On the Capacity of Latent Variable Private Information Retrieval

Islam Samy Mohamed A. Attia Ravi Tandon Loukas Lazos
Department of Electrical and Computer Engineering, University of Arizona
Email: {*islamsamy, madel, tandonr, llazos*}@email.arizona.edu

***Abstract*—In latent-variable private information retrieval (LV-PIR), a user wishes to retrieve one out of K messages (indexed by θ) without revealing any information about a sensitive *latent* attribute (modeled by a latent variable S correlated with θ). While conventional PIR protocols, which keep θ private, also suffice for hiding S, they can be too costly in terms of the download overhead. In this paper, we characterize the capacity (equivalently, the optimal download cost) of LV-PIR as a function of the distribution $P_{S|\theta}$. We present a converse proof that yields a lower bound on the optimal download cost, and a matching achievable scheme. The optimal scheme, however, involves an exhaustive search over subset queries and over all messages, which can be computationally prohibitive. We further present two low-complexity, albeit sub-optimal, schemes that also outperform the conventional PIR solution.**

I. Introduction

Private information retrieval (PIR) [1] allows a user to download one out of K messages from a curious database, without revealing the message index to the database. This is typically achieved at the expense of an increased communication cost, because more than one message has to be downloaded to preserve privacy. The PIR problem has been widely studied under information-theoretic privacy assumptions, [2]–[39], for different model setups. However, when a single database is considered or databases belong to a single operator as is the case for most online services, the majority of information-theoretic PIR models degenerate to impractical solutions requiring the download of the entire database.

We consider the latent-variable PIR (LV-PIR) problem which focuses on retrieving content while preserving the privacy of sensitive (latent) attributes. Clearly, a traditional PIR protocol that hides the identity of content (modeled by index θ) also preserves the privacy of the latent variable (modeled by S). However, this may not be necessary, as latent-variable privacy constraint is weaker than perfect message privacy, thereby providing an opportunity to reduce the download overhead. The LV-PIR problem was introduced in our prior work [40], where it was shown that it is possible to reduce the download cost beyond that of conventional PIR.

In this paper, we settle the LV-PIR problem by characterizing the optimal download cost as a function of the joint distribution $P_\theta \times P_{S|\theta}$. We focus on the case where the message index θ is uniformly distributed over the set $\{1, 2, \ldots, K\}$,

This work was supported by NSF grants CNS-1715947, CAREER-1651492, CCF-2100013, and CNS-1813401.

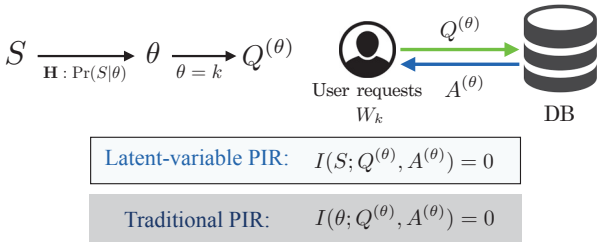

Fig. 1: Latent-variable PIR (LV-PIR).

whereas $P_{S|\theta}$ can be arbitrary, and described by a $|\mathcal{S}| \times |\theta|$ matrix $\mathbf{H}$. Our main contributions are summarized as follows:

- We derive a converse proof for the LV-PIR problem that gives a lower bound on the optimal download cost. We show that the cost reduction (compared to conventional PIR) depends on the correlation between the message index θ and S, as captured by matrix $\mathbf{H}$. We present an LV-PIR scheme that achieves the lower bound based on the idea of probabilistic subset queries over the messages. The minimum download cost is then obtained by searching over all possible probability assignments and over all subset queries that contain the desired message.
- The optimal probabilistic subset query scheme can be computationally prohibitive in practice, as it involves a search over a probability simplex of size 2^{K-1} (all subsets containing the desired message). Alternatively, we present two low-complexity schemes that, while not necessarily optimal, significantly outperform conventional PIR. Our first low-complexity scheme reduces the search space to a probability simplex of size $\binom{K}{\text{rank}(\mathbf{H})} = O(K^{\text{rank}(\mathbf{H})})$, and achieves a download cost which never exceeds rank($\mathbf{H}$). This scheme can be particularly useful when the number of latent attributes is much smaller compared to the number of messages in the database ($|S| << |\theta|$). Our second scheme further reduces the complexity to $O(K)$ and achieves a download cost not exceeding the number of unique columns in $\mathbf{H}$. This scheme can be beneficial in scenarios where content(s) (e.g., movies from a genre) reveal the same information about a sensitive attribute, which translates to a large number of repeated columns in $\mathbf{H}$.

