New Results on the Storage-Retrieval Tradeoff in Private Information Retrieval Systems

Tao Guo, *Member, IEEE*, Ruida Zhou, *Student Member, IEEE*, and Chao Tian, *Senior Member, IEEE*

***Abstract*—In a private information retrieval (PIR) system, the user needs to retrieve one of the possible messages from a set of storage servers, but wishes to keep the identity of the requested message private from any given server. Existing efforts in this area have made it clear that the efficiency of the retrieval will be impacted significantly by the amount of the storage space allowed at the servers. In this work, we consider the tradeoff between the storage cost and the retrieval cost. We first present three fundamental results: 1) a regime-wise approximate characterization of the optimal tradeoff with a factor of two, 2) a cyclic permutation lemma that can produce more sophisticated codes from simpler base codes, and 3) a relaxed entropic linear program (LP) lower bound that has a polynomial complexity. Equipped with the cyclic permutation lemma, we then propose two novel code constructions, and by applying the lemma, obtain new storage-retrieval points. Furthermore, we derive more explicit lower bounds by utilizing only a subset of the constraints in the relaxed entropic LP in a systematic manner. Though the new upper bound and lower bound do not lead to a better approximation factor uniformly, they are significantly tighter than the existing art in some regimes.**

***Index Terms*—Private information retrieval, storage cost, retrieval cost, tradeoff, linear program lower bound**

I. Introduction

The analysis of private information retrieval (PIR) systems from the information-theoretic perspective has drawn significant attention recently [1]–[35]. The canonical model, where the K messages are replicated over all the N servers, was studied extensively and well-understood. Particularly, the capacity of the canonical PIR system was characterized recently by Sun and Jafar [4], and a more efficient code construction was presented in [5] that achieves the optimal average performance.

Full replication of the messages at the storage servers can be costly, and the messages can be stored more efficiently by utilizing better storage codes. However, the amount of storage allowed at the servers will impact the efficiency of the retrieval. At one extreme, when the messages are replicated across all the servers, the retrieval can be made the most efficient; on the other hand, when no storage redundancy is allowed, the only possible strategy is to retrieve every message and thus highly inefficient.

The work of R.-D Zhou and C. Tian was supported in part by the National Science Foundation under Grants CCF-1816546 and CCF-2007067.

T. Guo was with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA; he is now with the Department of Electrical and Computer Engineering, the University of California, Los Angeles, CA, USA. (e-mail: guotao@ucla.edu)

R.-D Zhou and C. Tian are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. (e-mail: ruida@tamu.edu, chao.tian@tamu.edu)

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. Contact guotao@ucla.edu for further questions about this work.

There has been increasing interest in understanding the storage-retrieval tradeoff in PIR systems. Banawan and Ulukus [6] considered the case when each message is encoded by a maximum distance separable (MDS) code and stored across the servers, referred to as the MDS-PIR code, and characterized the capacity of this system. Sun and Tian presented two sets of codes where the messages are MDS-coded that can beat the separate MDS-PIR capacity by using joint storage coding for certain specific parameters [10]. Attia et al. considered the case when the storage servers can only store uncoded segments of the messages [7], [8], and derived the full storage-retrieval tradeoff in such systems. A generalized code construction unifying the two codes was presented more recently in [9]. Mathematically, we use α to denote the normalized average storage per server per message bit, and β for the normalized average download cost per server by message bit (the precise definitions are given in Section II). In this context, the MDS-PIR code in [6] achieve the following tradeoff points

$$(\alpha,\beta)=\left(\frac{K}{T},\frac{1}{N}\left(\sum_{i=0}^{K-1}\left(\frac{T}{N}\right)^i\right)\right),\ T=1,2,\cdots,N, \quad (1)$$

the uncoded storage PIR code [7], [8] achieves the following tradeoff points

$$(\alpha,\beta)=\left(\frac{KT}{N},\frac{1}{N}\left(\sum_{i=0}^{K-1}\frac{1}{T^i}\right)\right),\ T=1,2,\cdots,N, \quad (2)$$

and the unified code in [9] achieves

$$(\alpha,\beta)=\left(\frac{KT_2}{NT_1},\frac{1}{N}\sum_{i=0}^{K-1}\left(\frac{T_1}{T_2}\right)^i\right), \quad T_1,T_2\in\{1,2,\cdots,N\},T_1\le T_2. \quad (3)$$

Though significant progress has been made in the cases where structural restrictions are placed on the storage codes, our understanding on the fundamental tradeoff between the storage cost and the retrieval cost is quite limited when these restrictions are removed. In fact, even for the smallest case with two servers and two messages, this tradeoff is not known. A Shannon-theoretic approach [11] was used on this special case to improve the storage and download efficiency, and very specialized lower bounds were also given. Two general lower bounds were further given in [13] which focus on the two extreme points of the tradeoff curve.

In this work, we studied the tradeoff between the storage cost and the retrieval cost in PIR systems without any structural storage restrictions. Three fundamental results are presented,

1) A regime-wise 2-approximate characterization of the optimal tradeoff: The overall tradeoff can be partitioned into three regimes, where 2-approximation holds for the storage cost, the retrieval cost, or the sum-cost.
2) A cyclic permutation lemma that can produce more sophisticated codes from simpler ones: This is a general technique, and it can be shown that uncoded storage PIR code [7], [8] can be obtained directly from the code in [4] with this lemma, and the generalized MDS-PIR code [9] can be obtained from that in [6].
3) A relaxed entropic linear program (LP) lower bound that has a polynomial complexity: The generic entropic LP frame work [36]–[38] may be used to compute lower bounds in this problem, which however has exponential numbers of variables and constraints. By utilizing the specific structure in the PIR problem, we select a subset of these inequalities and formulate a simpler LP that is more amicable for computation.

With these results, we further seek to find improved upper bounds and lower bounds. We propose two novel code constructions, and by applying the cyclic permutation lemma, obtain a set of new storage-retrieval points. Then we derive a close-form lower bound by utilizing only a subset of the constraints in the relaxed entropic LP in a systematic manner. As a byproduct, we in fact obtain a set of lower bounds parametrized by a set of real values. Though the new upper bound and lower bound do not lead to a more precise approximate characterization in general, they are significantly tighter than the existing art.

The rest of the paper is organized as follows. We formally define the problem in Section II. The three fundamental results on the optimal tradeoff are presented in Section III. Section IV is mostly devoted to two new code constructions. A lower bound for the optimal tradeoff is then presented in Section V, with some numerical results. We conclude the paper in Section VI. Some technical proofs are given in the supplementary material.

II. Problem Formulation

We adopt the notation $[i:j] \triangleq \{i, i+1, \ldots, j\}$ when $i \leq j$, and define it to be $\emptyset$ if $i > j$; the brackets will be omitted when appeared in subscripts. An (N, K) *private information retrieval* (PIR) system can be described as follows. A total of K mutually independent equal-length messages $W_{1:K} = (W_1, W_2, \cdots, W_K)$ are coded and stored in N servers; the stored content at server n is denoted as $S_n \in \mathcal{S}_n$, where $\mathcal{S}_n$ is the domain of the content at server n. Let $\mathcal{Q}_n$ be the set of all possible queries sent to server n. When retrieving message W_k, the user sends a query $Q_n^{[k]} \in \mathcal{Q}_n$ to server n, from which an answer $A_n^{[k]}$ will be returned. After collecting the answers $A_{1:N}^{[k]}$ from all the servers, the user will recover the desired message W_k. The privacy requirement stipulates that any single server cannot derive any knowledge on the identity of the requested message based on the received query. In this work, we aim to study the tradeoff between the size of storage contents $S_{1:N}$ and that of the answers $A_{1:N}$.

Mathematically, a PIR system can almost be fully represented using information measures of the involved random variables alone. Each message W_k ($k \in [1:K]$) is comprised of L i.i.d. symbols uniformly distributed over a finite alphabet $\mathcal{X}$. In $\log_{|\mathcal{X}|}$-ary units, this is equivalent to

$$H(W_{1:K}) = \sum_{k=1}^{K} H(W_k), \tag{4}$$

$$H(W_k) = L, \ k \in [1:K]. \tag{5}$$

There are a total of N servers, and each can store coded or uncoded contents of the messages, which is equivalent to the condition that the stored content $S_n \in \mathcal{S}_n$ at server n satisfies

$$H(S_n|W_{1:K}) = 0, \ n \in [1:N]. \tag{6}$$

A user aims to retrieve a message W_k, $k \in [1:K]$ from the N severs without revealing the identity k to any individual server. A random key $\mathbf{F}$ is used to generate queries $Q_{1:N}^{[k]} = (Q_1^{[k]}, Q_2^{[k]}, \ldots, Q_N^{[k]})$, where $Q_n^{[k]} \in \mathcal{Q}_n$ for $n \in [1:N]$, which can be represented as

$$H(Q_1^{[k]}, Q_2^{[k]}, \ldots, Q_N^{[k]}|\mathbf{F}) = 0, \quad k = 1, 2, \ldots, K. \tag{7}$$

