# On the Storage Cost of Private Information Retrieval

Chao Tian
Department of Electrical and Computer Engineering
Texas A&M University
chao.tian@tamu.edu

*Abstract*—We consider the fundamental tradeoff between the storage cost and the download cost in private information retrieval systems, without any explicit structural restrictions on the storage codes, such as maximum distance separable codes or uncoded storage. Two novel outer bounds are provided, which have the following implications. When the messages are stored without any redundancy across the databases, the optimal PIR strategy is to download all the messages; on the other hand, for PIR capacity-achieving codes, each database can reduce the storage cost, from storing all the messages, by no more than one message on average. We then focus on the two-message two-database case, and show that a stronger outer bound can be derived through a novel pseudo-message technique.

## I. Introduction

Private information retrieval [1], or simply referred to as PIR, is a fundamental privacy-preserving information processing primitive. PIR has deep connections to other well-known communication and coding problems such as locally decodable codes [2], [3] and interference alignment [4]. The PIR capacity, i.e., the inverse of the minimum download cost, was recently characterized by Sun and Jafar [5], and the PIR capacities under other variations have also been considered [6]–[20].

Since the databases from which the information is retrieved are basically storage nodes, designing efficient storage strategies in PIR systems is of significant importance, and the problem has received considerable attention recently. Some of the existing efforts assume certain specific coding structures in the storage side, such as using maximum distance separable (MDS) codes [6], [15] or in an uncoded form [21]. In two recent works [11], [22], the tradeoff was considered without any structural constraints on the storage or retrieval codes for the special case of two databases and two messages, and it was found that non-linear codes can provide further improvement over linear codes. Other notable efforts can be found in [23]–[27] and references therein.

Despite these efforts, our understanding of the fundamental tradeoff between the storage cost $\alpha$ and the download cost $\beta$ is still quite limited, mainly due to the lack of general information theoretic converse results when the storage codes are not required to follow any potentially restricting structural constraint. In this work, we initialize such an effort and derive several information theoretical outer bounds of the fundamental tradeoff between the storage cost and the download cost. Instead of attempting to characterize the complete optimal tradeoff curve, in this work we focus on the two extreme

points: the point when the storage cost is minimal, and the point when the download cost is minimal. For the former, the question is when the storage has no redundancy, what the minimum download cost can be; for the latter, the question is for PIR capacity-achieving codes, what the minimum storage cost can be.

Our main result is a precise characterization of the first extreme point (no redundancy in storage), and an approximate characterization of the second (capacity-achieving PIR codes). More precisely, by providing two novel outer bounds, we show that for the former, the optimal PIR strategy is in fact to download all the messages, whereas for the latter, i.e., for PIR capacity-achieving codes, each database can reduce the storage cost by no more than one message on average, compared to the simple strategy of storing every message. In order to better understand the second extreme point, we then focus on the two-message two-database case, and show that a stronger outer bound can be derived using a novel pseudo-message technique, the origin of which can be traced back to Ozarow's bounding technique [28] of extending the original probability space through Markov coupling, and is also related to the derivation of non-Shannon type inequalities [29]–[31].

## II. Problem Definition

The problem setting of private information retrieval is well known, see e.g., [5], however in order to fix the notation and derive the outer bounds in question, we first introduce a rigorous version of the problem definition. We then provide the formal definitions of the operational download cost and storage cost, together with the informational version, in order to unify the different conventions in the literature and avoid confusion in subsequent discussions.