Due to space limitations, we present the proof of our main result (Theorem 1) in the paper and describe the main ideas behind the two low-complexity schemes via an example.

II. Problem Formulation

We consider the PIR setting in Fig. 1. A set $\mathcal{W} = \{W_1, W_2, \ldots, W_K\}$ of K independent messages, each of size L bits, are stored on a single database (DB). The user draws a message index θ uniformly from the set $\{1, 2, \ldots, K\} \triangleq [1:K]$. The user is interested in retrieving message W_θ while hiding a latent variable S, which can take one of T values from alphabet $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$. The latent variable S is correlated with the message index θ, as described by the conditional probability $\Pr(S = s_t|\theta = k)$, where $t \in [1:T]$. We capture this conditional probability via a $T \times K$ matrix $\mathbf{H}$ with $[\mathbf{H}]_{t,k} \triangleq h_{tk} = \Pr(S = s_t|\theta = k)$. We assume that the matrix $\mathbf{H}$ is fixed and publicly-known to the DB and the user. The pair (S, θ) is assumed to be independent of the messages.

To retrieve message W_θ, the user submits a query $Q^{(\theta)}$ to the DB. The DB determines the corresponding answer $A^{(\theta)}$ as a function of the query $Q^{(\theta)}$ and the K messages. We assume that the user does not exploit the knowledge of S in the query construction. Then, the variables $S \to \theta \to Q^{(\theta)}$ form a Markov chain as S is conditionally independent of $Q^{(\theta)}$ given the index θ, i.e.,

$$\Pr(S = s_t|\theta = k, Q^{(\theta)} = q) = \Pr(S = s_t|\theta = k). \quad (1)$$

An LV-PIR scheme must satisfy the following correctness and privacy constraints.

Correctness Constraint: The user must be able to recover the requested message W_θ from $Q^{(\theta)}$ and $A^{(\theta)}$, i.e.,

$$H(W_\theta|Q^{(\theta)}, A^{(\theta)}) = 0. \quad (2)$$

Latent-variable Privacy Constraint: The latent variable S should be independent of the query $Q^{(\theta)}$ and its corresponding answer $A^{(\theta)}$, i.e, $Q^{(\theta)}$ and $A^{(\theta)}$ should not leak any *additional* information about S than what is known by the prior distribution of S. This implies that $I(S; Q^{(\theta)}, A^{(\theta)}) = 0$. Equivalently, for any realization of $(Q^{(\theta)} = q, A^{(\theta)} = a)$, we must have

$$\Pr(S = s_t|Q^{(\theta)} = q, A^{(\theta)} = a) = \Pr(S = s_t),\ \forall s_t, q, a. \quad (3)$$

Let $\vec{h}_k$ be the k^{th} column of $\mathbf{H}$. Then, we can express the privacy constraint (3) in terms of the elements in matrix $\mathbf{H}$, as shown in the following Lemma:

Lemma 1. *For an LV-PIR problem characterized by matrix* $\mathbf{H} = [\vec{h}_1 \ \cdots \ \vec{h}_K]$ *and* K *equiprobable messages at the DB, the LV-PIR privacy constraint* (3) *is equivalent to*

$$\sum_{k=1}^{K} \big(\frac{1}{K} - p_{k|q}\big) \cdot \vec{h}_k = 0, \quad (4)$$

where $p_{k|q} \triangleq \Pr(\theta = k|Q^{(\theta)} = q)$.

The proof of Lemma 1 follows by substituting $h_{tk} = \Pr(S = s_t|\theta = k)$ in (3) followed by algebraic manipulations. The full proof is shown in the detailed version [41].

Download Cost: The average download cost $D_{\mathbf{H}}$ of a LV-PIR scheme is defined as follows:

$$D_{\mathbf{H}} = H(A^{(\theta)}|Q^{(\theta)}) = \sum_q \Pr(Q^{(\theta)} = q) D_{\mathbf{H}}(q), \quad (5)$$

where $D_{\mathbf{H}}(q) = H(A^{(\theta)}|Q^{(\theta)} = q)$ is the expected number of downloaded bits when query $Q^{(\theta)} = q$ is submitted. The pair $(L, D_{\mathbf{H}})$ is said to be achievable if there exists an LV-PIR scheme that satisfies the correctness and LV privacy constraint, and can retrieve a message of size L bits by downloading an average of $D_{\mathbf{H}}$ bits. Our goal is to characterize the minimum download cost $D^*(\mathbf{H})$,

$$D^*(\mathbf{H}) = \min\{D_{\mathbf{H}}/L : (L, D_{\mathbf{H}}) \text{ is achievable}\}. \quad (6)$$

III. Main Results and Discussion

The following theorem states the minimum download cost for the LV-PIR problem defined in (6).

