information

Article

A Shannon-Theoretic Approach to the Storage–Retrieval Trade-Off in PIR Systems

Chao Tian [1,*], Hua Sun [2] and Jun Chen [3]

[1] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77845, USA
[2] Department Electrical Engineering, University of North Texas, Denton, Texas 76203, USA
[3] Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada
* Correspondence: chao.tian@tamu.edu

Abstract: We consider the storage–retrieval rate trade-off in private information retrieval (PIR) systems using a Shannon-theoretic approach. Our focus is mostly on the canonical two-message two-database case, for which a coding scheme based on random codebook generation and the binning technique is proposed. This coding scheme reveals a hidden connection between PIR and the classic multiple description source coding problem. We first show that when the retrieval rate is kept optimal, the proposed non-linear scheme can achieve better performance over any linear scheme. Moreover, a non-trivial storage-retrieval rate trade-off can be achieved beyond space-sharing between this extreme point and the other optimal extreme point, achieved by the retrieve-everything strategy. We further show that with a method akin to the expurgation technique, one can extract a zero-error PIR code from the random code. Outer bounds are also studied and compared to establish the superiority of the non-linear codes over linear codes.

Keywords: capacity; information theory; multiple descriptions; privacy

check for updates

Citation: Tian, C.; Sun, H.; Chen, J. A Shannon-Theoretic Approach to the Storage–Retrieval Trade-Off in PIR Systems. *Information* **2023**, *14*, 44. https://doi.org/10.3390/info14010044

Academic Editor: Josiane Mothe

Received: 8 November 2022
Revised: 22 December 2022
Accepted: 5 January 2023
Published: 11 January 2023

1. Introduction

Private information retrieval (PIR) addresses the situation of storing K messages of L-bits each in N databases, with the requirement that the identity of any requested message must be kept private from any one (or any small subset) of the databases. The early works were largely computer science theoretic [1], where $L = 1$, and the main question is the scaling law of the retrieval rate in terms of (K, N).

The storage overhead in PIR systems has been studied in the coding and information theory community from several perspectives using mainly two problem formulations. Shah et al. [2] considered the problem when N is allowed to vary with L and K, and obtained some conclusive results. In a similar vein, for $L = 1$, Fazeli et al. [3] proposed a technique to convert any linear PIR code to a new one with low storage overhead by increasing N. Other notable results along this line can be found in [4–9].

An information theoretic formulation of the PIR problem was considered in [10], where L is allowed to increase, while (N, K) are kept fixed. Important properties on the trade-off between the storage rate and retrieval rate were identified in [10], and a linear code construction was proposed. In this formulation, even without any storage overhead constraint, characterizing the minimum retrieval rate in the PIR systems is nontrivial, and this capacity problem was settled in [11]. Tajeddine et al. [12] considered the capacity problem when the message is coded across the databases with a maximum-distance separable (MDS) code, which was later solved by Banawan and Ulukus [13]. Capacity-achieving code designs with optimal message sizes were given in [14,15]. Systems where servers can collude were considered in [16]. There have been various extensions and generalizations, and the recent survey article [17] provides a comprehensive overview on efforts following this information theoretic formulation.

In many existing works, the storage component and the PIR component are largely designed separately, usually by placing certain structural constraints on one of them, e.g., the MDS coding requirement for the storage component [13], or the storage is uncoded [18]; moreover, the code constructions are almost all linear. The few exceptions we are aware of are [19–21]. In this work, we consider the information theoretic formulation of the PIR problem, without placing any additional structural constraints on the two components, and explicitly investigate the storage–retrieval trade-off *region*. We mostly focus on the case $N = K = 2$ here since it provides the most important intuition; we refer to this as the $(2, 2)$ PIR system. Our approach naturally allows the joint design of the two components using either linear or *non-linear* schemes.

The work in [19] is of significant relevance to our work, where the storage overhead was considered in both single-round and multi-round PIR systems, when the retrieval rate must be optimal. Although multi-round PIR has the same capacity as single-round PIR, it was shown that at the minimum retrieval rate, a multi-round, ϵ-error, non-linear code can indeed break the storage performance barrier of an optimal single-round, zero error, linear code. The question of whether all the three differences are essential to overcome this barrier was left as an open question.

In this work, we show that a non-linear code is able to achieve better performance than the optimal linear code in the single-round zero-error $(2, 2)$ PIR system, over a range of the storage rates. This is accomplished by providing a Shannon-theoretic coding scheme based on random codebook generation and the binning technique. The proposed scheme at the minimum retrieval rate is conceptually simpler, and we present it as an explicit example. The general inner bound is then provided, and we show an improved trade-off can be achieved beyond space-sharing between the minimum retrieval rate code and the other optimal extreme point. By leveraging a method akin to the expurgation technique, we further show that one can extract a zero-error deterministic PIR code from the random ϵ-error PIR code. Outer bounds are also studied for both general codes and linear codes, which allow us to establish conclusively the superiority of non-linear codes over linear codes. Our work essentially answers the open question in [19], and shows that, in fact, only non-linearity is essential in breaking the aforementioned barrier.

A preliminary version of this work was presented first in part in [22]. In this updated article, we provide a more general random coding scheme, which reveals a hidden connection to the multiple description source coding problem [23]. Intuitively, we can view the retrieved message as certain partial reconstruction of the full set of messages, instead of a complete reconstruction of a single message. Therefore, the answers from the servers can be viewed as descriptions of the full set of messages, which are either stored directly at the servers or formed at the time of request, and the techniques seen in multiple description coding become natural in the PIR setting. Since the publication of the preliminary version [22], several subsequent efforts have been made in studying the storage–retrieval trade-off in the PIR setting, which provided stronger and more general information theoretic outer bounds and several new linear code constructions [20,21,24]. However, the Shannon-theoretic random coding scheme given in [22] remains the best-performing for the $(2, 2)$ case, which motivates us to provide the general coding scheme in this work and to make the connection to multiple description source coding more explicit. It is our hope that this connection may bring existing coding techniques for the multiple description problem to the study of the PIR problem.

2. Preliminaries

The problem we consider is essentially the same as that in [11], with the additional consideration on the storage overhead constraint at the databases. We provide a formal problem definition in the more traditional Shannon-theoretic language to facilitate subsequent treatment. Some relevant results on this problem are also reviewed briefly in this section.

2.1. Problem Definition

There are two independent messages, denoted as W_1 and W_2, in this system, each of which is generated uniformly at random in the finite field $\mathbb{F}_2^L$, i.e., each message is an L-bit sequence. There are two databases to store the messages, which are produced by two encoding functions operating on (W_1, W_2):

$$\phi_n : \mathbb{F}_2^L \times \mathbb{F}_2^L \to \mathbb{F}_2^{\alpha_n}, \quad n = 1, 2,$$

where α_n is the number of storage symbols at database-n, $n = 1, 2$, which is a deterministic function of L, i.e., we are using fixed length codes for storage. We write $S_1 = \phi_1(W_1, W_2)$ and $S_2 = \phi_2(W_1, W_2)$. When a user requests message-k, it generates two queries $(Q_1^{[k]}, Q_2^{[k]})$ to be sent to the two databases, randomly in the alphabet $\mathcal{Q} \times \mathcal{Q}$. Note that the joint distribution satisfies the condition

$$P_{W_1,W_2,Q_1^{[k]},Q_2^{[k]}} = P_{W_1,W_2} P_{Q_1^{[k]},Q_2^{[k]}}, \quad k = 1, 2, \tag{1}$$

i.e., the messages and the queries are independent. The marginal distributions P_{W_1,W_2} and $P_{Q_1^{[k]},Q_2^{[k]}}$, $k = 1, 2$, thus fully specify the randomness in the system.

