= H\left(A_{1:N}^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) - o(L) + H\left(W_{\mathcal{P}}|Z\right) \tag{40}$$

$$= H\left(A_{1:N}^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) - o(L) + PL(1-r)$$

where (36) follows from (8), (37) follows from the chain rule and conditioning reduces entropy, (39) follows from (7), and (40) follows from the uncoded caching scheme and the independence of the messages, the queries and index of the bits prefetched, i.e., (3) and (4).

The Induction Step

Suppose $\mathcal{P}_i$, $i = 0, 1, \cdots, I$ are all P-size subsets of $\{1, 2, \cdots, K\}$ with no intersections. Let $\mathcal{P}' = \bigcup_{i=1}^{I} \mathcal{P}_i$.

$$H\left(A_{1:N}^{[\mathcal{P}_i]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}, \mathbb{H}, Z\right)$$

$$= H\left(A_{1:N}^{[\mathcal{P}_i]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}\right) - H\left(\mathbb{H}, Z|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}\right) + H\left(\mathbb{H}, Z|A_{1:N}^{[\mathcal{P}_i]}, W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}\right)$$

$$\geq H\left(A_{1:N}^{[\mathcal{P}_i]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}\right) - (K - IP)Lr \tag{41}$$

$$\geq \frac{1}{N}\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}_i]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_i]}\right) - (K - IP)Lr \tag{42}$$

$$= \frac{1}{N}\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}_i]}|W_{[\mathcal{P}']}, Q_{n}^{[\mathcal{P}_i]}\right) - (K - IP)Lr \tag{43}$$

$$= \frac{1}{N}\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}_0]}|W_{[\mathcal{P}']}, Q_{n}^{[\mathcal{P}_0]}\right) - (K - IP)Lr \tag{44}$$

$$\geq \frac{1}{N} H\left(A_{1:N}^{[\mathcal{P}_0]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_0]}\right) - (K - IP)Lr \tag{45}$$

$$\geq \frac{1}{N} H\left(A_{1:N}^{[\mathcal{P}_0]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_0]}, \mathbb{H}, Z\right) - (K - IP)Lr \tag{46}$$

$$= \frac{1}{N}\left[H\left(A_{1:N}^{[\mathcal{P}_0]}|W_{[\mathcal{P}'\cup\mathcal{P}_0]}, Q_{1:N}^{[\mathcal{P}_0]}, \mathbb{H}, Z\right) + H\left(W_{[\mathcal{P}_0]}|W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_0]}, \mathbb{H}, Z\right)\right] \tag{47}$$

$$- \frac{1}{N} H\left(W_{[\mathcal{P}_0]}|A_{1:N}^{[\mathcal{P}_0]}, W_{[\mathcal{P}']}, Q_{1:N}^{[\mathcal{P}_0]}, \mathbb{H}, Z\right) - (K - IP)Lr \tag{48}$$

$$= \frac{1}{N} H\left(A_{1:N}^{[\mathcal{P}_0]}|W_{[\mathcal{P}'\cup\mathcal{P}_0]}, Q_{1:N}^{[\mathcal{P}_0]}, \mathbb{H}, Z\right) + \frac{1}{N} PL(1-r) + \frac{o(L)}{N} - (K - IP)Lr$$

where (41) follows from the uncoded nature of the cache and the nonnegativity of the entropy function, where (43) follows from the fact $A_n^{[\mathcal{P}]} \leftrightarrow (Q_n^{[\mathcal{P}]}, W_{\mathcal{P}}) \leftrightarrow Q_k^{[\mathcal{P}]} (k \neq n)$ is a Markov chain, where (44) follows from privacy constraint (6),

where (45) follows from conditioning reduces entropy, where (46) since conditioning reduces entropy, (49) follows from (3)(4)(7).