Theorem 1. *Define* $\mathcal{K}$ *as the powerset of* $[1:K]$ *excluding the empty set* $\emptyset$*, i.e.,* $\mathcal{K}$ *includes all non-empty subsets of indices* $[1:K]$*. Define* $\mathcal{K}_k$ *as the set of all subsets within* $\mathcal{K}$ *that include index* k*. Also, define* $\vec{\pi}_k = \{\pi_{q|k} \triangleq \Pr(Q^{(\theta)} = q|\theta = k)|\ \forall q \in \mathcal{K}\}$ *to be a valid PMF over the support* $\mathcal{K}$*. The optimal download cost of LV-PIR is given as follows*

$$D^*(\mathbf{H}) = \min_{\vec{\pi}_1, \vec{\pi}_2, \ldots, \vec{\pi}_K} \frac{1}{K} \sum_{k=1}^{K} \sum_{q \in \mathcal{K}} \pi_{q|k}\ |q| \quad (7)$$

$$\text{s.t.} \quad \pi_{q|k} = 0, \quad \forall q \notin \mathcal{K}_k, \forall k \in [1:K], \quad (8)$$

$$\sum_{k=1}^{K} \big(\pi_{q|k} - \frac{1}{K} \sum_{k'=1}^{K} \pi_{q|k'}\big) \cdot \vec{h}_k = 0, \quad \forall q \in \mathcal{K}, \quad (9)$$

$$0 \le \pi_{q|k} \le 1, \quad \sum_{q \in \mathcal{K}} \pi_{q|k} = 1, \quad \forall k \in [1:K]. \quad (10)$$

Proof of Theorem 1 is presented in Section IV. The query cardinality $|q|$ is the number of downloaded messages when the user submits query q. The objective function in (7) minimizes the average download cost over all probabilistic sub-set queries. The condition in (8) is used to satisfy correctness. The condition in (9) is used to ensure LV-PIR privacy, while (10) is to ensure valid distributions $\vec{\pi}_1, \vec{\pi}_2, \ldots, \vec{\pi}_K$. The achievability scheme for the optimal download cost in Theorem 1 requires searching over all subset queries that include the desired message index k. The number of these queries is 2^{K-1}, leading to an exponentially-increasing complexity with K.

In the following theorems, we propose two low-complexity achievability schemes. Due to space limitations, the two schemes are fully described in [41]. Here, we sketch the main ideas behind the schemes.

Probabilistic LV-PIR scheme: The probabilistic scheme utilizes a connection between valid subset queries and vectors $\vec{c} = \{c_1, \ldots, c_K\}$ inside the null-space of $\mathbf{H}$. For any query q that belongs to the powerset of $[1:K]$ and satisfies LV-privacy, there is a vector $\vec{c}$ where $c_k = {}^1\!/_K - p_{k|q}, \forall k$. This connection follows from the privacy condition in Lemma 1 where the equality only holds when there is some vector $\vec{c}$, inside the null-space of $\mathbf{H}$, with elements equal $c_k = {}^1\!/_K - p_{k|q}, \forall k$.

The scheme reduces the download cost by minimizing the size $|q|$ of the submitted query, hence minimizing the number of downloaded messages. This is performed by setting $\pi_{q|k} = 0$ for as many messages as possible. Having $c_k = 1/K$ (equivalently $p_{k|q} = 0$) ensures that $\pi_{q|k} = 0$ which means q will not be submitted to download W_k, then it is not required to include k. Using the above connection, the main idea of the scheme is to find vectors $\vec{c}$ that have the maximum number of elements set to $1/K$, then create a related subset query from each of them. The related query is formed such that it includes all indices k, where $c_k < \frac{1}{K}$ ($\pi_{q|k} > 0$). Let $R(\mathbf{H})$ be the rank of $\mathbf{H}$, then we have $K - R(\mathbf{H})$ free variables inside each vector $\vec{c}$. These variables can be controlled and set to $1/K$. Thus, each query can avoid downloading at least $K - R(\mathbf{H})$ messages. This explains the intuition behind the statement of Theorem 2 which states that the download cost of this scheme is upper-bounded by $R(\mathbf{H})$.