The random key is independent of messages, i.e.,

$$I(\mathbf{F}; W_{1:K}) = 0. \tag{8}$$

Server n uses the stored content S_n and the query $Q_n^{[k]}$ to construct an answer $A_n^{[k]}$, and then sends the answer to the user, which is represented by the relation

$$H(A_n^{[k]}|Q_n^{[k]}, S_n) = 0, \ n \in [1:N], \ k \in [1:K]. \tag{9}$$

The answer symbols are in a finite alphabet $\mathcal{Y}$, i.e., $A_n^{[k]} \in \mathcal{Y}^{\ell_n}$, where ℓ_n is the length of the answer. With the answers from all servers $A_{1:N}^{[k]}$, together with queries $Q_{1:N}^{[k]}$ and the identity of the desired message k, the user can recover the desired message W_k, i.e.,

$$H(W_k|A_{1:N}^{[k]}, Q_{1:N}^{[k]}) = 0. \tag{10}$$

The privacy requirement is more suitable to be represented using probability distribution relations, instead of information measures[1], i.e., for any $q \in \mathcal{Q}_n$

$$\Pr(Q_n^{[k]} = q) = \Pr(Q_n^{[k']} = q), \text{ for any } k \neq k' \in [1:K]. \tag{11}$$

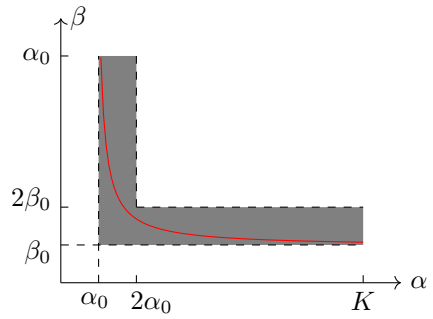
The *operational* normalized average storage cost and the *operational* normalized average download cost are defined as

$$\bar{\alpha} \triangleq \frac{1}{NL} \sum_{n=1}^{N} \log_{|\mathcal{X}|} |\mathcal{S}_n|, \tag{12}$$

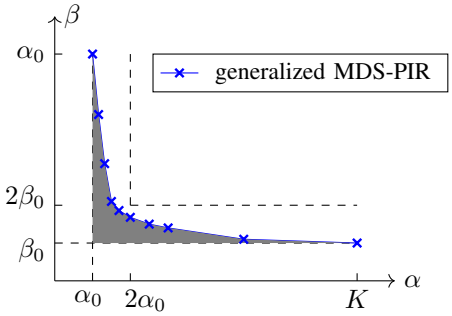
$$\bar{\beta} \triangleq \frac{\log_{|\mathcal{X}|} |\mathcal{Y}|}{NL} \sum_{n=1}^{N} \mathbb{E}(\ell_n), \tag{13}$$

which are the average amount of stored data per symbol of individual message and the expected amount of average

[1]Strictly speaking, it is possible to represent the privacy condition by introducing another random variable θ to represent the (random) index of the requested message, assuming the probability distribution of θ is known. The privacy requirement as represented by the probability distribution relations is more general, in the sense that there is no need to require the knowledge of the probability distribution of θ.



(a) Simple approximation



(b) Approximation vs. the generalized MDS-PIR bound

Fig. 1: Lower bound, upper bounds and simple approximation of the optimal tradeoff.

downloaded data per symbol of desired message, respectively. In the sequel, we shall simply refer to them as the storage cost and download cost, respectively. Note that $\bar{\beta}$ does not depend on the value of k, since the random variable ℓ_n has an identical distribution for all $k \in [1:K]$ due to the privacy requirement.

The definition $\bar{\beta}$ measures the retrieval download rate in terms of the expected retrieval download rate, which is in fact the prevailing performance measure taken in the information theoretic literature; see e.g., the problem definition in the pioneer work [4]. An alternative definition is to measure the worst case download cost among all possible random key realizations; see e.g., [39]. This two approaches are inherently related, and it was shown in [5] that any PIR code with a given expected case rate can be converted to a code that has the same worst case rate, at the expense of an increased message size; we therefore refer readers to that work for more details.

We say the storage-retrieval tradeoff point (α, β) is achievable, if there exists a PIR code whose operational storage cost $\bar{\alpha}$ and download cost $\bar{\beta}$ satisfy $\alpha \geq \bar{\alpha}$ and $\beta \geq \bar{\beta}$, respectively. The aim of this work is to characterize the set of all achievable pairs (α, β), or in other words, the optimal tradeoff between α and β. It is clear that

$$\bar{\alpha} \geq \frac{1}{NL} \sum_{n=1}^{N} H(S_n) \tag{14}$$

$$\bar{\beta} \geq \frac{1}{NL} \sum_{n=1}^{N} H(A_{1:N}^{[k]} | Q_{1:N}^{[k]}), \tag{15}$$

the right hand sides of which are referred to as the *informational* normalized storage cost and *informational* normalized download cost, respectively. We shall use the informational costs as surrogates for the operational costs in the rest of this work in order to derive meaningful lower bounds. A detailed discussion of these two definitions and their differences can be found in [13].

For some fixed download cost β, let $\alpha_{\min}(\beta)$ denote the minimum achievable storage cost for the download cost β, and $\beta_{\min}(\alpha)$ is defined similarly. It was established in [4] that

$$\beta_{\min}(\infty) = \beta_0 \triangleq \frac{1}{N} + \frac{1}{N^2} + \ldots + \frac{1}{N^K}, \tag{16}$$

and it is trivial to see

$$\alpha_{\min}(\infty) = \alpha_0 \triangleq \frac{K}{N}, \tag{17}$$

in order for the system to allow correct message retrieval. In fact the result in [4] implies that $\beta_{\min}(K) = \beta_0$, and it is not difficult to verify $\alpha_{\min}(\alpha_0) = \alpha_0$.

III. Three Fundamental Results

We first present three results that are not difficult from a technical point of view, but are of significant fundamental or instrumental importance. The first is a simple approximate characterization of the optimal (α, β) tradeoff, the second is a simple lemma which uses cyclic permutation to build more sophisticated codes from simpler ones, and the last is an extracted (low-complexity) linear programming lower bound that captures the most important constraints in the problem setting.

A. A Simple Approximate Characterization

The following proposition provides a simple approximate characterization of the achievable storage-retrieval tradeoff.

Proposition 1 (Regime-wise 2-approximation)**.** *For any (N, K) PIR system where $N \geq 2$,*

(i) The tradeoff point $(2\alpha_0, 2\beta_0)$ is achievable;

(ii) Conversely, any achievable (α, β) must satisfy $\alpha \geq \alpha_0$ and $\beta \geq \beta_0$.

The combination of the upper bound given in (i) and the lower bound given in (ii) provide an approximate characterization as shown in Fig. 1(a). In Fig. 1(b), we further include the upper bounds induced by the generalized MDS-PIR code [9] to illustrate this approximate characterization.

Proof of Proposition 1. The lower bounds in (ii) follow simply from the definition of α_0 and β_0, and thus we only need to prove the upper bounds in (i). This can be done by showing that the point $(2\alpha_0, 2\beta_0)$ is above the tradeoff curve achieved by the uncoded storage PIR code given in [7], [8], for which the storage cost and download cost are given by the lower convex envelope of the following points

$$(\bar{\alpha}, \bar{\beta}) = \left(\frac{KT}{N}, \frac{1}{N} \left(\sum_{i=0}^{K-1} \frac{1}{T^i} \right) \right), \; T = 1, 2, \cdots, N. \tag{18}$$

Taking $T = 2$, we obtain

$$(\bar{\alpha}, \bar{\beta}) = \left(2\alpha_0, \frac{1}{N} \cdot \frac{1 - \left(\frac{1}{2}\right)^K}{1 - \frac{1}{2}} \right). \tag{19}$$

For $N \geq 2$, it is seen that the

$$\bar{\beta} = \frac{1}{N} \frac{1 - \left(\frac{1}{2}\right)^K}{1 - \frac{1}{2}} < \frac{2}{N} \cdot \frac{1 - \left(\frac{1}{N}\right)^K}{1 - \frac{1}{N}} = 2\beta_0, \tag{20}$$

which completes the proof. □

Remark 1. It is also possible to utilize the upper bound induced by the MDS-PIR code [6] to prove this proposition, which we omit for brevity.

This approximate characterization shows that the storage-retrieval tradeoff can be divided into three regimes: a storage-bound regime, a retrieval-bound regime, and an intermediate regime. In the storage-bound regime $\beta \geq 2\beta_0$, the optimal storage cost is sandwiched between α_0 and $2\alpha_0$ for any fixed β; in the retrieval-bound regime $\alpha \geq 2\alpha_0$, the optimal retrieval cost is sandwiched between β_0 and $2\beta_0$ for any fixed α; in the intermediate regime, where $\alpha \leq 2\alpha_0$ and $\beta \leq 2\beta_0$, the optimal sum rate $\alpha+\beta$ is sandwiched between $\alpha_0+\beta_0$ and $2\alpha_0+2\beta_0$. Thus in the first regime, the potential loss of using the uncoded PIR code (or the MDS-PIR code) in terms of the storage cost is less than a multiplicative factor of 2, while in the second, the potential loss of using either of these two codes in terms of the retrieval cost is less than a factor of 2. In the intermediate regime, the sum-rate loss is less than a factor of 2 using these codes. This result makes clear what questions remain difficult: to find good approximate (or exact) characterization of the retrieval cost in the storage-bound regime, that on the storage cost in the retrieval-bound regime, and either direction in the intermediate regime. In [13], these questions were considered for the extreme cases when $\alpha = \alpha_0$ and $\beta = \beta_0$, and a precise characterization was given for the former and an approximate one for the latter. However, beyond those two extreme cases, the answers to these questions remain elusive. In the sections to follow, we shall provide further results toward answering these questions.

B. A Cyclic Permutation Lemma

We next introduce a general technique to produce more sophisticated codes from simpler codes, and present several immediate applications of this lemma. In Section IV we shall further utilize this technique to produce other achievable (α, β) tradeoff points based on several new code constructions.