5.1 Converse Proof for the Case of $P \geq \frac{K}{2}$

Note that in the case of $P \geq \frac{K}{2}$, there can not exists 2 non-intersecting sets $\mathcal{P}_0$ and $\mathcal{P}_1$ that are subsets of $\{1, 2, \cdots, K\}$. Hence, we take the induction step result of (49) for $I = 1$, and perform a union with the complement of set $\mathcal{P}$, denoted as $\bar{\mathcal{P}}$.

$$H\left(A_{1:N}^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}]}, \mathbb{H}, Z\right) \tag{49}$$

$$\geq \frac{1}{N}\left[H\left(A_{1:N}^{[\mathcal{P}^*]}|W_{1:K}, Q_{1:N}^{[\mathcal{P}^*]}, \mathbb{H}, Z\right) + H\left(W_{\bar{\mathcal{P}}}|W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}^*]}, \mathbb{H}, Z\right)\right] \tag{50}$$

$$- \frac{1}{N}H\left(W_{\mathcal{P}^*}|A_{1:N}^{[\mathcal{P}^*]}, W_{\mathcal{P}}, Q_{1:N}^{[\mathcal{P}^*]}, \mathbb{H}, Z\right) - (K-P)Lr$$

$$= \frac{1}{N}(K-P)L(1-r) - (K-P)Lr + \frac{o(L)}{N} \tag{51}$$

where (50) follows the step of induction, where $W_{\bar{\mathcal{P}}}$ corresponds to the complement set of messages of $W_{\mathcal{P}}$, $\bar{\mathcal{P}} \subseteq \mathcal{P}^*$, $|\mathcal{P}^*| = P$, where (51) follows from (5)(7).

Combining (40) and (51), we conclude the converse proof by dividing by PL and taking the limit as $L \to \infty$, the normalized download cost is lower bounded by maximum value

$$D_P^*(r) \geq \max\{1-r, \left(1 + \frac{K-P}{NP}\right) - \left[1 + \frac{(N+1)(K-P)}{NP}\right]r\} \tag{52}$$

5.2 Converse Proof for the Case of $P < \frac{K}{2}$

Now, we derive the inductive relation for $\frac{K}{P} > 2$. Suppose $\mathcal{P}_i$, $i = 0, 1, \cdots, I$ are all P-size subsets of $\{1, 2, \cdots, K\}$ with no intersections. Let $\mathcal{P}' = \bigcup_{i=1}^{I} \mathcal{P}_i$. Applying induction formula (49) repeatedly.

$$H\left(A_{1:N}^{[\mathcal{P}_1]}|W_{[\mathcal{P}_1]}, Q_{1:N}^{[\mathcal{P}_1]}, \mathbb{H}, Z\right) \tag{53}$$

$$= \frac{1}{N}H\left(A_{1:N}^{[\mathcal{P}_2]}|W_{[\mathcal{P}_1 \bigcup \mathcal{P}_2]}, Q_{1:N}^{[\mathcal{P}_2]}, \mathbb{H}, Z\right) + \frac{1}{N}PL(1-r) + \frac{o(L)}{N} - (K-P)Lr \tag{54}$$

$$\geq \frac{1}{N}\left[\frac{1}{N}H\left(A_{1:N}^{[\mathcal{P}_3]}|W_{[\mathcal{P}_1 \bigcup \mathcal{P}_2 \bigcup \mathcal{P}_3]}, Q_{1:N}^{[\mathcal{P}_3]}, \mathbb{H}, Z\right) + \frac{1}{N}PL(1-r) + \frac{o(L)}{N}\right] \tag{55}$$

$$-\frac{1}{N}(K-2P)Lr+\frac{1}{N}PL(1-r)+\frac{o(L)}{N}-(K-P)Lr \tag{56}$$

$$\geq \dots \tag{57}$$

$$\geq PL(1-r)\sum_{i=1}^{\lfloor \frac{K}{P} \rfloor -1}\frac{1}{N^i}+L(1-r)\left(K-\left\lfloor \frac{K}{P} \right\rfloor P\right)\frac{1}{N^{\lfloor \frac{K}{P} \rfloor}}$$
$$-Lr\sum_{i=0}^{\lfloor \frac{K}{P} \rfloor -1}\left(\frac{K-(i+1)P}{N^i}\right)+o(L)(\frac{1}{N}+\frac{1}{N^2}+\dots+\frac{1}{N^{\lfloor \frac{K}{P} \rfloor -1}}) \tag{58}$$