Theorem 2. *The Probabilistic LV-PIR scheme achieves a download cost not exceeding $R(\mathbf{H})$, and has a query search complexity of $\binom{K}{R(\mathbf{H})} = O(K^{R(\mathbf{H})})$, where $R(\mathbf{H})$ denotes the rank of $\mathbf{H}$.*

The probabilistic scheme that obtains the bound in Theorem 2 requires searching over $\binom{K}{R(\mathbf{H})} = O(K^{R(\mathbf{H})})$ subset queries. Although the complexity of the Probabilistic LV-PIR scheme grows polynomially with K, it can still be substantial if $\mathbf{H}$ has a high rank. We propose a grouping LV-PIR scheme which further reduces the query complexity as follows.

Grouping LV-PIR scheme: The grouping scheme deals with the case when groups of messages are equivalently correlated to the latent variable S. In this scheme, all messages with identical columns in $\mathbf{H}$ are grouped together. The subset queries are then designed such that one message is downloaded from each group. The number of downloaded messages for any query is $U(\mathbf{H})$, the number of unique columns in $\mathbf{H}$. Then, this always leads to a download cost of $U(\mathbf{H})$.

Theorem 3. *The Grouping LV-PIR scheme achieves a download cost not exceeding $U(\mathbf{H})$, where $U(\mathbf{H})$ denotes the number of unique columns in $\mathbf{H}$, and has a query search complexity of $O(K)$.*

We emphasize that this grouping scheme always yields a lower download cost compared to the grouping scheme proposed in our previous work [40]. In the following example, we compare conventional PIR with the proposed LV-PIR schemes and show how the download cost can be reduced.

Example 1. Consider the LV-PIR problem with $K = 5$ equiprobable messages and a latent variable S taking $T = 2$ values. The conditional distribution $P_{S|\theta}$ described by the matrix $\mathbf{H}$ is given as follows:

$$\mathbf{H} = \begin{bmatrix} 1/4 & 5/8 & 5/8 & 5/8 & 3/8 \\ 3/4 & 3/8 & 3/8 & 3/8 & 5/8 \end{bmatrix}. \tag{11}$$

Conventional PIR scheme [3]: For the PIR setting with one DB, the only solution that satisfies perfect privacy for θ (and subsequently for S) is the following: download all five messages, i.e., the download cost is:

$$D_{\text{PIR}} = K = 5. \tag{12}$$

TABLE I: Deterministic LV-PIR scheme [40]

Subset query	$\pi_{q\|1}$	$\pi_{q\|2}$	$\pi_{q\|3}$	$\pi_{q\|4}$	$\pi_{q\|5}$
$q_1 = \{1,2,3\}$	1	1	1	0	0
$q_2 = \{4,5\}$	0	0	0	1	1

TABLE II: Probabilistic LV-PIR scheme (Theorem 2)

Subset query	$\pi_{q\|1}$	$\pi_{q\|2}$	$\pi_{q\|3}$	$\pi_{q\|4}$	$\pi_{q\|5}$
$q_1 = \{1,2\}$	1/3	2/3	0	0	0
$q_2 = \{1,3\}$	1/3	0	2/3	0	0
$q_3 = \{1,4\}$	1/3	0	0	2/3	0
$q_4 = \{2,5\}$	0	1/3	0	0	1/3
$q_5 = \{3,5\}$	0	0	1/3	0	1/3
$q_6 = \{4,5\}$	0	0	0	1/3	1/3

Deterministic LV-PIR scheme [40]: In this scheme, messages are divided into disjoint subsets such that downloading any individual subset does not violate LV privacy (meets the prior distribution of S). In this example, we divide the five messages into subsets $q_1 = \{1, 2, 3\}$ and $q_2 = \{4, 5\}$. To download any desired message, the user submits the query that includes its index. Table I shows the query assignment for every message. The download cost for this scheme is

$$\begin{aligned} D^{(\text{LV})}_{\text{deterministic}} &= \sum_{k=1}^{3} \Pr(\theta = k) \cdot |q_1| + \sum_{k=4}^{5} \Pr(\theta = k) \cdot |q_2| \\ &= \frac{3}{5} \times 3 + \frac{2}{5} \times 2 = \frac{13}{5} = 2.6. \end{aligned} \tag{13}$$

This is clearly less than D_{PIR}. However, we can further reduce the download cost using a probabilistic scheme.