Lemma 1 (Cyclic permutation lemma)**.** *If an (N, K) PIR code (without any symmetry assumption) can achieve the tradeoff point (α, β), then there exists an (M, K) PIR code, $M \geq N$, that achieves the tradeoff point $\left(\frac{N}{M}\alpha, \frac{N}{M}\beta\right)$.*

Proof. We prove the lemma by generating an (M, K) PIR code from an (N, K) PIR code using round-robin, which is illustrated in Fig. 2. Let the message length in the (N, K) PIR (base) code be L, and in the (M, K) code to be constructed, the message length will be ML. Therefore, for the (M, K) PIR problem, we can partition each message W_k into M sub-messages $W_k^1, W_k^2, \cdots, W_k^M$, each has a message length L. For each $m \in [1 : M]$, the sub-messages $W_1^m, W_2^m, \cdots, W_K^m$ can be encoded by the original (N, K) code, and placed on a set of N consecutive servers, i.e., the base (N, K) PIR code is utilized on the servers in a round-robin manner. More precisely, for $m = 1, 2, \ldots, M$, the K sub-messages $W_1^m, W_2^m, \cdots, W_K^m$ are encoded using the (N, K) PIR storage code as $S_1^m, S_2^m, \cdots, S_N^m$; for notation simplicity, further define $S_n^m = \emptyset$ for $n \in [N+1, N+2, \cdots, M]$. Then server m stores the encoded messages $S^1_{(m+1)_N}, S^2_{(m+2)_N}, \cdots, S^M_{(m+M)_N}$ for $m = 1, 2, \ldots, M$, where $(x)_N$ is defined for any integer x as

$$(x)_N = \begin{cases} x \mod N, & \text{if } x \mod N \neq 0 \\ N, & \text{if } x \mod N = 0 \end{cases}. \tag{21}$$

The retrieval is done on each group of sub-messages, $(W_1^m, W_2^m, \ldots, W_K^m)$, which are stored on the corresponding server set that they are stored, for $m = 1, 2, \ldots, M$. Since for each group of sub-messages, the retrieval is private by the property of the base (N, K) PIR code, the overall retrieval is also private. Since the message lengths of the base code and the new code are L and ML, respectively, the resulting storage and download cost of the new code are $\alpha' = \frac{MN\alpha L}{MML} = \frac{N}{M}\alpha$ and $\beta' = \frac{MN\beta L}{MML} = \frac{N}{M}\beta$. The proof is complete. □

Remark 2. The lemma in fact also holds for the T-colluding (coded) PIR problem [29], [40]–[42] and the symmetric PIR problem [21]. Moreover, when the download cost is measured in terms of the worst case rate among all random key realizations [39], the lemma applies in a similar manner, i.e., we will obtain a new code with the storage cost and worst case download cost scaled by the same factor.

We can apply the cyclic permutation lemma on any existing codes, e.g., the Sun-Jafar code [4], the TSC code [5], the MDS-PIR code [6], and the uncoded storage PIR code [7], [8]. In fact, the performance of the uncoded storage PIR code in [7], [8] and the generalized MDS-PIR coded in [9] can be obtained in this way from the code [4] and [6], respectively, as we shall show next.

Application 1. *The (M, K) uncoded storage PIR code in [7], [8] can be produced from the Sun-Jafar code [4] using the cyclic permutation lemma. The storage cost and download cost of (N, K) Sun-Jafar code is*

$$(\alpha, \beta) = \left(K, \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^K} \right). \tag{22}$$

By applying Lemma 1 to an (N, K) Sun-Jafar code, the corresponding storage download cost can be obtained as

$$(\alpha', \beta') = \frac{N}{M}(\alpha, \beta) = \left(\frac{NK}{M}, \frac{1}{M} \sum_{i=0}^{K-1} \frac{1}{N^i} \right). \tag{23}$$

For different storage requirement, we can choose different Sun-Jafar base code by varying N. By taking $N = 1, 2, \cdots, M$, the storage-retrieval tradeoff of the uncoded storage PIR code in (18) *is obtained. We remark that the base code can be any other PIR capacity-achieving code, e.g., the TSC code [5], which can yield the same performance.*

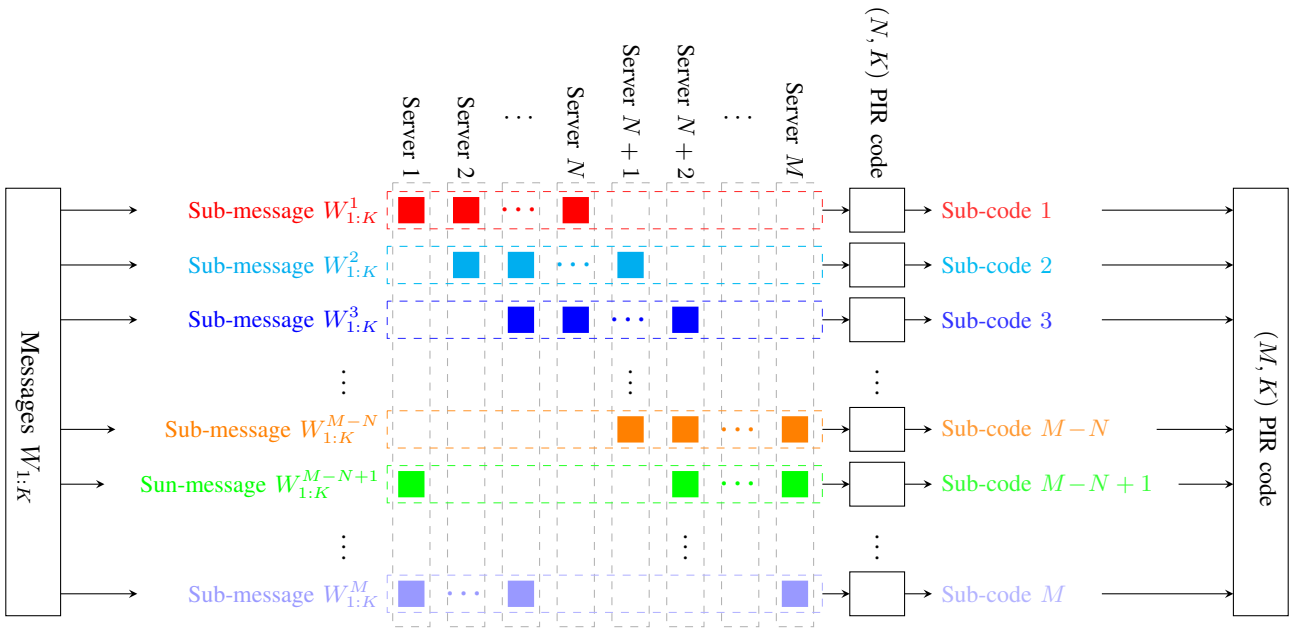


Fig. 2: Generation of (M, K) PIR code from (N, K) PIR code using round-robin.

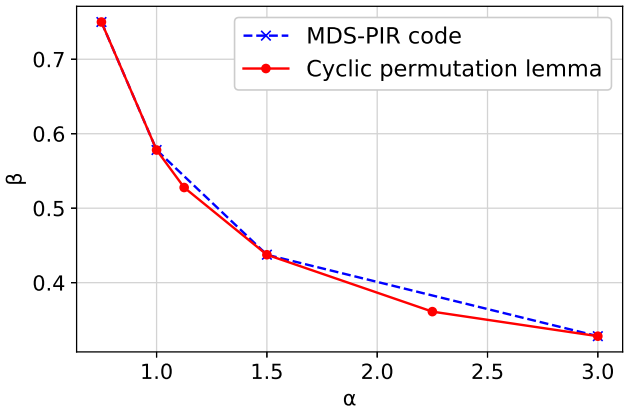


Fig. 3: MDS-PIR code upper bound vs. that obtained by cyclic permutation lemma.

Application 2. *The (M, K) generalized MDS-PIR code in [9] can be produced using the cyclic permutation lemma from the MDS-PIR codes in [6]. The average storage and download cost of MDS-PIR code with parameters (N, K) is given in* (1) *as*

$$(\alpha, \beta) = \left(\frac{K}{T}, \frac{1}{N} \left(\sum_{i=0}^{K-1} \frac{T^i}{N^i} \right) \right), \ T = 1, 2, \cdots, N. \quad (24)$$

By applying Lemma 1 to an (N, K) MDS-PIR code, the corresponding storage download cost can be obtained as

$$(\alpha', \beta') = \frac{N}{M}(\alpha, \beta) = \left(\frac{NK}{MT}, \frac{1}{M} \sum_{i=0}^{K-1} \frac{T^i}{N^i} \right), \quad T = 1, 2, \cdots, N. \quad (25)$$

By letting $N = 1, 2, \cdots, M$, we obtain the storage-retrieval tradeoff of the generalized MDS-PIR code [9] given in (3). We illustrate this storage-retrieval points for the codes obtained by applying the technique when $N = K = 3$ and $M = 4$ in Fig. 3, where it is seen that indeed new tradeoff points are obtained beyond those achieved by the MDS-PIR base code.

C. A Linear Programming Lower Bound

Fully characterizing the storage-retrieval tradeoff appears difficult, partly due to the lack of strong lower bounds, though some initial effort was reported in [11], [13], [33]. This problem can potentially be solved computationally in the generic entropic linear programming (LP) framework [36], similar to the approach discussed in [37], [38]. This generic approach however suffers from high complexity that is exponential in the number of random variables in the problem, since the variables in this generic entropic LP are the joint entropy values of all the possible subsets of these random variables. On the other hand, the PIR problem in fact has a very special structure, which can be well captured by a small class of inequalities. In the following, we use one special class of inequalities to formulate a relaxed linear program.