Combining (58) and ***the initialization step***, we have

$$D_P(r)+o(L)-PL(1-r) \tag{59}$$

$$\geq PL(1-r)\sum_{i=1}^{\lfloor \frac{K}{P} \rfloor -1}\frac{1}{N^i}+L(1-r)\left(K-\left\lfloor \frac{K}{P} \right\rfloor P\right)\frac{1}{N^{\lfloor \frac{K}{P} \rfloor}}$$
$$-Lr\sum_{i=0}^{\lfloor \frac{K}{P} \rfloor -1}\left(\frac{K-(i+1)P}{N^i}\right)+o(L)(\frac{1}{N}+\frac{1}{N^2}+\dots+\frac{1}{N^{\lfloor \frac{K}{P} \rfloor -1}}) \tag{60}$$

Consequently, the normalized download cost is lower bounded by

$$D_P(r)\geq \sum_{i=0}^{\lfloor \frac{K}{P} \rfloor -1}\frac{1}{N^i}+(1-r)\left(\frac{K}{P}-\left\lfloor \frac{K}{P} \right\rfloor\right)\frac{1}{N^{\lfloor \frac{K}{P} \rfloor}}-r\sum_{i=0}^{\lfloor \frac{K}{P} \rfloor -1}\left(\frac{\frac{K}{P}-i}{N^i}\right)+o(L) \tag{61}$$

We conclude the converse proof by dividing by PL and taking the limit as $L\to\infty$, the normalized download cost is lower bounded by maximum value

$$D^*(r)\geq \max_{i\in\{2,\cdots,\lfloor \frac{K}{P} \rfloor +1\}}(1-r)\sum_{j=0}^{\lfloor \frac{K}{P} \rfloor +1-i}\frac{1}{N^j}-r\sum_{j=0}^{\lfloor \frac{K}{P} \rfloor -i}\frac{\lfloor \frac{K}{P} \rfloor +1-i-j}{N^j} \tag{62}$$
$$+(1-r)\left(\frac{K}{P}-\left\lfloor \frac{K}{P} \right\rfloor\right)\frac{1}{N^{\lfloor \frac{K}{P} \rfloor}} \tag{63}$$

We note that when $M=\frac{K}{P}\in\mathbb{N}$, the result in (62) reduces to the known result of [5], which is

$$D^*(r)\geq \max_{i\in\{2,\cdots,M+1\}}(1-r)\sum_{j=0}^{M+1-i}\frac{1}{N^j}-r\sum_{j=0}^{M-i}\frac{M+1-i-j}{N^j} \tag{64}$$

6 Conclusion

In this paper, we developed the cache-aided MPIR problem from N non-colluding and replicated servers, when the cache stores uncoded bits that are unknown to

the servers. The problem generalizes the cache-aided PIR problem in [5] which retrieves a single message privately and the multi-message PIR scenario with no cache available at the user [1]. We determined inner and outer bounds for the optimal normalized download cost $D_P^*(r)$ as a function of the number of required message P, the total number of messages K, the number of servers N, and the caching ratio r. Both inner and outer bounds are piece-wise linear functions in r (for fixed P, N, K) that consist of two line segments. We determined the exact download cost for this problem when the number of desired messages is at least half of the number of total stored messages. Furthermore, the bounds match in two specific regimes for the case $\frac{K}{P} > 2$: the very high ratio regime,i.e., $r \geq \frac{1}{N+1}$ and the very low ratio regime,i.e., r satisfying (31).

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China under Grants 62071115, 61971135, National Key Research and Development Project under Grant 2019YFE0123600, the Research Fund of National Mobile Communications Research Laboratory, Southeast University (No. 2020A03), and the Six talent peaks project in Jiangsu Province.

A Proof of Corollary 1

Since both $\overline{D}_P(r)$ and $\underline{D}_P(r)$ are piecewise linear function of r, all we have to do to prove that they are equal at some interval is to prove that the vertices are the same. First, let us note that

$$r_1 = 0 \tag{65}$$

$$r_2 = \frac{(N-1)P\alpha_1}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1} \tag{66}$$

Then, we note that

$$\overline{D}_P(r_2) = \frac{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k\right)}{N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1} \tag{67}$$

$$= \frac{1}{(1-\frac{1}{N})U}\sum_{i=1}^{P}\gamma_i r_i^{K-P}\left[(N-1)(N^{\frac{K}{P}}-1) - \frac{PK(N-1)}{\gamma_i}\right] \tag{68}$$

where $U = N\left(\sum_{k=2}^{K}\binom{K}{k}\alpha_k - \sum_{k=2}^{K-P}\binom{K-P}{k}\alpha_k\right) + (N-1)P\alpha_1$, $\alpha_k = \sum_{i=1}^{P}\gamma_i r_i^{K-P-k}$.