Probabilistic LV-PIR scheme (Theorem 2): The first step in this scheme is to find vectors $\vec{c}$, inside the null-space of $\mathbf{H}$, that include the maximum number of elements set to $1/5 = 0.2$. As the matrix $\mathbf{H}$ in (11) is of rank $R(\mathbf{H}) = 2$, there can be three free variables inside $\vec{c}$ that can be set to 0.2. The free variables can be freely chosen, then we can obtain $\binom{5}{3} = 10$ different vectors. The next step is to exclude any vector $\vec{c}$ that includes elements exceeding 0.2 as this later can lead to negative probability assignment. Following that, we get only six remaining vectors. From each vector $\vec{c}$, we obtain a related subset query, $q = \{k : c_k < 0.2\}$. For instance, setting $c_3 = c_4 = c_5 = 0.2$ as free variables, we get the vector $\vec{c} = \{-2/15, -7/15, 0.2, 0.2, 0.2\}$. The related query for this vector is $q = \{k : c_k < 0.2\} = \{1, 2\}$. The remaining five vectors $\vec{c}$, and their related queries, can be obtained similarly. Then, we have a total of six queries: $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 5\}, \{3, 5\}$, and $\{4, 5\}$. For each desired message W_k, the user assigns a probability $\pi_{q|k} = \alpha_q(0.2 - c_k)$ to each of these queries, where α_q is a weighting factor used to ensure that the values for $\pi_{q|k}, \forall k, q$ create a valid PMFs for the distributions $\vec{\pi}_1, \vec{\pi}_2, \ldots, \vec{\pi}_K$. For instance, the factor α_q for the query $\{1, 2\}$ is calculated as $\alpha_q = 1$.

TABLE III: Grouping LV-PIR scheme (Theorem 3)

Subset query	$\pi_{q\|1}$	$\pi_{q\|2}$	$\pi_{q\|3}$	$\pi_{q\|4}$	$\pi_{q\|5}$
$q_1 = \{1, 2, 5\}$	1/3	1	0	0	1/3
$q_2 = \{1, 3, 5\}$	1/3	0	1	0	1/3
$q_3 = \{1, 4, 5\}$	1/3	0	0	1	1/3

Then, the query $\{1, 2\}$ is submitted to request messages W_1 and W_2 with probabilities $\pi_{q|1} = 1 \times (0.2 + 2/15) = 1/3$ and $\pi_{q|2} = 1 \times (0.2 + 7/15) = 2/3$, respectively. The query assignment for each message is shown in Table II.

As all subset queries are of size two, the download cost for this scheme can be written as

$$D^{(\text{LV})}_{\text{probabilistic}} = D(\mathbf{H}) = \sum_{k=1}^{5} \Pr(\theta = k) \cdot 2 = 2 = R(\mathbf{H}), \quad (14)$$

where $R(\mathbf{H})$ is the rank of matrix $\mathbf{H}$. We highlight that for this example, the scheme shown in Table II is optimal.

Grouping LV-PIR scheme (Theorem 3): For the matrix $\mathbf{H}$ in (11), since there are $U(\mathbf{H}) = 3$ unique columns, the following three groups are formed: $\{1\}, \{2, 3, 4\}$, and $\{5\}$. By picking one message from each group, we obtain the three queries shown in Table III. For any desired message, the user uniformly selects between queries that include the message index. For instance, the user selects each of the three queries with probability $1/3$ when message W_1 or W_5 is desired. All subset queries are of size three. The average download cost is

$$D^{(\text{LV})}_{\text{grouping}} = D(\mathbf{H}) = \sum_{k=1}^{5} \Pr(\theta = k) \cdot 3 = 3 = U(\mathbf{H}). \quad (15)$$

Although this scheme yields a higher download cost than the previous two LV-PIR schemes, it is still lower than the conventional PIR solution of downloading all messages. The advantage of this scheme is its lower complexity.

It is straightforward to verify the correctness of the proposed solutions for the three LV-PIR schemes where each message index k is included in any subset query q that can be requested to download W_k with a non-zero probability ($\pi_{q|k} > 0$). Furthermore, it can be shown that the three solutions satisfy the LV privacy condition in (9) by substituting the values of $\pi_{q|k}$ for different queries, inside Tables I, II, and III.

IV. Proof of Theorem 1

To prove Theorem 1, we present an LV-PIR scheme followed by a converse proof with a matching lower bound.