As a first step, we shall utilize the symmetry structure in this problem. As shown in [5], any PIR code can be symmetrized without sacrificing the storage and download cost to satisfy two symmetry relations: message symmetry and server symmetry. Let $\mathcal{P}_m$ be the set of all permutations of $[1 : m]$. A symmetrized PIR code satisfies the following condition for any $\mathcal{A}, \mathcal{B} \subset [1 : N]$ and $\mathcal{C} \subset [1 : K]$, $k \in [1 : K]$, and any $\pi \in \mathcal{P}_N$ and $\pi' \in \mathcal{P}_K$,

$$H(S_{\mathcal{A}}, A^{[k]}_{\mathcal{B}} | \mathbf{F}, W_{\mathcal{C}}) = H(S_{\pi(\mathcal{A})}, A^{[\pi'(k)]}_{\pi(\mathcal{B})} | \mathbf{F}, W_{\pi'(\mathcal{C})}). \quad (26)$$

In the sequel, we consider only such symmetrized codes without loss of optimality.

$$\text{minimize:} \quad a_0 y_1(1,0) + b_0 y_1(0,1) \tag{27}$$

$$\text{subject to:} \quad \text{(Submodular)} \quad x_k(|\mathcal{A}_1|,|\mathcal{B}_1|) + x_k(|\mathcal{A}_2|,|\mathcal{B}_2|) \geq x_k(|\mathcal{A}_1 \cup \mathcal{A}_2|, |\mathcal{B}_1 \cup \mathcal{B}_2 / (\mathcal{A}_1 \cup \mathcal{A}_2)|) + x_k(|\mathcal{A}_1 \cap \mathcal{A}_2|, |\mathcal{B}_1 \cap \mathcal{B}_2| + |\mathcal{A}_1 \cap \mathcal{B}_2| + |\mathcal{A}_2 \cap \mathcal{B}_1|), \quad \forall k \in [1:K-1], \forall \mathcal{A}_i, \mathcal{B}_i \in [1:N], \mathcal{A}_i \cap \mathcal{B}_i = \emptyset, i = 1,2 \tag{28}$$

$$y_k(|\mathcal{A}_1|,|\mathcal{B}_1|) + y_k(|\mathcal{A}_2|,|\mathcal{B}_2|) \geq y_k(|\mathcal{A}_1 \cup \mathcal{A}_2|, |\mathcal{B}_1 \cup \mathcal{B}_2 / (\mathcal{A}_1 \cup \mathcal{A}_2)|) + y_k(|\mathcal{A}_1 \cap \mathcal{A}_2|, |\mathcal{B}_1 \cap \mathcal{B}_2| + |\mathcal{A}_1 \cap \mathcal{B}_2| + |\mathcal{A}_2 \cap \mathcal{B}_1|), \quad \forall k \in [1:K], \forall \mathcal{A}_i, \mathcal{B}_i \in [1:N], \mathcal{A}_i \cap \mathcal{B}_i = \emptyset, i = 1,2 \tag{29}$$

$$\text{(Monotone)} \quad x_k(a,b) \geq x_k(a,b-1), \quad \forall k \in [1:K-1], \forall a \in [0:N-1], \forall b \in [1:N-a] \tag{30}$$

$$y_k(a,b) \geq y_k(a,b-1), \quad \forall k \in [1:K], \forall a \in [0:N-1], \forall b \in [1:N-a] \tag{31}$$

$$\text{(Decodable)} \quad y_k(a, N-a) \geq 1 + x_k(a, N-a), \ \forall k \in [1:K], \forall a \in [0:N-1] \tag{32}$$

$$\text{(Han's inequality)} \quad y_k(0,b) \geq \frac{b}{N} y_k(0,N), \quad \forall k \in [1:K], \forall b \in [1:N-1] \tag{33}$$

$$\text{(Privacy)} \quad x_k(a,1) = y_{k+1}(a,1), \quad \forall k \in [1:K-1], \forall a \in [0:N-1] \tag{34}$$

$$\text{(Invariance)} \quad x_k(a,0) = y_{k+1}(a,0), \quad \forall k \in [1:K-1], \forall a \in [1:N] \tag{35}$$

$$\text{(Boundary)} \quad x_K(a,b) = 0, \quad \forall a \in [0:N], \forall b \in [0:N-a]. \tag{36}$$

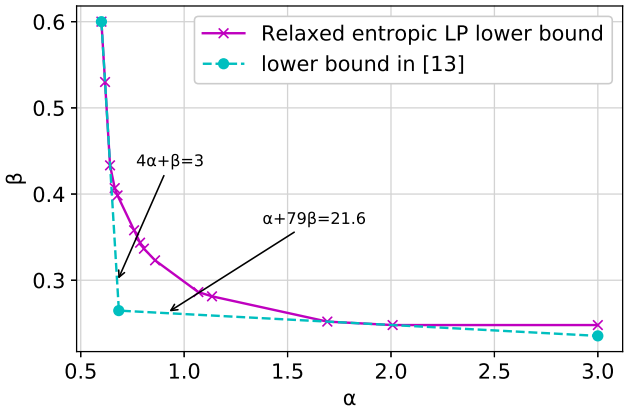


Fig. 4: Comparison of the lower bounds in Proposition 2 and [13] for $(N,K) = (5,3)$.

For any nonnegative integers a and b such that $a+b \leq N$, let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of $[1:N]$ with $|\mathcal{A}| = a$ and $|\mathcal{B}| = b$, we define (for any symmetrized code)

$$x_k(a,b) \triangleq H(S_{\mathcal{A}}, A_{\mathcal{B}}^{[k]} | \mathbf{F}, W_{1:k})/L$$
$$y_k(a,b) \triangleq H(S_{\mathcal{A}}, A_{\mathcal{B}}^{[k]} | \mathbf{F}, W_{1:k-1})/L.$$

It is straightforward to see that by definition

$$y_1(1,0) \geq \alpha$$
$$y_1(0,1) \geq \beta.$$

Thus in order to lower bound a linear combination of $a_0\alpha + b_0\beta$ where $a_0, b_0 \geq 0$, we can consider the following linear program, which we summarize as a proposition.

Proposition 2 (Relaxed entropic LP)**.** *For any achievable (α,β), the linear combination $a_0\alpha + b_0\beta$ where $a_0, b_0 \geq 0$ is lower-bounded by the solution of the linear program in* (27)-(36)*.*

In Fig. 4, we illustrate a set of bounds obtained by solving this linear program for the case $(N,K) = (5,3)$, which are considerably tighter than known bounds in the literature. The capacity bound $\beta \geq \beta_0 = 0.248$ is shown as a horizontal bound, which is indeed obtained through solving the relaxed entropic LP. The two constraints given in [13] are also obtained through the relaxed entropic LP. This is not surprising, since the insights used to formulate the relaxed entropic LP are partly motivated by the proof steps used there.

Proof. The constraints in this LP need to hold for any valid (symmetrized) PIR code for the reasons given below.

- Submodular: for any disjoint $\mathcal{A}_1, \mathcal{B}_1 \subset [1:N]$, disjoint $\mathcal{A}_2, \mathcal{B}_2 \subset [1:N]$ and any $k \in [0:K-1]$, $k' \in [1:K]$, by the submodular property of the entropy function, we have

$$\begin{aligned} &H(S_{\mathcal{A}_1}, A_{\mathcal{B}_1}^{[k']} | \mathbf{F}, W_{1:k}) + H(S_{\mathcal{A}_2}, A_{\mathcal{B}_2}^{[k']} | \mathbf{F}, W_{1:k}) \\ &\geq H((S_{\mathcal{A}_1}, A_{\mathcal{B}_1}^{[k']}) \cup (S_{\mathcal{A}_2}, A_{\mathcal{B}_2}^{[k']}) | \mathbf{F}, W_{1:k}) \\ &\quad + H((S_{\mathcal{A}_1}, A_{\mathcal{B}_1}^{[k']}) \cap (S_{\mathcal{A}_2}, A_{\mathcal{B}_2}^{[k']}) | \mathbf{F}, W_{1:k}) \\ &= H(S_{\mathcal{A}_1 \cup \mathcal{A}_2}, A_{\mathcal{B}_1 \cup \mathcal{B}_2 / (\mathcal{A}_1 \cup \mathcal{A}_2)}^{[k']} | \mathbf{F}, W_{1:k}) \\ &\quad + H(S_{\mathcal{A}_1 \cap \mathcal{A}_2}, A_{(\mathcal{B}_1 \cap \mathcal{B}_2) \cup (\mathcal{A}_1 \cap \mathcal{B}_2) \cup (\mathcal{A}_2 \cap \mathcal{B}_1)}^{[k']} | \mathbf{F}, W_{1:k}). \end{aligned} \tag{37}$$

  The constraints on $y_k(a,b)$ hold for the similar reason.
- Monotone: for any disjoint $\mathcal{A}, \mathcal{B} \subset [1:N]$ with $\mathcal{B}$ non-empty, let $\mathcal{B}' \subset \mathcal{B}$ with $|\mathcal{B}'| = |\mathcal{B}| - 1$, and for any $k \in [0:K-1]$ we have

$$H(S_{\mathcal{A}}, A_{\mathcal{B}}^{[k]} | \mathbf{F}, W_{1:k}) \geq H(S_{\mathcal{A}}, A_{\mathcal{B}'}^{[k]} | \mathbf{F}, W_{1:k}), \tag{38}$$

then the given linear constraints on $x_k(a,b)$ hold due to the symmetry relation mentioned earlier. The constraints on $y_k(a,b)$ hold for the same reason.