After receiving the queries, the databases produce the answers to the query via a set of deterministic functions:

$$\varphi_n^{(q)} : \mathbb{F}_2^{\alpha_n} \to \mathbb{F}_2^{\beta_n^{(q)}}, \quad q \in \mathcal{Q}, \, n = 1, 2. \tag{2}$$

We also write the answers $A_n^{[k]} = \varphi_n^{(Q_n^{[k]})}(S_n)$, $n = 1, 2$. The user, with the retrieved information, wishes to reproduce the desired message through a set of decoding functions

$$\psi^{(k,q_1,q_2)} : \mathbb{F}_2^{\beta_1^{(q_1)}} \times \mathbb{F}_2^{\beta_2^{(q_2)}} \to \mathbb{F}_2^L. \tag{3}$$

The outputs of the functions $\hat{W}_k = \psi^{(k,Q_1^{[k]},Q_2^{[k]})}(A_1^{[k]}, A_2^{[k]})$ are essentially the retrieved messages. We require the system to retrieve the message correctly (zero error), i.e., $\hat{W}_k = W_k$ for $k = 1, 2$.

Alternatively, we can require the system to have a small error probability. Denote the average probability of coding error of a PIR code as

$$P_e = 0.5 \sum_{k=1,2} P_{W_1,W_2,Q_1^{[k]},Q_2^{[k]}}(W_k \neq \hat{W}_k). \tag{4}$$

An $(L, \alpha_1, \alpha_2, \beta_1, \beta_2)$ ϵ-error PIR code is defined similar as a (zero-error) PIR code, except that the correctness condition is replaced by the condition that the probability of error $P_e \leq \epsilon$.

Finally, the privacy constraint stipulates that the identical distribution condition must be satisfied:

$$P_{Q_n^{[1]},A_n^{[1]},S_n} = P_{Q_n^{[2]},A_n^{[2]},S_n}, \quad n = 1, 2. \tag{5}$$

Note that one obvious consequence is that $P_{Q_n^{[1]}} = P_{Q_n^{[2]}} \triangleq P_{Q_n}$, for $n = 1, 2$.

We refer to the code, which is specified by two probability distributions $P_{Q_1^{[k]},Q_2^{[k]}}$, $k = 1, 2$, and a valid set of coding functions $\{\phi_n, \varphi_n^{(q)}, \psi^{k,q_1,q_2}\}$ that satisfy both the correctness and privacy constraints, as an $(L, \alpha_1, \alpha_2, \beta_1, \beta_2)$ PIR code, where $\beta_n = \mathbb{E}_{Q_n}[\beta_n^{(Q_n)}]$, for $n = 1, 2$.

Definition 1. *A normalized storage–retrieval rate pair* $(\bar{\alpha}, \bar{\beta})$ *is achievable, if for any* $\epsilon > 0$ *and sufficiently large* L*, there exists an* $(L, \alpha_1, \alpha_2, \beta_1, \beta_2)$ *PIR code, such that*

$$L(\bar{\alpha} + \epsilon) \geq \frac{1}{2}(\alpha_1 + \alpha_2), \ L(\bar{\beta} + \epsilon) \geq \frac{1}{2}(\beta_1 + \beta_2). \tag{6}$$

The collection of the achievable normalized storage–retrieval rate pair $(\bar{\alpha}, \bar{\beta})$ *is the achievable storage–retrieval rate region, denoted as* $\mathcal{R}$*.*

Unless explicitly stated, the rate region $\mathcal{R}$ is used for the zero-error PIR setting. In the definition above, we used the average rates $(\bar{\alpha}, \bar{\beta})$ across the databases instead of the individual rate vectors $\frac{1}{n}(\alpha_1, \alpha_2, \mathbb{E}_{Q_1}[\beta_1^{(Q_1)}], \mathbb{E}_{Q_2}[\beta_2^{(Q_2)}])$. This can be justified using the following lemma.

Lemma 1. *If an* $(L, \alpha_1, \alpha_2, \beta_1, \beta_2)$ *PIR code exists, then a* $(2L, \alpha, \alpha, \beta, \beta)$ *PIR code exists, where*

$$\alpha = \alpha_1 + \alpha_2, \quad \beta = \beta_1 + \beta_2. \tag{7}$$

This lemma can essentially be proved by a space-sharing argument, the details of which can be found in [19]. The following lemma is also immediate using a conventional space-sharing argument.

Lemma 2. *The region* $\mathcal{R}$ *is convex.*

2.2. Some Relevant Known Results

The capacity of a general PIR system with K messages and N databases is identified in [11] as

$$C = \frac{1 - 1/N}{1 - 1/N^K}, \tag{8}$$

which in our definition corresponds to the case when $\bar{\beta}$ is minimized, and the proposed linear code achieves $(\bar{\alpha}, \bar{\beta}) = (K, (1 - 1/N^K)/(N - 1))$. The capacity of MDS-code PIR systems was established in [13]. In the context of the storage–retrieval trade-off, this result can be viewed as providing the achievable trade-off pairs

$$(\bar{\alpha}, \bar{\beta}) = \left(t, \frac{1 - t^K/N^K}{N - t}\right), t = 1, 2, \ldots, N. \tag{9}$$

However, when specialized to the $(2, 2)$ PIR problem, this does not provide any improvement over the space-sharing strategy between the trivial code of retrieval-everything and the code in [11]. By specializing the code in [11], it was shown in [19] that for the $(2, 2)$ PIR problem, at the minimal retrieval value $\bar{\beta} = 0.75$, the storage rate $\bar{\alpha}_l = 1.5$ is achievable using a single-round, zero-error linear code, and in fact, it is the optimal storage rate that any single-round, zero-error linear code can achieve.

One of the key observations in [19] is that a special coding structure appears to be the main difficulty in the $(2, 2)$ PIR setting, which is illustrated in Figure 1. Here, message W_1 can be recovered from either (X_1, Y_1) or (X_2, Y_2), and message W_2 can be recovered from either (X_1, Y_2) or (X_2, Y_1); (X_1, X_2) is essentially S_1 and is stored at database-1, and (Y_1, Y_2) is essentially S_2 and is stored at database-2. It is clear that we can use the following strategy to satisfy the privacy constraint: when message W_1 is requested, with probability $1/2$, the user queries for either (X_1, Y_1) or (X_2, Y_2); for message 2, with probability $1/2$, the user queries for either (X_1, Y_2) or (X_2, Y_1). More precisely, the following probability distribution $P_{Q_1^{[1]}, Q_2^{[1]}}$ and $P_{Q_1^{[2]}, Q_2^{[2]}}$ can be used:

$$P_{Q_1^{[1]},Q_2^{[1]}} = \begin{cases} 0.5 & (Q_1^{[1]}, Q_2^{[1]}) = (11) \\ 0.5 & (Q_1^{[1]}, Q_2^{[1]}) = (22) \end{cases}, \tag{10}$$

and

$$P_{Q_1^{[2]},Q_2^{[2]}} = \begin{cases} 0.5 & (Q_1^{[2]}, Q_2^{[2]}) = (12) \\ 0.5 & (Q_1^{[2]}, Q_2^{[2]}) = (21) \end{cases}. \tag{11}$$

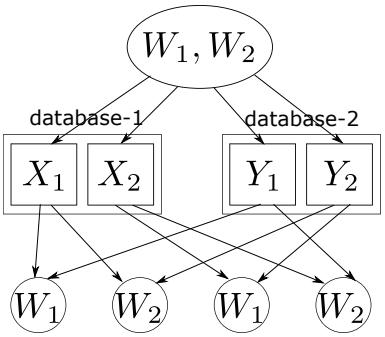

Figure 1. A possible coding structure.