A. Achievability: Exhaustive Search LV-PIR Scheme

We propose an exhaustive search scheme that minimizes the LV-PIR download cost by searching over all valid subset queries that satisfy both the correctness and LV privacy requirements. Let $\mathcal{K}$ be the powerset of the set $[1 : K]$, excluding the empty set $\emptyset$, i.e., $\mathcal{K}$ includes all non-empty subsets of the indices $[1 : K]$. Consider any subset $q \subseteq \mathcal{K}$, for which we define $\pi_{q|k}$ as follows:

$$\pi_{q|k} \triangleq \Pr(Q^{(\theta)} = q | \theta = k), \quad (16)$$

i.e., $\pi_{q|k}$ is probability of sending the subset query q if message k is desired by the user. For any query q, the LV privacy constraint in (4) can be written in terms of $\{\pi_{q|k}\}_{k=1}^{K}$ as

$$\sum_{k=1}^{K} \left(\frac{1}{K} - p_{k|q} \right) \cdot \vec{h}_k = \sum_{k=1}^{K} \left(\frac{1}{K} - \frac{\pi_{q|k} \cdot \frac{1}{K}}{\Pr(Q = q)} \right) \cdot \vec{h}_k = 0.$$

Rearranging the above equation and substituting $\Pr(Q = q) = \frac{1}{K} \sum_{k'=1}^{K} \pi_{q|k'}$ yields

$$\sum_{k=1}^{K} \left(\pi_{q|k} - \frac{1}{K} \sum_{k'=1}^{K} \pi_{q|k'} \right) \cdot \vec{h}_k = 0, \quad \forall q \in \mathcal{K}. \quad (17)$$

Any valid choice for the distribution $\vec{\pi}_k = \{\pi_{q|k} \mid \forall q \in \mathcal{K}\}$, over all possible queries has to satisfy two properties. First, each distribution $\vec{\pi}_k$ must be a valid PMF.

$$0 \leq \pi_{q|k} \leq 1, \quad \sum_{q \in \mathcal{K}} \pi_{q|k} = 1, \quad \forall k \in [1 : K]. \quad (18)$$

Second, let $\mathcal{K}_k$ be the set of all subsets within $\mathcal{K}$ that include index k. To satisfy decodability of message W_k, the subset query must be chosen such that $q \in \mathcal{K}_k$. That is,

$$\pi_{q|k} = 0, \quad \forall q \notin \mathcal{K}_k. \quad (19)$$

This is equivalent to $q = \{k \mid \pi_{q|k} > 0\}$.

Download cost: Let $|q|$ be the cardinality of the subset query q. We can express the number of downloaded bits when query q is submitted as $D_{\mathbf{H}}(q) = |q| \cdot L = L \sum_{k=1}^{K} I(\pi_{q|k} > 0)$, where $I(.)$ is an indicator function. The download cost $D_{\mathbf{H}}$, given specific distributions $\vec{\pi}_1$, $\vec{\pi}_2$, …, $\vec{\pi}_K$, can be written as

$$D_{\mathbf{H}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{q \in \mathcal{K}} \pi_{q|k} \cdot D_{\mathbf{H}}(q). \quad (20)$$

An exhaustive search LV-PIR scheme minimizes the download cost by searching over all possible distributions $\vec{\pi}_1$, $\vec{\pi}_2$, …, $\vec{\pi}_K$ that satisfy the requirements in (18), (19), and LV privacy constraint in (17). Thus, the exhaustive search scheme leads to the following upper bound on the optimal download cost,

$$\begin{aligned} D^*(\mathbf{H}) \leq \min_{\vec{\pi}_1, \vec{\pi}_2, \ldots, \vec{\pi}_K} & \frac{1}{K} \sum_{k=1}^{K} \sum_{q \in \mathcal{K}} \pi_{q|k} \ |q| \\ \text{s.t.} \quad & \pi_{q|k} = 0, \quad \forall q \notin \mathcal{K}_k, \forall k \in [1 : K], \\ & \sum_{k=1}^{K} \left(\pi_{q|k} - \frac{1}{K} \sum_{k'=1}^{K} \pi_{q|k'} \right) \cdot \vec{h}_k = 0, \quad \forall q \in \mathcal{K}, \\ & 0 \leq \pi_{q|k} \leq 1, \ \sum_{q \in \mathcal{K}} \pi_{q|k} = 1, \ \forall k \in [1 : K]. \quad (21) \end{aligned}$$