### A. Encoding and Decoding Functions

There are a total of $K$ messages $W_1, W_2, \ldots, W_K$ in the system, and there are a total of $N$ databases that the messages are to be retrieved; the messages $(W_i, W_{i+1}, \ldots, W_j)$ will sometimes be written together as $W_{i:j}$ for conciseness. Denote the set of possible queries that server-$n$ can accommodate as $\mathcal{Q}_n$, and denote its cardinality as $|\mathcal{Q}_n|$. The cardinality of a set $\mathcal{A}$ will be similarly denoted as $|\mathcal{A}|$ in the rest of the paper. Assume that a random key $\mathsf{F}$ is uniformly distributed on a certain finite set $\mathcal{F}$, which is used by the user to produce the (random) queries to the $N$ databases. The servers, after receiving the queries for message-$k$, denoted as $Q_n^{[k]}$, will reply with an answer $A_n^{[k]}$. A message $W_k$ consists of

$L$ symbols, each symbol belonging to a finite alphabet $\mathcal{X}$. The messages are mutually independent, each of which is uniformly distributed on $\mathcal{X}^L$. We further allow the query answers to be represented as a variable-length vector, whose elements are in the finite alphabet $\mathcal{Y}$. A mathematically precise description of the problem is given next via a set of encoding and decoding functions.

**Definition 1.** *An $N$-server private information retrieval (PIR) storage code for $K$ messages, each of $L$-symbols in the alphabet $\mathcal{X}$, consists of*

1) *$N$ storage encoding functions:*
$$\Phi_n : \mathcal{X}^{KL} \to \mathcal{S}_n, \qquad n \in \{1, 2, ..., N\}, \qquad (1)$$
*i.e., the stored information at database-$n$ is $S_n = \Phi_n(W_{1:K})$;*

2) *$N$ query functions:*
$$\phi_n : \{1, 2, \ldots, K\} \times \mathcal{F} \to \mathcal{Q}_n, \qquad n \in \{1, 2, ..., N\}, \qquad (2)$$
*i.e., the user chooses the query $Q_n^{[k]} = \phi_n(k, \mathsf{F})$ for server-$n$, using the index of the desired message and the random key $\mathsf{F}$;*

3) *$N$ answer length functions:*
$$\ell_n : \mathcal{Q}_n \to \{0, 1, 2, \ldots\}, \qquad n \in \{1, 2, ..., N\}, \quad (3)$$
*i.e., the length of the answer at each server, a non-negative integer, is a deterministic function of the query, but not the particular realization of the messages;*

4) *$N$ answer functions:*
$$\varphi_n : \mathcal{Q}_n \times \mathcal{S}_n \to \mathcal{Y}^{\ell_n}, \qquad n \in \{1, 2, ..., N\}, \qquad (4)$$
*where $\ell_n = \ell_n(q_n)$ with $q_n \in \mathcal{Q}_n$ being the (random) query for server-$n$, $\mathcal{Y}$ is the coded symbol alphabet, and in the sequel we write the query answer as $A_n^{[k]} \triangleq \varphi_n(Q_n^{[k]}, S_n)$ when the message index $k$ is relevant;*

5) *A reconstruction function using the answers from the servers together with the desired message index and the random key:*
$$\psi : \prod_{n=1}^{N} \mathcal{Y}^{\ell_n} \times \{1, 2, ..., K\} \times \mathcal{F} \to \mathcal{X}^L, \qquad (5)$$
*i.e., $\hat{W}_k = \psi(A_{1:N}^{[k]}, k, \mathsf{F})$ is the retrieved message.*

*These functions should satisfy the following two requirements:*

1) **Correctness:** *For any $k \in \{1, 2, ..., K\}$, $\hat{W}_k = W_k$.*
2) **Privacy:** *For every $k, k' \in \{1, 2, ..., K\}$, $n \in \{1, 2, ..., N\}$, and $q \in \mathcal{Q}_n$,*
$$\mathbf{Pr}(Q_n^{[k]} = q) = \mathbf{Pr}(Q_n^{[k']} = q). \qquad (6)$$

It should be noted that $A_n^{[k]}$ is a function of both the messages and the query $Q_n^{[k]}$. Sometimes we need to refer to the answer for a fixed query $Q_n^{[k]} = q$, and this shall be written as $A_n^{(q)}$.. As a consequence of the privacy requirement

and the coding requirement, the following joint distributions must also be identical for any $n = 1, 2, \ldots, N$,
$$(A_n^{[k]}, Q_n^{[k]}, W_{1;K}, S_{1:N}) \sim (A_n^{[k']}, Q_n^{[k']}, W_{1;K}, S_{1:N}),$$
$$k, k' \in \{1, 2, \ldots, K\}, \qquad (7)$$
which implies that their marginal distributions will also be identical.