- Decodable: for any $\mathcal{A} \subset [1:N]$, for any $k \in [1:K]$, we have

$$\begin{aligned} &H(S_{\mathcal{A}}, A_{\mathcal{A}^c}^{[k]}|\mathbf{F}, W_{1:k-1}) \\ &= H(W_k, S_{\mathcal{A}}, A_{\mathcal{A}^c}^{[k]}|\mathbf{F}, W_{1:k-1}) \\ &= H(W_k) + H(S_{\mathcal{A}}, A_{\mathcal{A}^c}^{[k]}|\mathbf{F}, W_{1:k}), \end{aligned} \tag{39}$$

  where $\mathcal{A}^c = [1:N] \setminus \mathcal{A}$.
- Han's inequality: for any $b \in [1:N-1]$ and $k \in [1:K]$, by the conditional version of the Han's inequality, we have

$$\begin{aligned} &\frac{1}{b\binom{N}{b}} \sum_{\mathcal{B}\subset[1:N]:|\mathcal{B}|=b} H(A_{\mathcal{B}}^{[k]}|\mathbf{F}, W_{1:k-1}) \\ &\quad \geq \frac{1}{N} H(A_{[1:N]}^{[k]}|\mathbf{F}, W_{1:k-1}). \end{aligned} \tag{40}$$

- Privacy: due to the Markov string $\mathbf{F} \leftrightarrow Q_n^{[k]} \leftrightarrow (A_n^{[k]}, W_{1:K}, S_{1:N})$, and the fact that $W_{1:K}$ is independent of $\mathbf{F}$, we have that for any $\mathcal{A} \subset [1:N]$ with $|\mathcal{A}| < N$ and any $n \in \mathcal{A}^c$,

$$\begin{aligned} H(S_{\mathcal{A}}, A_n^{[k]}|\mathbf{F}, W_{1:k}) &= H(S_{\mathcal{A}}, A_n^{[k]}|Q_n^{[k]}, W_{1:k}) \\ H(S_{\mathcal{A}}, A_n^{[k+1]}|\mathbf{F}, W_{1:k}) &= H(S_{\mathcal{A}}, A_n^{[k+1]}|Q_n^{[k+1]}, W_{1:k}). \end{aligned} \tag{41}$$

  By the privacy requirement, the distributions of $(Q_n^{[k]}, A_n^{[k]}, S_{\mathcal{A}}, W_{1:k})$ and $(Q_n^{[k+1]}, A_n^{[k+1]}, S_{\mathcal{A}}, W_{1:k})$ are identical, and it follows that

$$H(S_{\mathcal{A}}, A_n^{[k]}|\mathbf{F}, W_{1:k}) = H(S_{\mathcal{A}}, A_n^{[k+1]}|\mathbf{F}, W_{1:k}). \tag{42}$$

- Invariance: for any $\mathcal{A} \subset [1:N]$ with $|\mathcal{A}| = a$ and $k \in [0:K-1]$ we have

$$x_k(a,0) = H(S_{\mathcal{A}}|W_{1:k})/H(W_1) = y_{k+1}(a,0). \tag{43}$$

- Boundary: for any disjoint subsets $\mathcal{A}, \mathcal{B} \subset [1:N]$, we have

$$H(S_{\mathcal{A}}, A_{\mathcal{B}}^{[K]}|\mathbf{F}, W_{1:N}) = 0. \tag{44}$$

Since any valid symmetrized PIR code must satisfy these constraints, the optimal solution to this LP is indeed a lower bound for $a_0\alpha + b_0\beta$. □

Let us now consider the complexity of this LP. The variables in this LP are all the $x_k(a,b)$'s and $y_k(a,b)$'s for $k = 1, 2, \ldots, K$, and for integers a, b such that $a + b \leq N$, and it is straightforward to verify that there are a total of $K(N+1)(N+2)$ of them. It is more involved to count the total number of constraints. However, it is clear that the dominant component here is the submodular constraints, and thus let us focus on this set. Let $a, b, c, d, e, f, g, h \in [0:N]$ be

$$\begin{aligned} &a = |\mathcal{A}_1 \cap \mathcal{A}_2|; \quad b = |\mathcal{A}_1 \cap \mathcal{B}_2|; \quad c = |\mathcal{A}_1/(\mathcal{A}_2 \cup \mathcal{B}_2)| \\ &d = |\mathcal{A}_2 \cap \mathcal{B}_1|; \quad e = |\mathcal{B}_1 \cap \mathcal{B}_2|; \quad f = |\mathcal{B}_1/(\mathcal{A}_2 \cup \mathcal{B}_2)| \\ &g = |\mathcal{A}_2/(\mathcal{A}_1 \cup \mathcal{B}_1)|; \quad h = |\mathcal{B}_2/(\mathcal{A}_1 \cup \mathcal{B}_1)|, \end{aligned} \tag{45}$$

then the submodular inequalities can be rewritten as

$$\begin{aligned} &x_k(a+b+c, d+e+f) + x_k(a+g+d, b+e+h) \\ &\geq x_k(a+b+c+d+g, e+f+h) + x_k(a, b+d+e) \end{aligned} \tag{46}$$

$$\begin{aligned} &y_k(a+b+c, d+e+f) + y_k(a+g+d, b+e+h) \\ &\geq y_k(a+b+c+d+g, e+f+h) + y_k(a, b+d+e). \end{aligned} \tag{47}$$

Since each of these 8 parameters only takes values in $[0:N]$, the total number of (a,b,c,d,e,f,g,h) combinations is upper-bounded by $(N+1)^8$. Moreover, the combinations (a,b,c,d,e,f,g,h) and $(a,d,g,b,\ e,h,c,f)$ in fact represent the same inequality. Thus, there are fewer than $K(N+1)^8$ such submodular inequalities, though the number of valid combinations (a,b,c,d,e,f,g,h) is in fact even smaller due to their inherent relations. Therefore, the problem complexity in terms of the LP constraints is $O(KN^8)$.

In comparison, let us consider the complexity of the generic entropic LP approach [36] in the problem setting. The number of random variables there is at least $K+N+KN$, where the KN term is due to the answers from the N servers for the K messages. Thus there are a total of $2^{K+N+KN}-1$ joint entropy values as the variables in the generic entropic LP, and the $K+N+KN+\binom{K+N+KN}{2}2^{K+N+KN-2}$ elemental entropic constraints. It is possible to reduce the number of constraints using the symmetry reduction techniques introduced in [37], [38], but it will not change the exponential nature (see [43] for a more thorough analysis). In contrast, the complexity of the formulation in Proposition 2 is polynomial. The significant reduction in the number of constraints is due to the much more restricted set of submodular inequalities we include in this relaxed entropic LP, using the specific domain insights.

IV. Codes to Improve Known Upper Bounds

The lower convex envelop of the storage-retrieval pairs of the generalized MDS-PIR code provides an upper bound on the optimal tradeoff, which is the best known in the information theoretic PIR formulation. Equipped with the cyclic permutation lemma, in this section, we provide several new base code constructions which yield further improvements.

A. Construction-A: $N = 2$

We provide a construction for $N = 2$, which is based on the idea of compressing an existing code in [5]. Here the message length $L = 1$. We first provide an example, then present the general code construction.

Example: *Let $(K, N) = (3, 2)$. There 3 messages are a, b and c, respectively. The storage for each server is given in the following table,*

S_1	S_2
a	$a \oplus b$
b	$a \oplus c$
c	$\emptyset$

which implies that $\alpha = \frac{5}{2}$. Notice that $b \oplus c$ can be decoded by S_2 as $b \oplus c = (a \oplus b) \bigoplus (a \oplus c)$. Suppose the user desires

message a, it randomly chooses one row from the following table to retrieve the answers.

prob.	server 1	server 2
0.25	a	$\emptyset$
0.25	b	$a \oplus b$
0.25	c	$a \oplus c$
0.25	$a \oplus b \oplus c$	$(a \oplus b) \bigoplus (a \oplus c)$

Similarly, to retrieve $W_2 = b$, the following table is used.

prob.	server 1	server 2
0.25	b	$\emptyset$
0.25	a	$a \oplus b$
0.25	c	$b \oplus c$
0.25	$a \oplus b \oplus c$	$a \oplus c$

And to retrieve $W_3 = c$, the user uses the following table.

prob.	server 1	server 2
0.25	c	$\emptyset$
0.25	a	$a \oplus c$
0.25	b	$(a \oplus b) \bigoplus (a \oplus c)$
0.25	$a \oplus b \oplus c$	$a \oplus b$

It is seen that $\beta = \frac{7}{8}$. Compared with the capacity achieving code in [10] or [5], where $(\alpha, \beta) = \left(3, \frac{7}{8}\right)$, the storage is compressed while the download costs remains the same.

We next provide the code construction for more general K and $N = 2$.