B. Converse Proof

We now derive a lower bound on the minimum download cost $D^*(\mathbf{H})$ that matches the upper bound in (21). Consider any arbitrary scheme $\mathcal{B}$ that satisfies both the correctness and privacy conditions in (2) and (3), respectively. The minimum download cost $D^*(\mathbf{H})$ for LV-PIR can be represented as

$$D^*(\mathbf{H}) = \min_{\mathcal{B}} D_{\mathcal{B}}(\mathbf{H}), \quad (22)$$

where $D_{\mathcal{B}}(\mathbf{H})$ is the download cost when scheme $\mathcal{B}$ is used. Since $\theta \to Q^{(\theta)} \to A^{(\theta)}$, we define $A(q)$ to be the answer when a query q is submitted, regardless of the desired message. $D_{\mathcal{B}}(\mathbf{H})$ can be expressed as follows:

$$\begin{aligned}
D_{\mathcal{B}}(\mathbf{H}) &= \frac{1}{L} H(A^{(\theta)}|Q^{(\theta)}, \theta) \\
&= \frac{1}{LK} \sum_{k=1}^{K} H(A^{(k)}|Q^{(k)}, \theta = k) \\
&= \frac{1}{LK} \sum_{k=1}^{K} \sum_{q} \Pr(Q^{(k)} = q|\theta = k) H(A(q)|Q^{(k)} = q, \theta = k) \\
&\overset{(a)}{=} \frac{1}{LK} \sum_{k=1}^{K} \sum_{q} \pi_{q|k}\ H(A(q)|Q^{(k)} = q), \qquad (23)
\end{aligned}$$

where (a) is due to the Markov chain in (1). We next state a lemma that gives a necessary condition for decodability.

Lemma 2. *For any query q that can be submitted to download message W_k with non-zero probability, i.e., $\pi_{q|k} = \Pr(Q^{(k)} = q|\theta = k) > 0$, the message W_k must be decodable from the corresponding answer $A(q)$, i.e.,*

$$H(W_k|A(q), Q^{(k)} = q) = 0. \qquad (24)$$

The proof of this Lemma follows readily by expanding the correctness constraint (2) and is omitted. We next define the *image set* of query q, as $\mu^{(q)} \triangleq \{k|\ \pi_{q|k} > 0\}$, as the set of message indices for which q can be submitted. Let $\mathcal{W}_{\mu^{(q)}}$ be the set of messages whose indices are included in $\mu^{(q)}$. Using Lemma 2, we can lower bound the term in (23) as follows:

$$\begin{aligned}
&H(A(q)|Q^{(k)} = q) \\
&= H(\mathcal{W}_{\mu^{(q)}}, A(q)|Q^{(k)} = q) - H(\mathcal{W}_{\mu^{(q)}}|A(q), Q^{(k)} = q) \\
&\overset{(a)}{=} H(\mathcal{W}_{\mu^{(q)}}, A(q)|Q^{(k)} = q) \\
&\geq H(\mathcal{W}_{\mu^{(q)}}|Q^{(k)} = q) = |\mu^{(q)}| \cdot L, \qquad (25)
\end{aligned}$$

where (a) is due to Lemma 2, and $|\mu^{(q)}|$ is the cardinality of the image set $\mu^{(q)}$ for query q. Substituting (25) in (23) yields

$$D_{\mathcal{B}}(\mathbf{H}) \geq \frac{1}{K} \sum_{k=1}^{K} \sum_{q} \pi_{q|k}\ |\mu^{(q)}|. \qquad (26)$$

Furthermore, any query q using scheme $\mathcal{B}$, must also satisfy the LV privacy constraint, which can be written as follows:

$$\begin{aligned}
&\Pr(S = s_t) = \Pr(S = s_t|Q^{(k)} = q) \\
&= \sum_{k=1}^{K} \Pr(\theta = k|Q^{(k)} = q_j) \Pr(S = s_t|\theta = k, Q^{(k)} = q) \\
&= \sum_{k=1}^{K} \frac{\Pr(\theta = k)\pi_{q|k}}{\sum_{k'=1}^{K} \Pr(\theta = k)\pi_{q|k'}} \Pr(S = s_t|\theta = k), \qquad (27)
\end{aligned}$$

which can be rearranged in the following relation

$$\sum_{k=1}^{K} \pi_{q|k} (\Pr(S = s_t) - \Pr(S = s_t|\theta = k)) = 0. \qquad (28)$$

Our main idea for proving the optimality of subset queries is as follows: given an arbitrary query q for a scheme $\mathcal{B}$ (which may or may not be a subset query), we can construct a sub-set query $\tilde{q}$ which also satisfies decodability and LV-privacy, and does not exceed the download cost of q. Specifically, we let $\tilde{q} = \mu^{(q)}$, i.e., the derived sub-set query only requests to download the messages in the image set $\mu^{(q)}$ of query q, and define the query probabilities as follows:

$$\pi_{\tilde{q}|k} = \pi_{\mu^{(q)}|k} = \begin{cases} 0, & k \notin \mu^{(q)}, \\ \pi_{q|k} & k \in \mu^{(q)}. \end{cases} \qquad (29)$$