*B. Operational and Informational Costs*

The *operational* normalized storage cost for database-$n$ is
$$\alpha_n \triangleq \frac{\log_2 |\mathcal{S}_n|}{L \log_2 |\mathcal{X}|}, \qquad n = 1, 2, \ldots, N, \qquad (8)$$
which is the amount of stored data per bit of individual message; the average storage (per node) cost is defined as
$$\alpha = \frac{1}{N} \sum_{n=1}^{N} \alpha_n. \qquad (9)$$

The *operational* normalized download cost for database-$n$ is
$$\beta_n \triangleq \frac{\log_2 |\mathcal{Y}| \mathbb{E}(\ell_n)}{L \log_2 |\mathcal{X}|}, \qquad n = 1, 2, \ldots, N, \qquad (10)$$
which is the expected amount of downloaded data per bit desired message at database-$n$, and the average per node download cost is
$$\beta = \frac{1}{N} \sum_{n=1}^{N} \beta_n. \qquad (11)$$
Note that $\beta_n$ does not depend on $k$, since the privacy requirement stipulates that the random variable $\ell_n$ has an identical distribution for all $k = 1, 2, \ldots, K$.

In the literature (e.g., [11]), the *informational* storage cost is sometimes used directly in place of the operational storage cost, i.e.,
$$\alpha_n' \triangleq \frac{H(S_n)}{L \log_2 |\mathcal{X}|}, \qquad n = 1, 2, \ldots, N, \qquad (12)$$
and correspondingly the average informational storage cost can be defined as
$$\alpha' = \frac{1}{N} \sum_{n=1}^{N} \alpha_n'. \qquad (13)$$
Similarly the *informational* download cost can be given as
$$\beta_n' \triangleq \frac{H(A_n^{[k]}|\mathsf{F})}{L \log_2 |\mathcal{X}|}, \qquad n = 1, 2, \ldots, N, \qquad (14)$$
and the average informational download cost
$$\beta' = \frac{1}{N} \sum_{n=1}^{N} \beta_n'. \qquad (15)$$
Note $\beta_n'$ does not depend on $k$, which can be justified as
$$\frac{H(A_n^{[k]}|\mathsf{F})}{L \log_2 |\mathcal{X}|} = \frac{H(A_n^{[k]}|Q_n^{[k]})}{L \log_2 |\mathcal{X}|}$$
$$= \frac{H(A_n^{[k']}|Q_n^{[k']})}{L \log_2 |\mathcal{X}|} = \frac{H(A_n^{[k']}|\mathsf{F})}{L \log_2 |\mathcal{X}|}, \qquad (16)$$

where the first and last equality are due the Markov string $F \leftrightarrow Q_n^{[k]} \leftrightarrow A_n^{[k]}$, and the equality in the middle is due to the privacy condition (7).

It is clear that

$$\alpha_n \geq \alpha'_n, \quad n = 1, 2, \ldots, N. \qquad (17)$$

The relation between $\beta_n$ and $\beta'_n$ is slightly more subtle. It can be seen that for any $n = 1, 2, \ldots, N$,

$$\beta_n = \frac{\log_2 |\mathcal{Y}| \mathbb{E}(\ell_n)}{L \log_2 |\mathcal{X}|} = \frac{\mathbb{E}\left[\mathbb{E}(\ell_n|F)\right] \log_2 |\mathcal{Y}|}{L \log_2 |\mathcal{X}|}$$
$$\geq \frac{\mathbb{E}H(A_n^{[k]}|F = f)}{L \log_2 |\mathcal{X}|} = \frac{H(A_n^{[k]}|F)}{L \log_2 |\mathcal{X}|}. \qquad (18)$$

As a consequence, we have

$$\alpha \geq \alpha', \qquad \beta \geq \beta', \qquad (19)$$

but they may not be equal. In this work, we adopt from a first principle the operational definitions, from which the informational definitions will emerge as the substitutes naturally.