2.3. Multiple Description Source Coding

The multiple description source coding problem [23] considers compressing a memoryless source S into a total of M descriptions, i.e., M compressed bit sequences such that the combinations of any subset of these descriptions can be used to reconstruct the source S to guarantee certain quality requirements. The motivation of this problem is mainly to address the case when packets can be dropped randomly on a communication network.

Denote the coding rate for each description as R_i, $i = 1, 2, \ldots, M$. A coding scheme was proposed in [25], which leads to the following rate region. Let $U_1, U_2, \ldots, U_M$ be M random variables jointly distributed with S, then the following rates $(R_1, R_2, \ldots, R_M)$ and distortions $(D_{\mathcal{A}}, \mathcal{A} \subseteq \{1, 2, \ldots, M\})$ are achievable:

$$\sum_{i \in \mathcal{A}} R_i \geq \sum_{i \in \mathcal{A}} H(U_i) - H(\{U_i, i \in \mathcal{A}\}|S), \quad \mathcal{A} \subseteq \{1, 2, \ldots, M\}, \tag{12}$$

$$D_{\mathcal{A}} \geq \mathbb{E}[d(S, f_{\mathcal{A}}(U_i, i \in \mathcal{A}))], \quad \mathcal{A} \subseteq \{1, 2, \ldots, M\}. \tag{13}$$

Here, $f_{\mathcal{A}}$ is a reconstruction mapping from the random variables $\{U_i, i \in \mathcal{A}\}$ to the reconstruction domain, $d(\cdot, \cdot)$ is a distortion metric that is used to measure the distortion, and $D_{\mathcal{A}}$ is the distortion achievable using the descriptions in the set $\mathcal{A}$. Roughly speaking, the coding scheme requires generating approximately 2^{nR_i} length-n codewords in an i.i.d. manner using the marginal distribution U_i for each $i = 1, 2, \ldots, M$, and the rate constraints ensure that when n is sufficiently large, with overwhelming probability there is a tuple of M codewords $(u_1^n, u_2^n, \ldots, u_M^n)$, one in each codebook constructed earlier, that are jointly typical with the source vector S^n. In this coding scheme, the descriptions are simply the codeword indices of these codewords in these codebooks. For a given joint distribution $(S, U_1, U_2, \ldots, U_M)$, we refer to the rate region in (12) as the MD rate region $\mathcal{R}_{MD}(S, U_1, U_2, \ldots, U_M)$, and the corresponding random code construction the MD codebooks associated with $(S, U_1, U_2, \ldots, U_M)$.

The binning technique [26] can be applied in the multiple description problem to provide further performance improvements, particularly when not all the combinations of the descriptions are required to satisfy certain performance constraints, but only a subset of them are; this technique was previously used in [27,28] for this purpose. Assume that only the subsets of descriptions $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_T \subseteq \{1, 2, \ldots, M\}$ have distortion requirements associated with the reconstructions using these descriptions, which are denoted as $D_{\mathcal{A}_i}$, $i = 1, 2, \ldots, T$. Consider the MD codebooks associated with $(S, U_1, U_2, \ldots, U_M)$ at rates $(R_1', R_2', \ldots, R_M') \in \mathcal{R}_{MD}(S, U_1, U_2, \ldots, U_M)$, then assign the codewords in the i-th

codebook uniformly at random into 2^{nR_i} bins with $0 \le R_i \le R'_i$. The coding rates and distortions that satisfy the following constraints simultaneously for all $\mathcal{A}_i, i = 1, 2, \ldots, T$ are achievable:

$$\sum_{j\in\mathcal{J}} (R'_j - R_j) \le \sum_{j\in\mathcal{J}} H(U_j) - H\Big(\{U_j, j \in \mathcal{J}\}\Big|\Big\{U_{j'}, j' \in \mathcal{A}_i \setminus \mathcal{J}\Big\}\Big), \quad \forall \mathcal{J} \subseteq \mathcal{A}_i, \tag{14}$$

$$D_{\mathcal{A}_i} \ge \mathbb{E}[d(S, f_{\mathcal{A}_i}(U_j, j \in \mathcal{A}_i))]. \tag{15}$$

We denote the collection of such rate vectors $(R_1, R_2, \ldots, R_M, R'_1, R'_2, \ldots, R'_M)$ as $\mathcal{R}^*_{MD}((S, U_1, U_2, \ldots, U_M), (\{U_j, j \in \mathcal{A}_i\}, i = 1, 2, \ldots, T))$, and refer to the corresponding codebooks as the MD* codebooks associated with the random variables $(S, U_1, U_2, \ldots, U_M)$ and the reconstruction sets $(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_T)$.

3. A Special Case: Slepian–Wolf Coding for Minimum Retrieval Rate

In this section, we consider the minimum-retrieval-rate case, and show that non-linear and Shannon-theoretic codes are beneficial. We will be rather cavalier here and ignore some details, in the hope of better conveyance of the intuition. In particular, we ignore the asymptotic-zero probability of error that is usually associated with a random coding argument, but this will be addressed more carefully in Section 4.

Let us rewrite the L-bit messages as

$$W_k = (V_k[1], \ldots, V_k[L]) \triangleq V_k^L, \quad k = 1, 2. \tag{16}$$

The messages can be viewed as being produced from a discrete memoryless source $P_{V_1,V_2} = P_{V_1} \cdot P_{V_2}$, where V_1 and V_2 are independent uniform-distributed Bernoulli random variables.

Consider the following auxiliary random variables:

$$\begin{aligned} X_1 &\triangleq V_1 \wedge V_2, \quad X_2 \triangleq (\neg V_1) \wedge (\neg V_2), \\ Y_1 &\triangleq V_1 \wedge (\neg V_2), \quad Y_2 \triangleq (\neg V_1) \wedge V_2, \end{aligned} \tag{17}$$

where $\neg$ is the binary negation, and $\wedge$ is the binary "and" operation. This particular distribution satisfies the coding structure depicted in Figure 1, with (V_1, V_2) taking the role of (W_1, W_2), and the relation is non-linear. The same distribution was used in [19] to construct a multiround PIR code. This non-linear mapping appears to allow the resultant code to be more efficient than linear codes.

We wish to store (X_1^L, X_2^L) at the first database in a lossless manner, however, store only certain necessary information regarding Y_1^L and Y_2^L to facilitate the recovery of W_1 or W_2. For this purpose, we will encode the message as follows:

- At database-1, compress and store (X_1^L, X_2^L) losslessly;
- At database-2, encode Y_1^L using a Slepian–Wolf code (or more precisely Sgarro's code with uncertainty side information [29]), with either X_1^L or X_2^L at the decoder, whose resulting code index is denoted as C_{Y_1}; encode Y_2^L in the same manner, independent of Y_1^L, whose code index is denoted as C_{Y_2}.