- **Storage:** Let $S_1 = W_{1:K}$, and $S_2 = \{W_1 \oplus W_k\}_{k=2}^K$. It can be interpreted as server 1 stores all the messages, and server 2 stores all the even sum of messages, because the summation of any even number of messages can be constructed by S_2. The normalized average storage can be calculated simply as
$$\alpha = \frac{1}{2}(K + K - 1) = K - \frac{1}{2}.$$
- **Retrieval:** To retrieve message W_k, $k \in [1 : K]$, we randomly choose a length-K vector $\mathbf{v}$ of 0 and 1. Then retrieve $\bigoplus_{\mathbf{v}(j)=1} W_j$ and $W_k \oplus \bigoplus_{\mathbf{v}(j)=1} W_j$, each from one server. If the vector $\mathbf{v}$ has odd number of 1's, retrieve the former from server 2; otherwise, retrieve the latter from server 2. The user can recover the desired message by $W_k = A_1^{[k]} \oplus A_2^{[k]}$. The user will retrieve 2 bits unless $\mathbf{v}$ is consisted of all 0's or only the k-th position of $\mathbf{v}$ is 1, and in these two cases, the user only retrieve 1 bit. It follows that
$$\beta = \frac{1}{2 \times 2^K}(2^K \times 2 - 2) = \frac{2^K - 1}{2^K}.$$
To see that the protocol is private, observe that server 1 receives queries uniformly distributed over $\{\bigoplus_{\mathbf{v}(j)=1} W_j : \mathbf{v}$ has odd number of 1's$\}$; similarly, server 2 receives queries uniformly distributed over the set $\{\bigoplus_{\mathbf{v}(j)=1} W_j : \mathbf{v}$ has even number of 1's$\}$.

B. Construction-B: $K|(N-1)$

Next we provide a construction by generalizing the code in [10]. Here the message length L is 1, and $K = T(N-1)$ for some positive integer T. An example is given first, and then the general construction will be presented.

Example: *Let $(K, N) = (4, 3)$, and thus $T = 1$ in this example. There are four messages a, b, c, d. The contents are stored as in the following table.*

S_1	S_2	S_3
a	c	$a \oplus c$
b	d	$b \oplus d$

It is clear that $\alpha = 2$. Suppose the user desires message a, a row from the table below is chosen, uniformly at random, as the queries for the servers.

prob.	server 1	server 2	server 2
0.25	a	$\emptyset$	$\emptyset$
0.25	b	$c \oplus d$	$a \oplus c \oplus b \oplus d$
0.25	$\emptyset$	c	$a \oplus c$
0.25	$a \oplus b$	d	$b \oplus d$

Similarly, to retrieve $W_3 = c$, the following table is used.

prob.	server 1	server 2	server 2
0.25	$\emptyset$	c	$\emptyset$
0.25	$a \oplus b$	d	$a \oplus c \oplus b \oplus d$
0.25	a	$\emptyset$	$a \oplus c$
0.25	b	$c \oplus d$	$b \oplus d$

It is straightforward to see that code is private because the set of queries at one server is the same for different retrieving requirements and the query for retrieving any message is uniformly distributed over that set. Clearly $\beta = 0.75$. Comparing with $(4, 3)$-MDS coded PIR in [6], where $(\alpha, \beta) = (2, 0.802)$, the code given above has a smaller download cost.

In the general code construction, the index of a message can be represented either as $(N-1)i + j$ for $i \in [0 : T-1]$ and $j \in [1 : T]$, or as $Ta + b$ for $a \in [0 : N-2]$ and $b \in [1 : T]$.

- **Storage:** For $n \in [1 : N-1]$, server n stores $S_n = \{W_{T(n-1)+i}\}_{i=1}^T$; server N stores $S_N = \{\bigoplus_{j=0}^{N-2} W_{Tj+i}\}_{i=1}^T$. As a consequence $\alpha = T$.
- **Retrieval:** To retrieve $W_{T(a-1)+b}$, where $a \in [1 : N-1]$ and $b \in [1 : T]$. We randomly generate a vector $\mathbf{v}$ of 0 and 1 with length T. Let $\mathbf{v}'$ be a vector such that the only difference between $\mathbf{v}$ and $\mathbf{v}'$ is the b-th position, which is $\mathbf{v}'(b) = 1 \oplus \mathbf{v}(b)$. The user retrieves $A_n^{[k]} = \bigoplus_{\mathbf{v}'(i)=1} W_{T(n-1)+i}$ from server $n \in [1 : N-1]/\{a\}$; $A_a^{[k]} = \bigoplus_{\mathbf{v}(i)=1} W_{T(a-1)+i}$ from server a; and $A_N^{[k]} = \bigoplus_{\mathbf{v}'(i)=1} \bigoplus_{j=0}^{N-2} W_{Tj+i}$ from server N. Then the user can decode $W_k = \bigoplus_{n=1}^N A_n^{[k]}$. Thus
$$\begin{aligned}\beta &= \frac{1}{N}\left(\frac{1}{2^T} + \frac{1}{2^T}(N-1) + (1 - \frac{1}{2^{T-1}})N\right) \\ &= \frac{2^T - 1}{2^T}.\end{aligned}$$

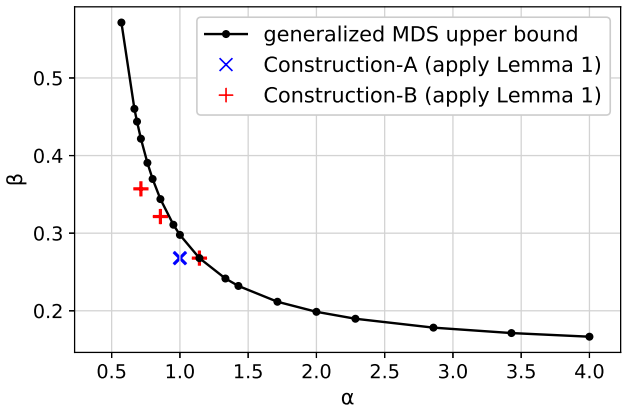


Fig. 5: Comparison of new upper bound and generalized MDS upper bound for $(N, K) = (7, 4)$.

To see that the protocol is private, observe that for any server $n \in [1 : N-1]$, the received query is uniformly distributed over $\left\{\bigoplus_{\mathbf{v}(i)=1} W_{Ta+i} : \mathbf{v} \in [0:1]^T\right\}$; server N receives queries uniformly distributed over $\left\{\bigoplus_{\mathbf{v}(i)=1}\bigoplus_{j=0}^{N-2} W_{Tj+i} : \mathbf{v} \in [0:1]^T\right\}$.

C. Applying the Cyclic Permutation Lemma

By applying the cyclic permutation lemma to the base codes given as construction-A and construction-B, we can obtain further improvement on the storage-retrieval tradeoff which is given in the following proposition.

Proposition 3. *For $N, K \geq 2$, the following tradeoff points are achievable:*
(a) $(\alpha, \beta) = \frac{2}{N}\left(K - \frac{1}{2}, \frac{2^K-1}{2^K}\right)$;
(b) $(\alpha, \beta) = \frac{K/T+1}{N}\left(T, \frac{2^T-1}{2^T}\right)$ *for all T being a factor of K so that $\frac{K}{T} + 1 \leq N$.*

In Fig. 5, we show the improvement of the tradeoff points in Proposition 3 compared with the best known upper bound, i.e. the generalized MDS tradeoff curve. Note that there is always a point generated from construction-B lies on the generalized MDS tradeoff curve. This can be seen by letting $T = K$ in Proposition 3 (b), then the resulting storage-retrieval tradeoff is obtained as $(\alpha, \beta) = \left(\frac{2K}{N}, \frac{1}{N}\frac{2^K-1}{2^{K-1}}\right)$ which is exactly the tradeoff point of the generalized MDS code in (3) for $(T_1, T_2) = (1, 2)$.

V. Explicit Lower Bounds

In this section, we derive more explicit lower bounds by further relaxing the linear program given in Proposition 2. Such derived bounds are more explicit, and moreover, we show numerically that the loss from those obtained using Proposition 2 is small, for the cases that the relaxed entropic LP can be effectively computed.

A. A Lower Bound for $(N-m)\alpha + m\beta$

For notation convenience, define the following function for any integer $m \in [1 : N]$,

$$B_N(K, m) \triangleq \inf_{\text{achievable } (\alpha,\beta)} \{(N-m)\alpha + m\beta\}. \tag{48}$$

The following boundary conditions are immediate (the last two are from (16) and (17)):

$$B_N(1, m) = 1, \text{ for } m \in [1 : N], \tag{49}$$
$$B_N(K, 1) = K, \tag{50}$$
$$B_N(K, N) = N\beta_0. \tag{51}$$

Let $\mathbb{C}$ be the set of positive real-valued vectors $\boldsymbol{c} = \left(c_j^n : j \in [1 : m], n \in [0 : N-j+1]\right)$ satisfying the conditions

$$\sum_{n=0}^{N-j+1} c_j^n = 1, \text{ for } j \in [1 : m], \tag{52}$$
$$0 \leq c_j^n \leq 1, \text{ for } j \in [1 : m], n \in [0 : N-j+1], \tag{53}$$

and the conditions

$$\sum_{i=j}^{N}(N-i+1)d_i \geq \sum_{i=j}^{m}(N-i+1), \text{ for all } j \in [2 : m], \tag{54}$$

where

$$d_j = \begin{cases} \sum_{i=2}^{j-1} c_i^{j-i}\left(\frac{1}{j-i} - \frac{1}{j-1}\right) + c_{j-1}^0 + \sum_{n=1}^{N-j+1}\frac{(n-1)c_j^n}{n}, \\ \qquad\qquad \text{if } j \in [2 : m] \\ c_m^0 + \sum_{i=2}^{m} c_i^{j-i}\left(\frac{1}{j-i} - \frac{1}{j-1}\right), \quad \text{if } j = m+1 \\ \sum_{i=2}^{m} c_i^{j-i}\left(\frac{1}{j-i} - \frac{1}{j-1}\right), \quad \text{if } j \in [m+2 : N]. \end{cases} \tag{55}$$

The lower bound is derived by a linear combination of Shannon-type inequalities. Specifically, c_j^n $(j \in [1 : m], n \in [0 : N-j+1])$ are the coefficients of the submodular inequalities and d_j $(j \in [2, m])$ are the coefficients for the residual terms of $B_N(K, m)$ after applying submodular inequalities. The above conditions (52)-(55) are chosen to maintain the validity of the bound. The detailed proof can be found in the supplementary material.