It was shown [5] that the minimum download cost is

$$\min \beta = \frac{1}{N} + \ldots + \frac{1}{N^K} = \frac{N^K - 1}{N^K(N - 1)}, \qquad (20)$$

and codes that can achieve this minimal value are often referred to as optimal private information retrieval codes, or capacity-achieving private information retrieval codes. The codes are optimal in the sense that the download cost is minimal. Note that the capacity achieving code given in [5] assumed fully replicated messages at all databases, however in our setting the databases are not necessarily replicated.

## III. EXTREME POINT CHARACTERIZATIONS

We consider the two extreme points in question in the following two subsections, respectively.

### A. Minimizing Download Cost without Storage Redundancy

The first main result we present is for the extreme case when the messages are stored across the databases without any redundancy, i.e., $N\alpha = K$. With such compressed storage, we shall provide a converse to confirm the folklore that the best PIR download strategy is to download everything.

**Theorem 1.** *The per-node retrieval cost $\beta$ and per-node storage cost $\alpha$ must satisfy*

$$(N - 1)\alpha + \beta \geq K. \qquad (21)$$

We omit the proof here, which can be found in [32]. This bound is illustrated in Fig. 1, together with the achievable tradeoffs with MDS-coded storage [9] and uncoded storage [21]. It is seen that this bound is almost vertical, but it is sufficient to characterize one of two extreme points. The next corollary now follows directly from the theorem, which states that when the storage code has no redundancy, the optimal strategy is to download every message.
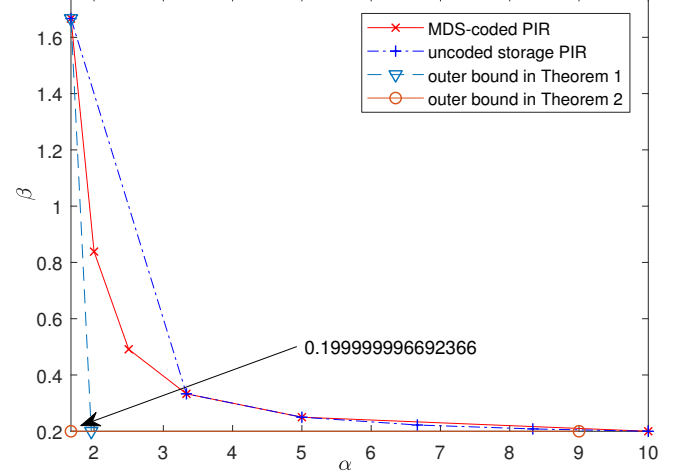


Fig. 1. Bounds in Theorem 1 and Theorem 2 when $N = 6$ and $K = 10$, where the arrow indicates the intercept of the bound in Theorem 2 and $\beta = \frac{1}{5} - \frac{1}{5*6^{10}}$ is the minimum download cost.

**Corollary 1.** *At the minimum storage point $\alpha = \frac{K}{N}$, we must have*

$$\beta \geq \frac{K}{N}. \qquad (22)$$

*Moreover, the equality can be achieved by downloading all the information from the databases.*

### B. Minimum Storage Overhead for Capacity-Achieving Codes

The next main result is a novel lower bound on the storage cost and download cost tradeoff, which leads to an approximate characterization of the extreme point for capacity-achieving codes.