It is clear that for database-1, we need roughly $\bar{\alpha}_1 = H(X_1, X_2)$. At database-2, in order to guarantee successful decoding of the Slepian-Wolf code, we can chose roughly

$$\begin{aligned} \bar{\alpha_2} &= \max(H(Y_1|X_1), H(Y_1|X_2)) + \max(H(Y_2|X_1), H(Y_2|X_2)) \\ &= 2H(Y_1|X_1), \end{aligned} \tag{18}$$

where the second equality is due to the symmetry in the probability distribution. Thus we find that this code achieves

$$\begin{aligned}\bar{\alpha}_{nl} &= 0.5[H(X_1, X_2) + 2H(Y_1|X_1)] \\ &= 0.75 + 0.75H(1/3, 2/3) \\ &= 0.25 + 0.75\log_2 3 \approx 1.4387. \end{aligned} \tag{19}$$

The retrieval strategy is immediate from the coding structure in Figure 1, with $(V_1^L, V_2^L, X_1^L, X_2^L, C_{Y_1}, C_{Y_2})$ serving the roles of $(W_1, W_2, X_1, X_2, Y_1, Y_2)$, and thus indeed the privacy constraint is satisfied. The retrieval rates are roughly as follows:

$$\bar{\beta}_1^{(1)} = \bar{\beta}_1^{(2)} = H(X_1) = H(X_2), \tag{20}$$

$$\bar{\beta}_2^{(1)} = \bar{\beta}_2^{(2)} = H(Y_1|X_1), \tag{21}$$

implying

$$\bar{\beta} = 0.5[H(X_1) + H(Y_1|X_1)] = 0.5H(Y_1, X_1) = 0.75.$$

Thus, at the optimal retrieval rate $\bar{\beta} = 0.75$, we have

$$\bar{\alpha}_l = 1.5 \text{ vs. } \bar{\alpha}_{nl} \approx 1.4387, \tag{22}$$

and clearly the proposed non-linear Shannon-theoretic code is able to perform better than the optimal linear code. We note that it was shown in [19] by using a multround approach, the storage rate $\bar{\alpha}$ can be further reduced; however, this issue is beyond the scope of this work. In the rest of the paper, we build on the intuition in this special case to generalize and strengthen the coding scheme.

4. Main Result

4.1. A General Inner Bound

We first present a general inner bound to the storage–retrieval trade-off region. Let (V_1, V_2) be independent random variables uniformly distributed on $\mathbb{F}_2^t \times \mathbb{F}_2^t$. Define the region $\mathcal{R}_{in}^{(t)}$ to be the collection of $(\bar{\alpha}, \bar{\beta})$ pairs for which there exist random variables $(X_0, X_1, X_2, Y_1, Y_2)$ jointly distributed with (V_1, V_2) such that the following hold:

1. There exist deterministic functions $f_{1,1}$, $f_{1,2}$, $f_{2,1}$, and $f_{2,2}$ such that

$$V_1 = f_{1,1}(X_0, X_1, Y_1) = f_{2,2}(X_0, X_2, Y_2), \quad V_2 = f_{1,2}(X_0, X_1, Y_2) = f_{2,1}(X_0, X_2, Y_1); \tag{23}$$

2. There exist non-negative coding rates

$$\begin{aligned}&(\beta_1^{(0)}, \beta_1^{(1)}, \beta_1^{(2)}, \beta_2^{(1)}, \beta_2^{(2)}, \gamma_1^{(0)}, \gamma_1^{(1)}, \gamma_1^{(2)}, \gamma_2^{(1)}, \gamma_2^{(2)}) \\ &\quad \in \mathcal{R}_{MD}^*(((V_1, V_2), X_0, X_1, X_2, Y_1, Y_2), \\ &\qquad (\{X_0, X_1, Y_1\}, \{X_0, X_1, Y_2\}, \{X_0, X_2, Y_1\}, \{X_0, X_2, Y_2\})); \end{aligned} \tag{24}$$

3. There exist non-negative storage rates $(\alpha_1^{(0)}, \alpha_1^{(1)}, \alpha_1^{(2)}, \alpha_2^{(1)}, \alpha_2^{(2)})$ such that

$$\alpha_1^{(0)} \le \beta_1^{(0)}, \alpha_1^{(1)} \le \beta_1^{(1)}, \alpha_1^{(2)} \le \beta_1^{(2)}, \alpha_2^{(1)} \le \beta_2^{(1)}, \alpha_2^{(2)} \le \beta_2^{(2)}, \tag{25}$$

and if

$$\gamma_1^{(0)} - \beta_1^{(0)} + \gamma_1^{(1)} - \beta_1^{(1)} + \gamma_1^{(2)} - \beta_1^{(2)} < H(X_1) + H(X_2) + H(X_3) - H(X_0, X_1, X_2), \tag{26}$$

choose

$$(\alpha_1^{(0)}, \alpha_1^{(1)}, \alpha_1^{(2)}, \gamma_1^{(0)}, \gamma_1^{(1)}, \gamma_1^{(2)}) \in \mathcal{R}^*_{MD}(((V_1, V_2), X_0, X_1, X_2), (\{X_0, X_1, X_2\})); \quad (27)$$

otherwise, choose $(\alpha_1^{(0)}, \alpha_1^{(1)}, \alpha_1^{(2)}) = (\beta_1^{(0)}, \beta_1^{(1)}, \beta_1^{(2)})$. Similarly, if

$$\gamma_2^{(1)} - \beta_2^{(1)} + \gamma_2^{(2)} - \beta_2^{(2)} < I(Y_1; Y_2), \quad (28)$$

choose

$$(\alpha_2^{(1)}, \alpha_2^{(2)}, \gamma_2^{(1)}, \gamma_2^{(2)}) \in \mathcal{R}^*_{MD}(((V_1, V_2), Y_1, Y_2), (\{Y_1, Y_2\})), \quad (29)$$

otherwise $(\alpha_2^{(1)}, \alpha_2^{(2)}) = (\beta_1^{(1)}, \beta_1^{(2)})$;

4. The normalized average retrieval and storage rates

$$2t\bar{\alpha} \geq \alpha_1^{(0)} + \alpha_1^{(1)} + \alpha_1^{(2)} + \alpha_2^{(1)} + \alpha_2^{(2)}, \quad (30)$$

$$4t\bar{\beta} \geq 2\beta_1^{(0)} + \beta_1^{(1)} + \beta_1^{(2)} + \beta_2^{(1)} + \beta_2^{(2)}. \quad (31)$$

Then, we have the following theorem.

Theorem 1. $\mathcal{R}_{in}^{(t)} \subseteq \mathcal{R}$.

We can, in fact, potentially enlarge the achievable region by taking $\cup_{t=1}^{\infty} \mathcal{R}_{in}^{(t)}$. However, unless $\mathcal{R}_{in}^{(t+1)} \subseteq \mathcal{R}_{in}^{(t)}$ for all $t \geq 1$, the region $\cup_{t=1}^{\infty} \mathcal{R}_{in}^{(t)}$ is even more difficult to characterize. Nevertheless, for each fixed t, we can identify inner bounds by specifying a feasible set of random variables X_0, X_1, X_2, Y_1, Y_2.