For any $\boldsymbol{c} \in \mathbb{C}$, we provide a general lower bound on $B_N(K, m)$ for $m \in [2 : N-1]$, through a recursive relation:

$$\widetilde{B_N}(K, m, \boldsymbol{c}) \triangleq 1 + c_1^1(K-1) + \sum_{j=2}^{m}\sum_{n=1}^{N-j+1}\frac{c_j^n}{j+n-1}\widetilde{B_N}(K-1, j+n-1, \boldsymbol{c}), \tag{56}$$

where the initial conditions are given as $\widetilde{B_N}(K, m, \boldsymbol{c}) = B_N(K, m)$ for i) $K = 1$; ii) $m = 1$; iii) $m = N$; i.e., the boundary conditions in (49)-(51).

Moreover, for $K \geq 2$, $N \geq 2$, and $m \in [1 : N]$, define

$$\underline{B_N}(K, m) \triangleq \max_{\boldsymbol{c}\in\mathbb{C}} \widetilde{B_N}(K, m, \boldsymbol{c}). \tag{57}$$

Theorem 1. *For $K \geq 2$, $N \geq 2$, and $m \in [1 : N]$, we have*

$$B_N(K, m) \geq \underline{B_N}(K, m) \geq \widetilde{B_N}(K, m, \boldsymbol{c}), \ \forall\ \boldsymbol{c} \in \mathbb{C}. \tag{58}$$

Proof overview for Theorem 1. The lower bound can be obtained by further relaxing the linear program in Proposition 2,

which is to minimize the objective function under a chosen subset of constraints in a systematic manner. The idea is to first specify which and how the submodular inequalities are applied (i.e., utilize an even smaller subset of the possible submodular inequalities), and then apply the other inequalities accordingly. More specifically, we use only the submodular constraints for $j \in [2:N-1],\ i \in [j:N]$ that

$$\begin{aligned} &H(A_{1:j}^{[1]}, S_{j+1:N}|\mathbf{F}, W_1) + H(S_{j:N}|\mathbf{F}, W_1) \\ &\geq H(A_j^{[1]}, S_{j+1:N}|\mathbf{F}, W_1) + H(A_{1:j-1}^{[1]}, S_{j:N}|\mathbf{F}, W_1), \quad (59) \\ &H(A_{j:i-1}^{[2]}, S_{i:N}|\mathbf{F}, W_1) + H(A_i^{[2]}, S_{j:i-1}, S_{i+1:N}|\mathbf{F}, W_1) \\ &\geq H(A_{j:i}^{[2]}, S_{i+1:N}|\mathbf{F}, W_1) + H(S_{j:N}|\mathbf{F}, W_1). \quad (60) \end{aligned}$$

The detailed proof can be found in supplementary material. □

The bound given in Theorem 1 is still not explicit, and next we specialize it even further, in order to obtain a more explicit form. This is accomplished by choosing a specific set of $\boldsymbol{c} \in \mathbb{C}$. More precisely, we will show that the following value of $\underline{\underline{B_N}}(K,m)$, $m \in [2:N-1]$ is feasible,

$$\underline{\underline{B_N}}(K,m) = \begin{cases} 1 + \frac{m-j^*+1}{N} + \sum_{j=j^*}^{m-1} \frac{(m-j)(m+j-1)}{2j(j-1)N} \\ \quad + \frac{N-j^*}{N-j^*+1}\left(1 - \frac{1}{N-j^*}\frac{(m-j^*)(m+j^*-1)}{2(j^*-1)}\right), & \text{if } K=2 \\ 1 + \sum_{j=j^*}^{m} \frac{1}{j} \cdot \underline{\underline{B_N}}(K-1,j) \\ \quad + \frac{1}{N-j^*+1}\left[\frac{(N-m)(m-1)}{m} - \sum_{i=j^*+1}^{m} \frac{N-i+1}{i-1}\right] \\ \quad \cdot \underline{\underline{B_N}}(K-1,j^*-1), & \text{if } K \geq 3, \end{cases} \quad (61)$$

where the initial conditions are given by

$$\underline{\underline{B_N}}(1,m) = 1, \text{ for } m \in [1:N], \quad (62)$$

$$\underline{\underline{B_N}}(K,1) = K, \quad (63)$$

$$\underline{\underline{B_N}}(K,N) = N\beta_0, \quad (64)$$

and $j^* \in [2:m]$ is defined as follows:

- For $K=2$, j^* is given as

$$j^* = \max\left\{2, \left\lceil (N+\frac{1}{2}) - \sqrt{(N-m)(N+m-1)+\frac{1}{4}} \right\rceil\right\}. \quad (65)$$

- For $K \geq 3$, j^* is the minimum j such that

$$\sum_{i=j+1}^{m} \frac{N-i+1}{i-1} \leq \frac{(m-1)(N-m)}{m}, \quad (66)$$

which is equivalent to

$$N\sum_{i=j}^{m} \frac{1}{i} + j \leq N+1. \quad (67)$$

If the above inequality holds for all $j \in [2:m]$, then let $j^*=2$.

We can upper bound the LHS of (67) by

$$N\sum_{i=j}^{m} \frac{1}{i} + j \leq N \cdot \ln\left(\frac{m+1/2}{j-1/2}\right) + j, \quad (68)$$

where $f(j) = N \cdot \ln\left(\frac{m+1/2}{j-1/2}\right) + j$ is obtained by the convexity of the reciprocal function $\frac{1}{x}$, which is

$$\sum_{i=j}^{m} \frac{1}{i} \leq \int_{j-\frac{1}{2}}^{m+\frac{1}{2}} \frac{1}{x}dx = \ln\left(\frac{m+1/2}{j-1/2}\right). \quad (69)$$

The following theorem is our main result of this section.

Theorem 2. *For $K \geq 2$, $N \geq 2$, and $m \in [1:N]$, we have*

$$B_N(K,m) \geq \underline{\underline{B_N}}(K,m). \quad (70)$$

Proof. The lower bound $\underline{\underline{B_N}}(K,m)$ defined in (61) can be obtained by simply assigning a proper feasible value to $\boldsymbol{c}$ and substituting this $\boldsymbol{c}$ into the lower bound $\widetilde{B_N}(K,m,\boldsymbol{c})$ in (56). Then the theorem follows directly from Theorem 1. The details can be found in the supplementary material. □

Remark 3. The definition of $\underline{\underline{B_N}}(K,m)$ in (61) is defined for $m \in [2:N-1]$. We can verify that for $m=N$, we have $j^*=N$, and then (61) gives $\underline{\underline{B_N}}(K,N) = N\beta_0$ which is consistent with the initial condition in (64). However, for $m=1$, we have $j^*=2$, and (61) gives $\underline{\underline{B_N}}(K,1) = 1 <= K$, which is not consistent with the initial condition in (63).

B. Bounding $\alpha + m\beta$ with Large Integer m

Similar to the previous case, we can also further relax the relaxed entropic LP to find the following lower bound.

Theorem 3. *For $K \geq 2$, $N \geq 2$, and $m = (N-1) + (N-2)N^{K-k}$ for $k \in [1:K]$, the term $\alpha + m\beta$ can be lower bounded as follows:*

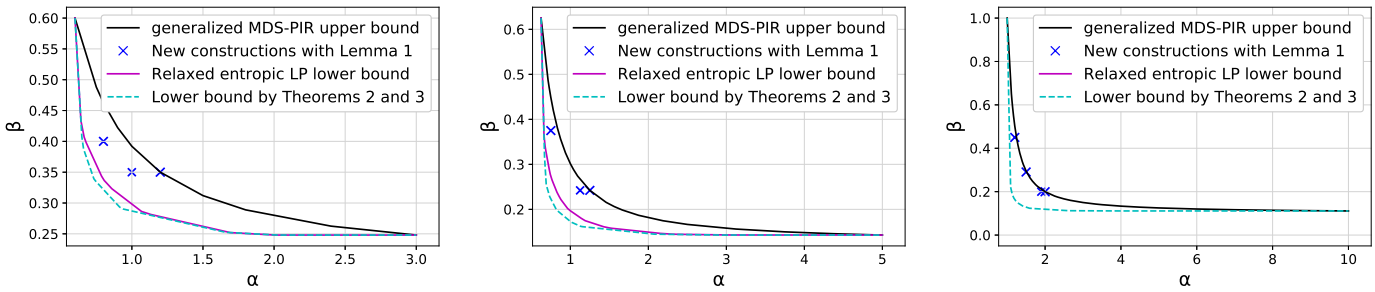
1) If $k \leq \frac{K}{2}$, then

$$\begin{aligned} \alpha + m\beta \geq &\left(k + \frac{N^{K-2k}(N^k-1)(N-2)}{(N-1)}\right) \\ &+ \frac{1}{N^{k-1}}\left[\frac{(N^{K-k}-1)(N-2)}{N(N-1)} + (K-k)\right] \\ &+ \left(1 - \frac{1}{N^{k-1}}\right)\Big(B_N(K-k+1, N-1) - 1\Big), \end{aligned}$$