**Theorem 2.** *The per-node retrieval cost $\beta$ and per-node storage cost $\alpha$ when $N \geq 3$ must satisfy*

$$\frac{\alpha + (N - 1)\beta}{N - 2} + N^{K-1}\beta \geq \frac{K}{N - 2} + \frac{N^K - 1}{N(N - 1)}. \qquad (23)$$

The proof of this theorem can be found in [32]. The most important implication of this theorem is the following corollary, which states that for capacity-achieving PIR codes, each database must store at least $K - 1$ messages on average.

**Corollary 2.** *At the PIR capacity point $\beta = \frac{N^K - 1}{N^K(N-1)}$, we have*

$$\alpha \geq \left(K - \frac{N^K - 1}{N^K}\right) > (K - 1). \qquad (24)$$

An illustration of this bound can also be found in Fig. 1. It can be seen that the proposed bound in Theorem 2 is almost horizontal, and its intersection with the horizontal axis gives a lower bound on the minimum storage cost when the code is capacity-achieving. Since a trivial storage solution is to replicate all the messages at all the databases, this corollary in fact provides a characterization of the minimum storage cost for capacity-achieving codes within an additive gap of one message.

In the proof of Theorem 2, the following auxiliary quantities and their relation are important :

$$T^k \triangleq H(A_{1:N}^{[k]}|W_{1:k}, \mathsf{F}), \quad k = 1, 2, \ldots, K, \qquad (25)$$

$$V_n^k \triangleq H(A_{1:n-1,n+1:N}^{[k]}, S_n|W_{1:k}, \mathsf{F}),$$
$$n = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, K, \qquad (26)$$

$$V^k = \frac{\sum_{n=1}^{N} V_n^k}{N}, \quad k = 1, 2, \ldots, K. \qquad (27)$$

Due to the definitions above, it is clear that

$$T^K = V^K = 0. \qquad (28)$$

The following two auxiliary lemmas are instrumental to the proof of the theorem, whose proofs can be found in [32]. The first lemma is a recursive relation on $T^k$.

**Lemma 1.** *For any* $k = 1, 2 \ldots, K - 1$

$$T^k \geq \frac{L \log_2 |\mathcal{X}|}{N} + \frac{T^{k+1}}{N}. \qquad (29)$$

The second lemma is a refined recursive relation involving both $V^k$ and $T^k$.

**Lemma 2.** *For any* $k = 1, 2 \ldots, K - 1$ *and* $n = 1, 2, \ldots, N$,

$$\frac{V^k}{N-2} + T^k \geq \left(\frac{1}{N-2} + \frac{1}{N}\right) L \log_2 |\mathcal{X}| + \frac{V^{k+1}}{N-2} + \frac{T^{k+1}}{N}. \qquad (30)$$

Using the recursive relations among $T^k$'s and $V^k$'s in these two lemmas, an induction can be used to prove Theorem 2, the details of which can be found in [32].

## IV. AN IMPROVED OUTER BOUND VIA PSEUDO MESSAGES

In this section, we shall take a closer look at the special case $N = K = 2$. Theorem 1 in this case specializes to the bound

$$\alpha + \beta \geq 2. \qquad (31)$$

As a consequence, the minimum download cost $\beta$ when the storage cost is minimal is clearly $\beta = 1$, and this settles one of the two extreme points. Since the messages must be held in the databases as a whole, it is clear that $\alpha \geq 1$ regardless $\beta$, however, the more sophisticated bound in Theorem 2 does not apply to $N = 2$. The question we wish to address in this section is for this special case $N = K = 2$, whether the other extreme point, where the download cost is minimal, can be more accurately approximated. For this purpose, we present a novel outer bound based on a pseudo-message technique.

The main result of this section is the following theorem, which provides an improved outer bound for the storage cost and download cost tradeoff.