Further, we note from (20), by choosing $r = r_2$, we have

$$\underline{D}_P(r_2) \geq \sum_{i=0}^{\frac{K}{P}-1} \frac{1}{N^i}\left[1 - r_2\left(\frac{K}{P} - i\right)\right] \tag{69}$$

$$= \frac{1-(\frac{1}{N})^{\frac{K}{P}}}{1-\frac{1}{N}} - \frac{(N-1)P\alpha_1}{U}\left[\frac{\frac{K}{P} - \frac{1}{N}(1-(\frac{1}{N})^{\frac{K}{P}})}{(1-\frac{1}{N})^2}\right] \tag{70}$$

$$= \frac{1}{(1-\frac{1}{N})U}\left[\left(1-\frac{1}{N^{\frac{K}{P}}}\right)U - (N-1)P\alpha_1\left(\frac{\frac{K}{P} - \frac{1}{N}(1-(\frac{1}{N})^{\frac{K}{P}})}{(1-\frac{1}{N})}\right)\right] \tag{71}$$

$$= \frac{1}{(1-\frac{1}{N})U}\left[(N-1)P\alpha_1\left(\left(1+\frac{1}{N-1}\right)\left(1-\frac{1}{N^{\frac{K}{P}}}\right) - \frac{K}{P}\right)\right] \tag{72}$$

$$+ \frac{1}{(1-\frac{1}{N})U}\left[N\left(1-\frac{1}{N^{\frac{K}{P}}}\right)\sum_{i=1}^{P}\gamma_i r_i^{K-P}\left(N^{\frac{K}{P}} - N^{\frac{K}{P}-1} - \frac{P}{\gamma_i}\right)\right] \tag{73}$$

$$= \frac{1}{(1-\frac{1}{N})U}\sum_{i=1}^{P}\gamma_i r_i^{K-P}\left[(N-1)(N^{\frac{K}{P}}-1) - \frac{PK(N-1)}{\gamma_i}\right] \tag{74}$$

$$= \overline{D}_P(r_2) \tag{75}$$

Thus, since $\underline{D}_P(r_2) \leq \overline{D}_P(r_2)$ by definition, (75) implies $\underline{D}_P(r_2) = \overline{D}_P(r_2)$. We also note that $\underline{D}_P(r_1) = \overline{D}_P(r_1)$. Since both $\overline{D}_P(r)$ and $\underline{D}_P(r)$ are piecewise linear function of r, and since $\underline{D}_P(r_2) = \overline{D}_P(r_2)$ and $\underline{D}_P(r_1) = \overline{D}_P(r_1)$, we have $\underline{D}_P(r) = \overline{D}_P(r) = D_P^*(r)$ for $r_1 \leq r \leq r_2$. Thus we complete the proof of corollary 1.

References

1. Banawan, K., Ulukus, S.: Multi-message private information retrieval: capacity results and near-optimal schemes. IEEE Trans. Inform. Theory **64**(10), 6842–6862 (2018). https://doi.org/10.1109/TIT.2018.2828310
2. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: Proceedings of IEEE 36th Annual Foundations of Computer Science, pp. 41–50, October 1995. https://doi.org/10.1109/SFCS.1995.492461
3. Shah, N.B., Rashmi, K.V., Ramchandran, K.: One extra bit of download ensures perfectly private information retrieval. In: 2014 IEEE International Symposium on Information Theory, pp. 856–860, June 2014. https://doi.org/10.1109/ISIT.2014.6874954
4. Sun, H., Jafar, S.A.: The capacity of private information retrieval. IEEE Trans. Inform. Theory **63**(7), 4075–4088 (2017). https://doi.org/10.1109/TIT.2017.2689028
5. Wei, Y., Banawan, K.A., Ulukus, S.: Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. CoRR abs/1709.01056 (2017). http://arxiv.org/abs/1709.01056