Instead of directly establishing this theorem, we shall prove the following theorem which establishes the existence of a PIR code with diminishing error probability, and then use an expurgation technique to extract a zero-error PIR code.

Theorem 2. *Consider any* $(\bar{\alpha}, \bar{\beta}) \in \mathcal{R}_{in}^{(t)}$. *For any* $\epsilon > 0$ *and sufficiently large* L, *there exists an* $(L, L(\bar{\alpha} + \epsilon), L(\bar{\alpha} + \epsilon), L(\bar{\beta} + \epsilon), L(\bar{\beta} + \epsilon))$ ϵ*-error PIR code with the query distribution given in (10) and (11).*

The key observation to establish this theorem is that there are five descriptions in this setting; however, the retrieval and storage place different constraints on different combination of descriptions, and some descriptions can, in fact, be stored, recompressed, and then retrieved. Such compression and recompression may lead to storage savings. The description based on X_0 can be viewed as some common information to X_1 and X_2, which allows us to trade-off the storage and retrieval rates.

Proof of Theorem 2. *Codebook generation:* Codebooks are built using the MD codebooks based on the distribution $((V_1, V_2), X_0, X_1, X_2, Y_1, Y_2)$.

Storage codes: The bin indices of the codebooks are stored in the two servers: those of X_0, X_1, and X_2 are stored at server-1 at rates $\alpha_1^{(0)}$, $\alpha_1^{(1)}$, and $\alpha_1^{(2)}$, respectively; those of Y_1 and Y_2 are stored at server-2 at rates $\alpha_2^{(1)}$ and $\alpha_2^{(2)}$. Note that at such rates, the codewords for X_0, X_1, and X_2 can be recovered jointly with overwhelming probability, while those for Y_1 and Y_2 can also be recovered jointly with overwhelming probability.

Retrieval codes: A different set of bin indices of the codebooks are retrieved during the retrieval process, again based on the MD* codebooks: those of X_0, X_1, and X_2 are retrieved at server-1 at rates $\beta_1^{(0)}$, $\beta_1^{(1)}$, and $\beta_1^{(2)}$, respectively; those of Y_1 and Y_2 are retrieved at server-2 at rates $\beta_2^{(1)}$ and $\beta_2^{(2)}$. Note that at such rates, the codewords of X_0, X_1, and Y_1 can

be jointly recovered such that using the three corresponding codewords, the required V_1 source vector can be recovered with overwhelming probability. Similarly, the three retrieval patterns of $(X_0, X_1, Y_2) \to V_2$, $(X_0, X_2, Y_1) \to V_2$, and $(X_0, X_2, Y_2) \to V_2$ will succeed with overwhelming probabilities.

Storage and retrieval rates: The rates can be computed straightforwardly, after normalization by the parameter t. □

Next we use it to prove Theorem 1.

Proof of Theorem 1. Given an $\epsilon > 0$, according to Proposition 2, we can find an $(L, L(\bar{\alpha} + \epsilon), L(\bar{\alpha} + \epsilon), L(\bar{\beta} + \epsilon), L(\bar{\beta} + \epsilon))$ ϵ-error PIR code for some sufficient large L. The probability of error of this code can be rewritten as

$$P_e = 0.5 \sum_{k=1,2} \sum_{(w_1,w_2)} 2^{-2L} P_{Q_1^{[k]}, Q_2^{[k]} | (w_1,w_2)} (w_k \neq \hat{W}_k).$$

For a fixed (w_1, w_2) pair, denote the event that there exists a $(q_1, q_2) \in \{(11), (22)\}$, i.e., when $(Q_1^{[1]}, Q_2^{[1]}) = (q_1, q_2)$, such that $\hat{w}_1 \neq w_1$ as $E^{(1)}_{w_1,w_2}$, and there exists a $(q_1, q_2) \in \{(12), (21)\}$ such that $\hat{w}_2 \neq w_2$ as $E^{(2)}_{w_1,w_2}$. Since $(Q_1^{[k]}, Q_2^{[k]})$ is independent of (W_1, W_2), if $P(E^{(k)}_{w_1,w_2}) \neq 0$, we must have $P(E^{(k)}_{w_1,w_2}) \geq 0.5$. It follows that

$$P_e \geq 0.25 \sum_{(w_1,w_2)} 2^{-2L} \mathbf{1}(E^{[1]}_{w_1,w_2} \cup E^{[2]}_{w_1,w_2}), \tag{32}$$

where $(\cdot)$ is the indicator function. This implies that for any $\epsilon \leq 0.125$, there are at most 2^{2L-1} pairs of (w_1, w_2) that will induce any coding error. We can use any 2^{2L-2} of the remaining 2^{2L-1} pairs of L-bit sequence pairs to instead store a pair of $(L-1)$-bit messages, through an arbitrary but fixed one-to-one mapping. This new code has a factor of $1 + 1/(L-1)$ increase in the normalized coding rates, which is negligible when L is large. Thus a zero-error PIR code is found with the same normalized rates as the ϵ-error code asymptotically, and this completes the proof. □

4.2. Outer Bounds

We next turn our attention to the outer bounds for $\mathcal{R}$, summarized in the following theorem.

Theorem 3. *Any $(\bar{\alpha}, \bar{\beta}) \in \mathcal{R}$ must satisfy*

$$\bar{\beta} \geq 0.75, \quad \bar{\alpha} + \bar{\beta} \geq 2, \quad 3\bar{\alpha} + 8\bar{\beta} \geq 10. \tag{33}$$

Moreover, if $(\bar{\alpha}, \bar{\beta}) \in \mathcal{R}$ can be achieved by a linear code, it must satisfy

$$\bar{\alpha} + 6\bar{\beta} \geq 6. \tag{34}$$

The inequality $\bar{\beta} \geq 0.75$ follows from [11], while the two other bounds in (33) were proved in [24]. Therefore, we only need to prove (34).

Proof of Theorem 3. Following [19], we make the following simplifying assumptions that have no loss of generality. Define $\mathbb{Q} = \{Q_1^{[1]}, Q_1^{[2]}, Q_2^{[1]}, Q_2^{[2]}\}$.

1. $Q_1^{[1]} = Q_1^{[2]} \Rightarrow A_1^{[1]} = A_1^{[2]}$, (35)
2. $H(A_1^{[1]}|\mathbb{Q}) = H(A_2^{[1]}|\mathbb{Q}) = H(A_2^{[2]}|\mathbb{Q}), \quad H(S_1) = H(S_2)$ (36)

   $\Rightarrow H(A_1^{[1]}|\mathbb{Q}) \leq \beta \leq (\bar{\beta} + \epsilon)L, \quad H(S_2) \leq \alpha \leq (\bar{\alpha} + \epsilon)L.$ (37)

Assumption 1 states that the query to the first database is the same regardless of the desired message index. This is justified by the privacy condition that the query to one database is independent of the desired message index. Assumption 2 states that the scheme is symmetric after the symmetrization operation in Lemma 1 (the proof is referred to Theorem 3 in [19]). Then, (37) follows from the fact that to describe $S_2, A_1^{[1]}$, the number of bits needed cannot be less than the entropy value, and Definition 1.