**Theorem 3.** *For $N = K = 2$, we must have $3\alpha + 8\beta \geq 10$.*

**Corollary 3.** *For $N = K = 2$, any capacity achieving codes, i.e., codes with $\beta = 0.75$, must satisfy $\alpha \geq 4/3$.*

This bound is illustrated in Fig. 2, together with the best known achievable tradeoff discovered in [22]. It can be seen
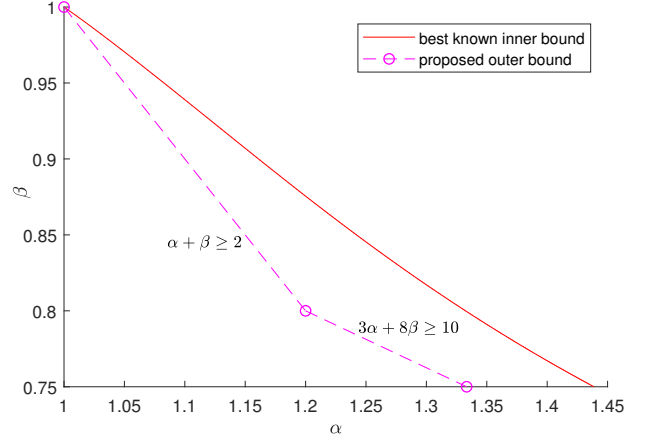


Fig. 2. The proposed outer bound and the best known inner bound when $N = K = 2$.

that this new bound can indeed improve the accuracy of the approximation for the extreme point on the horizontal axis. The proof of this bound is rather technical, the details of which can be found in [32], but the main proof idea is given and discussed in the remainder of this section. The new bound in Theorem 3 appears difficult to generalize to larger values of $N$ and $K$. Nevertheless, it can be viewed as a strong piece of evidence that the outer bounds we have at this point are likely not tight, and more sophisticated techniques involving non-Shannon type inequalities may provide additional improvement. In fact, without using the pseudo-message technique, we have indeed applied the computational approach [33], [34] on this two-message two-database problem, i.e., invoking all Shannon type inequalities, which did not produce any bound stronger than that in Theorem 1.

In order to prove this bound, we need to first introduce some necessary simplifications on the code structure used in the derivation. With such simplification, we shall discuss the main pseudo-message technique in some more details.

### A. Symmetry Assumptions

It was shown in [19] that there are three types of symmetry relations in this private information retrieval problem. By applying the three types of symmetry relations, it can be shown that in the setting of identifying the optimal storage cost and the download cost, it is without loss of optimality to consider only codes such that

$$H(S_n) = H(S_{n'}), \quad n, n' \in \{1, 2, \ldots, N\}$$
$$H(A_n^{(q)}) = H(A_{n'}^{(q')}),$$
$$n, n' \in \{1, 2, \ldots, N\}, \quad q \in \mathcal{Q}_n, \quad q' \in \mathcal{Q}_{n'}$$
$$H(A_n^{(q)}, W_k) = H(A_{n'}^{(q')}, W_{k'}), \quad n, n' \in \{1, 2, \ldots, N\},$$
$$q \in \mathcal{Q}_n, \quad q' \in \mathcal{Q}_{n'}, \quad k, k' \in \{1, 2, \ldots, K\}. \qquad (32)$$

In other words, for such symmetric codes, the databases use the same amount of storage, the answers all have the same entropy, and the combinations of any single answer and any single message all have the same joint entropy. Moreover, for

such symmetrized codes, we also have

$$\beta = \beta_n, \quad n = 1, 2, \ldots, N. \tag{33}$$

As a consequence, we have

$$\beta \geq H(A_n^{(q)}), \quad q \in \mathcal{Q}_n, \quad n = 1, 2, \ldots, N. \tag{34}$$

The symmetry in terms of the joint entropy can be extended beyond that in (32), however in this work there is no need for such generality. The discussion in the following in effect utilizes the concept of answer variety introduced in [19], though we will not explicitly invoke the concept, and the discussion will be self-contained.