2) If $k > \frac{K}{2}$, then

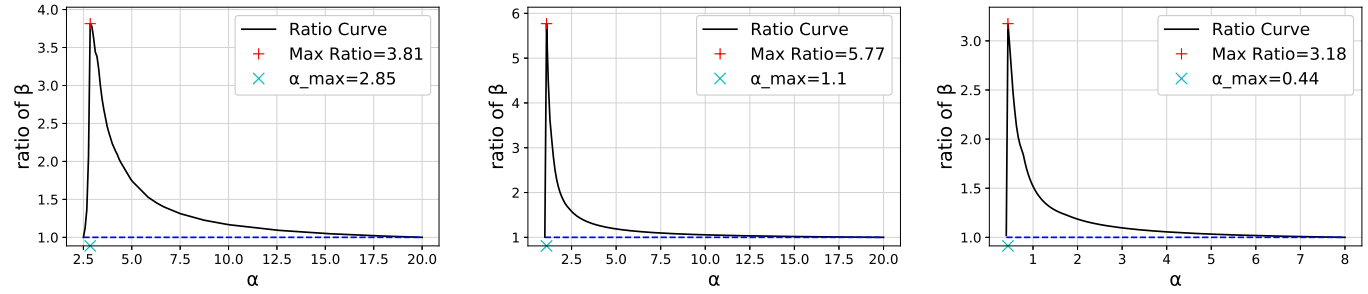
$$\begin{aligned} \alpha + m\beta \geq &\left(K - k + \frac{(N-2)(N^{K-k}-1)}{N-1}\right) \\ &+ \frac{2(N^{2k-K}-1)}{N^{2k-K}} \\ &+ (N-1)\sum_{i=1}^{2k-K-1} \frac{1}{N^i}\Big(B_N(k-i+1, N-1) - 1\Big) \\ &+ \frac{1}{N^{k-1}}\left[\frac{(N^{K-k}-1)(N-2)}{N(N-1)} + (K-k)\right] \\ &+ \frac{N^{K-k}-1}{N^{k-1}}\Big(B_N(K-k+1, N-1) - 1\Big). \end{aligned}$$

Proof. Similar to Theorem 1, the lower bound can be obtained by minimizing the objective function in Proposition 2 under a



(a) (b) (c)

Fig. 6: Comparison of the lower bounds and upper bounds for $(N, K) = (5, 3), (8, 5), (10, 10)$.



(a) (b) (c)

Fig. 7: The ratio of the upper bound and lower bound for $(N, K) = (8, 20), (20, 20), (20, 8)$.

chosen subset of constraints. Specifically, we use the submodular constraints for $j \in [2 : N-1],\ i \in [1 : j]$ that

$$\begin{aligned}
&H(A_{1:N}^{[k]}|\mathbf{F}, W_{1:k}) + H(S_{j+1:N}|\mathbf{F}, W_{1:k}) \\
&\geq H(A_{1:j}^{[k]}, S_{j+1:N}|\mathbf{F}, W_{1:k}) + H(A_{j+1:N}^{[k]}|\mathbf{F}, W_{1:k}), \\
&H(A_{1:i}^{[k+1]}, S_{j+1:N}|\mathbf{F}, W_{1:k}) + H(A_{i+1}^{[k+1]}, S_{j+1:N}|\mathbf{F}, W_{1:k}) \\
&\geq H(A_{1:i+1}^{[k+1]}, S_{j+1:N}|\mathbf{F}, W_{1:k}) + H(S_{j+1:N}|\mathbf{F}, W_{1:k}).
\end{aligned}$$

The details can be found as supplementary material. □

Remark 4. From Fig. 4, we see that the two constraints given in [13] are also obtained through the relaxed entropic LP. Now these two constraints are also included in the further relaxed explicit expression, where the first one (Theorem 1 in [13]) is simply (50) and the second one (Theorem 2 in [13]) can be obtained from Theorem 3 by letting $k = 1$ which becomes

$$\begin{aligned}
&\alpha + \left[(N-1) + (N-2)N^{K-1}\right]\beta \\
&\geq K + \frac{(N-2)(N^K - 1)}{N(N-1)}.
\end{aligned} \tag{71}$$

C. Numerical Results

We first compare the proposed lower bounds and the upper bounds of the storage-retrieval tradeoff in Fig. 6. The upper bounds in Section IV evidently outperform the generalized MDS-PIR upper bound at the storage-bound regime. We further observe that the lower bound in Theorem 2 and Theorem 3 is close to the relaxed entropic LP lower bound in Section III-C. Recall that the relaxed entropic LP has constraints on the order of $O(KN^8)$, which becomes unmanageable for larger N, K, e.g., we found $K = N = 10$ can not be effectively computed in a reasonable amount of time. However, the lower bound in Theorem 2 and Theorem 3 is obtained only by direct calculation and can be solved quickly for large K and N. These upper bounds and lower bounds help to further refine the approximation.

We next further analyze the difference in Fig. 7, and consider whether the new bounds will be able to provide a tighter approximation ratio. Since both the lower bounds and the upper bounds are small for large K and N, we plot the ratio of the upper and lower bounds instead of their difference. Unfortunately, though the new bounds indeed provide improvement over the existing art, it appears they are not sufficient to yield a better approximation of β in the storage-bound regime, and the largest ratio gap appears to be just above $\frac{K}{N}$; the precise positions of the largest gap are given in the respective figures.

VI. Conclusion

We studied the tradeoff between the storage cost and the download cost in private information retrieval systems. Three fundamental results are first presented: a regime-wise 2-approximation, a cyclic permutation lemma, and a relaxed entropic LP with polynomial complexity. Equipped with these results, we then provide improved upper bounds and lower bounds. Though these results provide significant new insights into the storage-retrieval tradeoff in PIR systems, the characterization is not tight in general. As a future work, we plan

to further investigate the relaxed entropic LP and derive an improved lower bound that can yield better approximations or precise characterizations.

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proceedings of IEEE 36th Annual Foundations of Computer Science*, Oct. 1995, pp. 41–50.

[2] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 965–981, Nov. 1998.

[3] N. B. Shah, K. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *2014 IEEE International Symposium on Information Theory (ISIT)*, Honolulu, HI, Jul. 2014, pp. 856–860.

[4] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

[5] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Trans. Inf. Theory*, vol. 65, pp. 7613–7627, Nov. 2019.

[6] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[7] R. Tandon, M. Abdul-Wahid, F. Almoualem, and D. Kumar, "PIR from storage constrained databases - coded caching meets PIR," in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–7.

[8] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6617–6634, Nov. 2020.

[9] K. Banawan, B. Arasli, and S. Ulukus, "Improved storage for efficient private information retrieval," in *2019 IEEE Information Theory Workshop (ITW)*, Visby, Gotland, Sweden, Aug. 2019, pp. 1–5.

[10] H. Sun and C. Tian, "Breaking the MDS-PIR capacity barrier via joint storage coding," *Information*, vol. 10, no. 9, p. 265, Aug. 2019.

[11] C. Tian, H. Sun, and J. Chen, "A Shannon-theoretic approach to the storage-retrieval tradeoff in PIR systems," pp. 1904–1908, Jun. 2018.

[12] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4904–4916, Aug. 2020.

[13] C. Tian, "On the storage cost of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7539–7549, Dec. 2020.

[14] H.-Y. Lin, S. Kumar, E. Rosnes, and A. G. i Amat, "An MDS-PIR capacity-achieving protocol for distributed storage using non-MDS linear codes," in *2018 IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, Jun. 2018, pp. 966–970.

[15] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 2852–2856.

[16] T. H. Chan, S. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 2842–2846.

[17] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics and Security*, vol. 12, no. 12, pp. 2920–2932, Dec. 2017.

[18] H. Yang, W. Shin, and J. Lee, "Private information retrieval for secure distributed storage systems," *IEEE Trans. Inf. Forensics and Security*, vol. 13, no. 12, pp. 2953–2964, Dec. 2018.

[19] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[20] S. Kumar, H.-Y. Lin, E. Rosnes, and A. Graell i Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4243–4273, Jul. 2019.

[21] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[22] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "Weakly-private information retrieval," in *2019 IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1257–1261.

[23] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Trans. Inf. Forensics and Security*, vol. 15, pp. 2999–3012, Mar. 2020.

[24] Q. Wang and M. Skoglund, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3183–3197, May 2019.

[25] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.

[26] Q. Wang and M. Skoglund, "Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5160–5175, Aug. 2019.

[27] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *2019 IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1262–1266.

[28] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric," in *2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, Jun. 2020, pp. 1–6.

[29] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.

[30] K. Banawan and S. Ulukus, "The capacity of private information retrieval from byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[31] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al." *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[32] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[33] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

[34] Y. Zhang, X. Wang, H. Wei, and G. Ge, "On private information retrieval array codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5565–5573, Sep. 2019.

[35] R. Tandon, "The capacity of cache aided private information retrieval," in *2017 55th Annual Allerton Conference*, Oct. 2017, pp. 1078–1082.

[36] R. W. Yeung, "A framework for linear information inequalities," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1924–1934, Nov. 1997.

[37] C. Tian, "Characterizing the rate region of the (4, 3, 3) exact-repair regenerating codes," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 5, pp. 967–975, May 2014.

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

[39] Z. Zhang and J. Xu, "The optimal sub-packetization of linear capacity-achieving PIR schemes with colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2723–2735, May 2019.

[40] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM Journal on Applied Algebra and Geometry*, vol. 1, no. 1, pp. 647–664, Nov. 2017.

[41] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al." *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[42] R. G. L. D'Oliveira and S. El Rouayheb, "One-shot PIR: Refinement and lifting," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2443–2455, Apr. 2020.

[43] K. Zhang and C. Tian, "On the symmetry reduction of information inequalities," *IEEE Trans. Communications*, vol. 66, no. 6, pp. 2396–2408, Jun. 2017.