In the following, we use (c) to refer to the correctness condition, (i) to refer to the constraint that queries are independent of the messages, (a) to refer to the constraint that answers are deterministic functions of the storage variables and corresponding queries, and (p) to refer to the privacy condition.

From $A_1^{[1]}, A_2^{[1]}, \mathbb{Q}$, we can decode W_1.

$$\begin{aligned} H(A_1^{[1]}, A_2^{[1]}|W_1, \mathbb{Q}) &= H(A_1^{[1]}, A_2^{[1]}, W_1|\mathbb{Q}) - H(W_1|\mathbb{Q}) && (38)\\ &\overset{(c)(i)}{=} H(A_1^{[1]}, A_2^{[1]}|\mathbb{Q}) - L && (39)\\ &\overset{(36)}{\leq} 2H(A_1^{[1]}|\mathbb{Q}) - L. && (40) \end{aligned}$$

Next, consider Ingleton's inequality.

$$\begin{aligned} I(A_2^{[1]}; A_2^{[2]}|\mathbb{Q}) &\leq I(A_2^{[1]}; A_2^{[2]}|W_1, \mathbb{Q}) + I(A_2^{[1]}; A_2^{[2]}|W_2, \mathbb{Q}) && (41)\\ &= 2I(A_2^{[1]}; A_2^{[2]}|W_1, \mathbb{Q}) && (42)\\ &= 2\big(H(A_2^{[1]}|W_1, \mathbb{Q}) + H(A_2^{[2]}|W_1, \mathbb{Q}) - H(A_2^{[1]}, A_2^{[2]}|W_1, \mathbb{Q})\big) && (43)\\ &\overset{(p)}{=} 2\big(2H(A_2^{[1]}|W_1, \mathbb{Q}) - H(A_2^{[1]}, A_2^{[2]}|W_1, \mathbb{Q})\big) && (44)\\ &\leq 2\big(2H(A_2^{[1]}|W_1, \mathbb{Q}) + H(A_1^{[1]}, A_2^{[1]}|W_1, \mathbb{Q})\\ &\quad - H(A_1^{[1]}, A_2^{[1]}, A_2^{[2]}|W_1, \mathbb{Q}) - H(A_2^{[1]}|W_1, \mathbb{Q})\big) && (45)\\ &\overset{(c)(35)}{=} 2\big(H(A_2^{[1]}|W_1, \mathbb{Q}) + H(A_1^{[1]}, A_2^{[1]}|W_1, \mathbb{Q})\\ &\quad - H(A_1^{[1]}, A_2^{[1]}, A_2^{[2]}, W_2|W_1, \mathbb{Q})\big) && (46)\\ &\overset{(i)}{\leq} 2\big(2H(A_1^{[1]}, A_2^{[1]}|W_1, \mathbb{Q}) - H(W_2)\big) && (47)\\ &\overset{(40)}{\leq} 2\big(2(2H(A_1^{[1]}|\mathbb{Q}) - L) - L\big) && (48) \end{aligned}$$

where (42) follows from the observation that the second term can be bounded using the same method as that bounds the first term by switching the message index. A more detailed derivation of (44) appears in (79) of [19]; (45) is due to the sub-modularity of entropy.

Note that

$$\begin{aligned} I(A_2^{[1]}; A_2^{[2]}|\mathbb{Q}) &= H(A_2^{[1]}|\mathbb{Q}) + H(A_2^{[2]}|\mathbb{Q}) - H(A_2^{[1]}, A_2^{[2]}|\mathbb{Q}) && (49)\\ &\overset{(36)}{\geq} 2H(A_1^{[1]}|\mathbb{Q}) - (\bar{\alpha} + \epsilon)L && (50) \end{aligned}$$

where in (50), and the second term is bounded as follows:

$$H(A_2^{[1]}, A_2^{[2]}|\mathbb{Q}) \leq H(A_2^{[1]}, A_2^{[2]}, S_2|\mathbb{Q}) \overset{(a)}{=} H(S_2|\mathbb{Q}) \overset{(37)}{\leq} (\bar{\alpha} + \epsilon)L. \quad (51)$$

Combining (48) and (50), we have

$$\begin{aligned} & 2H(A_1^{[1]}|\mathbb{Q})/L - (\bar{\alpha}+\epsilon) \geq 2(4H(A_1^{[1]}|\mathbb{Q})/L - 3) \\ \Rightarrow \quad & \bar{\alpha}+\epsilon+6H(A_1^{[1]}|\mathbb{Q})/L \geq 6 \qquad (52) \\ \overset{(37)}{\Rightarrow} \quad & \bar{\alpha}+6\bar{\beta} \geq 6. \qquad (53) \end{aligned}$$

The proof is complete. □

4.3. Specialization of the Inner Bound

The inner bound given in Theorem 1 is general but more involved, and we can specialize it in multiple ways in order to simplify it. One particularly interesting approach is as follows. Define the region $\tilde{\mathcal{R}}_{in}^{(t)}$ to be the collection of $(\bar{\alpha}, \bar{\beta})$ pairs such that there exists random variables $(X_0, X_1, X_2, Y_1, Y_2)$ jointly distributed with (V_1, V_2) such that the following hold:

1. The distribution factorizes as follows

$$P_{V_1,V_2,X_0,X_1,X_2,Y_1,Y_2} = P_{V_1,V_2}P_{X_0|V_1,V_2}P_{X_1|V_1,V_2}P_{X_2|V_1,V_2}P_{Y_1|V_1,V_2}P_{Y_2|V_1,V_2};$$

2. There exist deterministic functions $f_{1,1}$, $f_{1,2}$, $f_{2,1}$, and $f_{2,2}$ such that

$$V_1 = f_{1,1}(X_0, X_1, Y_1) = f_{2,2}(X_0, X_2, Y_2), \qquad (54)$$
$$V_2 = f_{1,2}(X_0, X_1, Y_2) = f_{2,1}(X_0, X_2, Y_1); \qquad (55)$$

3. A set of rates

$$\gamma_1^{(0)} = I(V_1, V_2; X_0),\ \gamma_1^{(1)} = I(V_1, V_2; X_1),\ \gamma_1^{(2)} = I(V_1, V_2; X_2), \qquad (56)$$
$$\gamma_2^{(1)} = I(V_1, V_2; Y_1),\ \gamma_2^{(2)} = I(V_1, V_2; Y_2), \qquad (57)$$
$$\beta_1^{(0)} = \gamma_1^{(0)},\ \beta_1^{(1)} = I(V_1, V_2; X_1|X_0),\ \beta_1^{(2)} = I(V_1, V_2; X_2|X_0), \qquad (58)$$
$$\beta_2^{(1)} = \max(I(V_1, V_2; Y_1|X_0, X_1), I(V_1, V_2; Y_1|X_0, X_2)), \qquad (59)$$
$$\beta_2^{(2)} = \max(I(V_1, V_2; Y_2|X_0, X_1), I(V_1, V_2; Y_2|X_0, X_2)), \qquad (60)$$

and $(\alpha_1^{(0)} = \gamma_1^{(0)}, \alpha_1^{(1)}, \alpha_1^{(2)}, \alpha_2^{(1)}, \alpha_2^{(2)})$ as defined in item 3 for the general region $\mathcal{R}^{(t)}$;

4. The normalized average retrieval and storage rates

$$2t\bar{\alpha} \geq \alpha_1^{(0)} + \alpha_1^{(1)} + \alpha_1^{(2)} + \alpha_2^{(1)} + \alpha_2^{(2)}, \qquad (61)$$
$$4t\bar{\beta} \geq 2\beta_1^{(0)} + \beta_1^{(1)} + \beta_1^{(2)} + \beta_2^{(1)} + \beta_2^{(2)}. \qquad (62)$$

Then we have the following corollary.