### B. A Subtle Dependence Structure

Our first step is to consider the relation among different answers:

1) First consider an answer at database-1 for an arbitrary but fixed $q_1 \in \mathcal{Q}_1$ which can be used to retrieve $W_1$, denoted as $X_1 = A_1^{(q_1)}$. Clearly there exists a query $q_2 \in \mathcal{Q}_2$ such that together with the answer $Y_1 = A_2^{(q_2)}$, the message $W_1$ can recovered, i.e., $H(W_1|X_1, Y_1) = 0$.

2) Because of the privacy constraint, the answer $X_1 = A_1^{(q_1)}$ from database-1 can be used, together with an query $q_2' \in \mathcal{Q}_2$, to retrieve $W_2$, i.e., with the answer $Y_2 = A_2^{(q_2')}$ we have $H(W_2|X_1, Y_2) = 0$. Note that the answer $Y_1$ and $Y_2$ are not necessarily distinct.

3) Continuing the same argument, an answer $X_2 = A_1^{(q_1')}$ must exist such that $H(W_2|X_2, Y_1) = 0$.

4) Finally, an answer $X_3 = A_1^{(q_1'')}$ must also exist such that $H(W_1|X_3, Y_2) = 0$.

Clearly, $(X_1, X_2, X_3, Y_1, Y_2)$ are functions of $W_1, W_2$, i.e., $H(X_1, X_2, X_3, Y_1, Y_2|W_1, W_2) = 0$.

Since the answers must come from the stored content, $H(X_1, X_2, X_3|S_1) = 0$, we must have

$$\alpha \geq H(S_1) \geq H(X_1, X_2, X_3), \tag{35}$$

and similarly

$$\alpha \geq H(S_2) \geq H(Y_1, Y_2). \tag{36}$$

This is the dependence structure and constraints in the original problem setting that we shall avail. Note that in this line of proof, the random key $\mathsf{F}$ is not playing any significant role, unlike in the proofs for Theorem 1 and 2.

### C. A Pseudo Message Technique

Our next step is to extend the random variable space, by introducing some random variables not in the original problem. The random variables $V_1, V_2$, referred to as the pseudo messages, are introduced into the setting using the so-called Markov coupling

$$(V_1, V_2) \leftrightarrow (Y_1, Y_2) \leftrightarrow (W_1, W_2, X_1, X_2, X_3). \tag{37}$$

Moreover, the two sets of random variables have the identical marginal distribution

$$(Y_1, Y_2, V_1, V_2) \sim (Y_1, Y_2, W_1, W_2). \tag{38}$$

Similarly, the second set of pseudo message random variables $(U_1, U_2)$ are also introduced such that

$$(U_1, U_2) \leftrightarrow (X_1, X_2, X_3) \leftrightarrow (W_1, W_2, Y_1, Y_2, V_1, V_2), \tag{39}$$

and the two sets of random variables have the identical marginal distribution

$$(X_1, X_2, X_3, U_1, U_2) \sim (X_1, X_2, X_3, W_1, W_2). \tag{40}$$

As a consequence, the extended set of random variables can be factorized as

$$P(W_1, W_2, X_1, X_2, X_3, Y_1, Y_2) \\ \cdot P(U_1, U_2|X_1, X_2, X_3) P(V_1, V_2|Y_1, Y_2). \tag{41}$$

The proof of Theorem 3 utilizes the symmetry, the Markov condition in the extended random variable space, as well as the encoding and decoding constraints. The basic idea is to bound or substitute the conditional entropy involving $(W_1, W_2, X_1, X_2, X_3, Y_1, Y_2)$ using conditional entropy involving subsets of $(U_1, U_2, V_1, V_2, W_1, W_2, X_1, X_2, X_3, Y_1, Y_2)$ in matching forms, and then to cancel terms using the identical distribution relations.

The proof technique of introducing pseudo messages can be viewed as being closely related to non-Shannon type inequalities, since all known non-Shannon type inequalites are essentially produced by introducing certain mirrored copies. This proof was obtained with the assistance of the computational tool that the author developed previously, which was found valuable in the investigation of the regenerating code problem [33] and the coded caching problem [35]. The main difference between the technique we use in this work and those seen in generating non-Shannon type inequalities is that instead of introducing a single-sided mirrored set, we introduce mirrors on two sides–one side being $(V_1, V_2)$, and the other being $(U_1, U_2)$–both of which through the Markov coupling.

## V. Conclusion

We investigated the fundamental tradeoff between the storage cost and the download cost in general private information retrieval systems. Several novel outer bounds are provided. On the one hand, we were able to confirm the folklore that when the messages are stored without any redundancy, the retrieval must download all the messages. On the other hand, when the code is PIR capacity-achieving, we establish the somewhat surprising result that the storage cost cannot be too much lower than storing all the messages at each database. Moreover, we show that for the two-message two-database case, a more elaborate pseudo-message technique can be used to derive a stronger outer bound. As an ongoing work, we are investigating more general outer bounds of the fundamental tradeoff between the storage cost and the download cost.

## REFERENCES

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

[2] O. Goldreich, H. Karloff, L. J. Schulman, and L. Trevisan, "Lower bounds for linear locally decodable codes and private information retrieval," in *Proceedings 17th IEEE Annual Conference on Computational Complexity*, 2002, pp. 175–183.

[3] H. Sun and S. A. Jafar, "On the capacity of locally decodable codes," *arXiv preprint arXiv:1812.05566*, 2018.

[4] ——, "Blind interference alignment for private information retrieval," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 560–564.

[5] ——, "The capacity of private information retrieval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

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

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

[8] ——, "The capacity of symmetric private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 322–329, 2018.

[9] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1945–1956, 2018.

[10] ——, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 1206–1219, 2018.

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

[12] ——, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al." *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[13] 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.

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

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

[16] ——, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Transactions on Information Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[17] Z. Chen, Z. Wang, and S. Jafar, "The capacity of private information retrieval with private side information," *IEEE Transactions on Information Theory*, to appear, 2020.

[18] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8222–8231, Dec. 2019.

[19] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7613–7627, 2019.

[20] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," *IEEE Transactions on Information Theory*, to appear, 2020.

[21] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," *arXiv preprint arXiv:1805.04104*, 2018.

[22] C. Tian, H. Sun, and J. Chen, "A Shannon-theoretic approach to the storage-retrieval tradeoff in PIR systems," in *2018 Proceedings of IEEE International Symposium on Information Theory (ISIT)*, Jun. 2018, pp. 1904–1908.

[23] S. Rao and A. Vardy, "Lower bound on the redundancy of PIR codes," *arXiv preprint arXiv:1605.01869*, 2016.

[24] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," *Proceedings of 2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2842–2846, Jun. 2015.

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

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

[27] K. Banawan, B. Arasli, and S. Ulukus, "Improved storage for efficient private information retrieval," in *2019 IEEE Information Theory Workshop (ITW)*, 2019.

[28] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell System Technical Journal*, vol. 59, pp. 1909–1921, Dec. 1980.

[29] Z. Zhang and R. W. Yeung, "A non-Shannon-type conditional inequality of information quantities," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1982–1986, 1997.

[30] R. Dougherty, C. Freiling, and K. Zeger, "Non-Shannon information inequalities in four random variables," *arXiv preprint arXiv:1104.3602*, 2011.

[31] E. Gürpınar and A. Romashchenko, "How to use undiscovered information inequalities: Direct applications of the copy lemma," *arXiv preprint arXiv:1901.07476*, 2019.

[32] C. Tian, "On the storage cost of private information retrieval," *arXiv preprint arXiv:1910.11973*, 2019.

[33] ——, "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, 2014.

[34] C. Tian, J. S. Plank, and B. Hurst, "An open-source toolbox for computer-aided investigation on the fundamental limits of information systems, version 0.1," *arXiv preprint arXiv:1910.08567*, 2019.

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