Corollary 1. $\tilde{\mathcal{R}}_{in}^{(t)} \subseteq \mathcal{R}$.

This inner bound is illustrated together with the outer bounds in Figure 2.

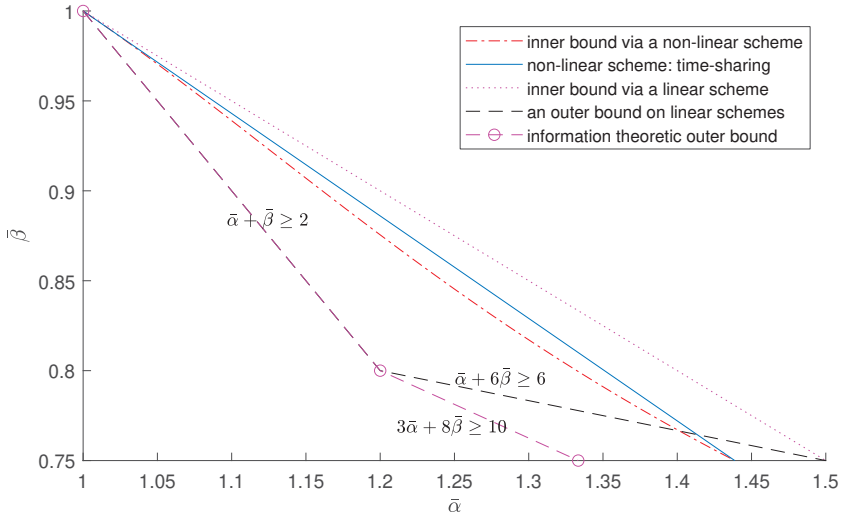

Figure 2. Illustration of inner bounds and outer bounds.

Proof. The main difference from Theorem 1 is in the special dependence structure of $(X_0, X_1, X_2, Y_1, Y_2)$ jointly distributed with (V_1, V_2), i.e., the Markov structure. We verify that the rate assignments satisfy all the constraints in Theorem 1. Due to the special dependence structure of $(X_0, X_1, X_2, Y_1, Y_2)$ jointly distributed with (V_1, V_2), it is straightforward to verify that

$$(\gamma_1^{(0)}, \gamma_1^{(1)}, \gamma_1^{(2)}, \gamma_2^{(1)}, \gamma_2^{(2)}) \in \mathcal{R}_{MD}((V_1, V_2), X_0, X_1, X_2, Y_1, Y_2).$$

We next verify that (24) holds with the choice given above. Due to the symmetry in the structure, we only need to confirm one subset of random variables, i.e., $\{X_0, X_1, Y_1\}$, and the three other subsets $\{X_0, X_1, Y_2\}$, $\{X_0, X_2, Y_1\}$, and $\{X_0, X_2, Y_2\}$ follow similarly. There are a total of 7 conditions in the form of (14) associated with this subset $\{X_0, X_1, Y_1\}$. Notice that

$$\gamma_1^{(0)} - \beta_1^{(0)} = 0,\ \gamma_1^{(1)} - \beta_1^{(1)} = I(X_1; X_0),\ \gamma_2^{(2)} - \beta_2^{(2)} \leq I(Y_1; X_0, X_1),$$

which in fact confirm three of the seven conditions when $\mathcal{J}$ is a singleton. Next, when $\mathcal{J}$ has two elements, we verify that

$$\begin{aligned}
\gamma_1^{(0)} - \beta_1^{(0)} + \gamma_1^{(1)} - \beta_1^{(1)} &= I(X_1; X_0) = H(X_0) + H(X_1) - H(X_0, X_1) \\
&\leq H(X_0) + H(X_1) - H(X_0, X_1|Y_1), && (63)\\
\gamma_1^{(0)} - \beta_1^{(0)} + \gamma_2^{(1)} - \beta_2^{(1)} &\leq I(Y_1; X_0, X_1) = H(Y_1) + H(X_0, X_1) - H(X_0, X_1, Y_1) \\
&\leq H(Y_1) + H(X_0) + H(X_1) - H(X_0, X_1, Y_1) \\
&= H(X_0) + H(Y_1) - H(X_0, Y_1|X_1), && (64)\\
\gamma_1^{(1)} - \beta_1^{(1)} + \gamma_2^{(1)} - \beta_2^{(1)} &\leq I(X_1; X_0) + I(Y_1; X_0, X_1) = H(X_1) + H(Y_1) - H(X_1, Y_1|X_0). && (65)
\end{aligned}$$

Finally when $\mathcal{J}$ has all the three elements, we have

$$\begin{aligned}
&\gamma_1^{(0)} - \beta_1^{(0)} + \gamma_1^{(1)} - \beta_1^{(1)} + \gamma_2^{(1)} - \beta_2^{(1)} \\
&= I(X_0; X_1) + I(V_1, V_2; X_1) - \max(I(V_1, V_2; Y_1|X_0, X_1), I(V_1, V_2; Y_1|X_0, X_2)) && (66)\\
&\leq I(X_0; X_1) + I(V_1, V_2; X_1) - I(V_1, V_2; Y_1|X_0, X_1) && (67)\\
&= H(X_0) + H(X_1) + H(Y_1) - H(X_0, X_1, Y_1). && (68)
\end{aligned}$$

Thus, (24) is indeed true with the assignments (56)–(60). This, in fact, completes the proof. □

We can use any explicit distribution $(X_0, X_1, X_2, Y_1, Y_2)$ to obtain an explicit inner bound to $\tilde{\mathcal{R}}_{in}^{(t)}$, and the next corollary provides one such non-trivial bound. For convenience, we write the entropy function of a probability mass $(p_1, \ldots, p_t)$ as $H(p_1, \ldots, p_t)$.

Corollary 2. *The following* $(\bar{\alpha}, \bar{\beta}) \in \mathcal{R}$ *for any* $p \in [0,1]$*:*

$$\begin{aligned}\bar{\alpha} =& \frac{9}{4} - H(\frac{1}{4}, \frac{3}{4}) + \frac{1}{4}H(\frac{1-p}{2}, \frac{1-p}{2}, \frac{p}{2}, \frac{p}{2}) \\ &+ \frac{1}{2}H(\frac{2-p}{4}, \frac{2-p}{4}, \frac{p}{2}) - \frac{3}{4}H(\frac{3-2p}{6}, \frac{3-2p}{6}, \frac{p}{3}, \frac{p}{3}), \\ \bar{\beta} =& \frac{5}{8} + \frac{1}{4}H(\frac{2-p}{4}, \frac{2-p}{4}, \frac{p}{2}) - \frac{1}{8}H(\frac{1-p}{2}, \frac{1-p}{2}, p).\end{aligned}$$

Proof. These trade-off pairs are obtained by applying Corollary 1, and choosing $t = 1$ and setting (X_1, X_2, Y_1, Y_2) as given in (17), and letting X_0 be defined as in Table 1. Note that the joint distribution indeed satisfies the required Markov structure, and in this case $\alpha_2^{(1)} = \beta_2^{(1)}$ and $\alpha_2^{(2)} = \beta_2^{(2)}$. □

Table 1. Conditional distribution $P_{X_0|W_1,W_2}$ used in Corollary 2.

(w_1, w_2)	$x_0 = (00)$	$x_0 = (01)$	$x_0 = (10)$	$x_0 = (11)$
(00)	1/2			1/2
(10)	$(1-p)/2$	p		$(1-p)/2$
(01)	$(1-p)/2$		p	$(1-p)/2$
(11)	1/2			1/2

5. Conclusions

We consider the problem of private information retrieval using a Shannon-theoretic approach. A new coding scheme based on random coding and binning is proposed, which reveals a hidden connection to the multiple description problem. It is shown that for the $(2,2)$ PIR setting, this non-linear coding scheme is able to provide the best known tradeoff between retrieval rate and storage rate, which is strictly better than that achievable using linear codes. We further investigate the relation between zero-error PIR codes and ϵ-error PIR codes in this setting and show that they do not cause any essential difference in this problem setting. We hope that the hidden connection to multiple description coding can provide a new avenue to design more efficient PIR codes.

Author Contributions: Conceptualization, C.T., H.S. and J.C.; Methodology, C.T., H.S. and J.C.; Investigation, C.T., H.S. and J.C.; Writing—original draft, C.T.; Writing—review and editing, C.T., H.S. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science Foundation through grants CCF-18-16518, CCF-18-16546, CCF-20-07067, CCF-20-07108, and CCF-20-45656.

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

References

1. Chor, B.; Goldreich, O.; Kushilevitz, E.; Sudan, M. Private information retrieval. In Proceedings of the 36th Annual Symposium on Foundations of Computer Science, Milwaukee, WI, USA, 23–25 October 1995; pp. 41–50.
2. Shah, N.; Rashmi, K.; Ramchandran, K. One extra bit of download ensures perfectly private information retrieval. In Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 856–860.
3. Fazeli, A.; Vardy, A.; Yaakobi, E. Codes for distributed PIR with low storage overhead. In Proceedings of the 2015 Proceedings of IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 2852–2856.
4. Rao, S.; Vardy, A. Lower bound on the redundancy of PIR codes. *arXiv* **2016**, arXiv:1605.01869.
5. Blackburn, S.R.; Etzion, T. PIR array codes with optimal virtual server rate. *IEEE Trans. Inf. Theory* **2019**, *65*, 6136–6145. [CrossRef]

6. Blackburn, S.R.; Etzion, T.; Paterson, M.B. PIR schemes with small download complexity and low storage requirements. *IEEE Trans. Inf. Theory* **2019**, *66*, 557–571. [CrossRef]
7. Zhang, Y.; Wang, X.; Wei, H.; Ge, G. On private information retrieval array codes. *IEEE Trans. Inf. Theory* **2019**, *65*, 5565–5573. [CrossRef]
8. Vajha, M.; Ramkumar, V.; Kumar, P.V. Binary, shortened projective Reed Muller codes for coded private information retrieval. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 2648–2652.
9. Asi, H.; Yaakobi, E. Nearly optimal constructions of PIR and batch codes. *IEEE Trans. Inf. Theory* **2018**, *65*, 947–964. [CrossRef]
10. Chan, T.H.; Ho, S.W.; Yamamoto, H. Private information retrieval for coded storage. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 2842–2846.
11. Sun, H.; Jafar, S.A. The capacity of private information retrieval. *IEEE Trans. Inf. Theory* **2017**, *63*, 4075–4088. [CrossRef]
12. Tajeddine, R.; Gnilke, O.W.; El Rouayheb, S. Private information retrieval from MDS coded data in distributed storage systems. *IEEE Trans. Inf. Theory* **2018**, *64*, 7081–7093. [CrossRef]
13. Banawan, K.; Ulukus, S. The capacity of private information retrieval from coded databases. *IEEE Trans. Inf. Theory* **2018**, *64*, 1945–1956. [CrossRef]
14. Tian, C.; Sun, H.; Chen, J. Capacity-achieving private information retrieval codes with optimal message size and upload cost. *IEEE Trans. Inf. Theory* **2019**, *65*, 7613–7627. [CrossRef]
15. Zhou, R.; Tian, C.; Sun, H.; Liu, T. Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size. *IEEE Trans. Inf. Theory* **2020**, *66*, 4904–4916. [CrossRef]
16. Sun, H.; Jafar, S.A. The capacity of robust private information retrieval with colluding databases. *IEEE Trans. Inf. Theory* **2018**, *64*, 2361–2370. [CrossRef]
17. Ulukus, S.; Avestimehr, S.; Gastpar, M.; Jafar, S.; Tandon, R.; Tian, C. Private retrieval, computing and learning: Recent progress and future challenges. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 729–748. [CrossRef]
18. Attia, M.A.; Kumar, D.; Tandon, R. The capacity of private information retrieval from uncoded storage constrained databases. *IEEE Trans. Inf. Theory* **2020**, *66*, 6617–6634. [CrossRef]
19. Sun, H.; Jafar, S.A. Multiround private information retrieval: Capacity and storage overhead. *IEEE Trans. Inf. Theory* **2018**, *64*, 5743–5754. [CrossRef]
20. Sun, H.; Tian, C. Breaking the MDS-PIR capacity barrier via joint storage coding. *Information* **2019**, *10*, 265. [CrossRef]
21. Guo, T.; Zhou, R.; Tian, C. New results on the storage-retrieval tradeoff in private information retrieval systems. *IEEE J. Sel. Areas Inf. Theory* **2021**, *2*, 403–414. [CrossRef]
22. Tian, C.; Sun, H.; Chen, J. A Shannon-theoretic approach to the storage-retrieval tradeoff in PIR systems. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–20 June 2018; pp. 1904–1908.
23. Gamal, A.; Cover, T. Achievable rates for multiple descriptions. *IEEE Trans. Inf. Theory* **1982**, *28*, 851–857. [CrossRef]
24. Tian, C. On the storage cost of private information retrieval. *IEEE Trans. Inf. Theory* **2020**, *66*, 7539–7549. [CrossRef]
25. Venkataramani, R.; Kramer, G.; Goyal, V.K. Multiple description coding with many channels. *IEEE Trans. Inf. Theory* **2003**, *49*, 2106–2114. [CrossRef]
26. Wyner, A.; Ziv, J. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inf. Theory* **1976**, *22*, 1–10. [CrossRef]
27. Pradhan, S.S.; Puri, R.; Ramchandran, K. n-channel symmetric multiple descriptions-part I: (n, k) source-channel erasure codes. *IEEE Trans. Inf. Theory* **2004**, *50*, 47–61. [CrossRef]
28. Tian, C.; Chen, J. New coding schemes for the symmetric K-description problem. *IEEE Trans. Inf. Theory* **2010**, *56*, 5344–5365. [CrossRef]
29. Sgarro, A. Source coding with side information at several decoders. *IEEE Trans. Inf. Theory* **1977**, *23*, 179–182. [CrossRef]