We denote the resulting subset query scheme as $\tilde{B}$. It is clear that $\tilde{q}$ satisfies decodability, and it follows from (28), (29) that every query $\tilde{q}$ in scheme $\tilde{B}$ also satisfies LV-privacy. The download cost of $\tilde{B}$ can be exactly written as follows:

$$\begin{aligned}
D_{\tilde{\mathcal{B}}}(\mathbf{H}) &= \frac{1}{K} \sum_{k=1}^{K} \sum_{\tilde{q}} \pi_{\tilde{q}|k}\ |\tilde{q}| = \frac{1}{K} \sum_{k=1}^{K} \sum_{\mu^{(q)}} \pi_{\mu^{(q)}|k}\ |\mu^{(q)}| \\
&= \frac{1}{K} \sum_{k=1}^{K} \sum_{q} \pi_{q|k}\ |\mu^{(q)}|. \qquad (30)
\end{aligned}$$

Hence, from (26) and (30), it follows that $D_{\mathcal{B}}(\mathbf{H}) \geq D_{\tilde{\mathcal{B}}}(\mathbf{H})$. Thus, we can now obtain a lower bound for the optimal download cost by minimizing over all valid subset queries, and arrive at the following lower bound on $D^*(\mathbf{H})$:

$$\begin{aligned}
D^*(\mathbf{H}) \geq \min_{\vec{\pi}_1, \vec{\pi}_2, \ldots, \vec{\pi}_K} & \frac{1}{K} \sum_{k=1}^{K} \sum_{q \in \mathcal{K}} \pi_{q|k}\ |q| \\
\text{s.t.} \quad & \pi_{q|k} = 0, \quad \forall q \notin \mathcal{K}_k, \forall k \in [1:K], \\
& \sum_{k=1}^{K} \Big(\pi_{q|k} - \frac{1}{K} \sum_{k'=1}^{K} \pi_{q|k'}\Big) \cdot \vec{h}_k = 0, \quad \forall q \in \mathcal{K}, \\
& 0 \leq \pi_{q|k} \leq 1,\ \sum_{q \in \mathcal{K}} \pi_{q|k} = 1,\ \forall k \in [1:K]. \qquad (31)
\end{aligned}$$

This completes the proof of Theorem 1.

V. Conclusion

In this paper, we studied the problem of latent-variable PIR, where information-theoretic privacy is preserved with respect to a latent variable. We characterized the capacity for the LV-PIR under the practical assumption of a single database. We derived a converse proof that gives a lower bound on the optimal download cost, and proposed an achievability scheme that matches the derived bound. Furthermore, we introduced a simplified general scheme that decreases the complexity of obtaining efficient retrieval queries. We utilized the structure of $\mathbf{H}$ to achieve a low-complexity LV-PIR grouping construction when messages are partitioned in groups with the same statistical properties with respect to the latent variable.

REFERENCES

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

[2] 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*. IEEE, 2014, pp. 856–860.

[3] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4075–4088, 2017.

[4] ——, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2920–2932, 2017.

[5] ——, "The capacity of robust private information retrieval with colluding databases," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2361–2370, 2018.

[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] 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.

[8] Q. Wang and M. Skoglund, "Symmetric private information retrieval for MDS coded distributed storage," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[9] 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, 2017.

[10] 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 Transactions on Information Theory*, vol. 64, no. 2, pp. 1000–1022, 2018.

[11] Z. Jia and S. A. Jafar, "X-secure t-private information retrieval from mds coded storage with byzantine and unresponsive servers," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7427–7438, 2020.

[12] 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)*. IEEE, 2018, pp. 966–970.

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

[14] R. Zhou, C. Tian, T. Liu, and H. Sun, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 370–374.

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

[16] Y. Zhang and G. Ge, "Private information retrieval from MDS coded databases with colluding servers under several variant models," *arXiv preprint arXiv:1705.03186*, 2017.

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

[18] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti, "Robust private information retrieval from coded systems with Byzantine and colluding servers," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2451–2455.

[19] K. Banawan and S. Ulukus, "Private information retrieval through wiretap channel ii: Privacy meets security," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4129–4149, 2020.

[20] Q. Wang and M. Skoglund, "Secure private information retrieval from colluding databases with eavesdroppers," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2456–2460.

[21] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3198–3214, 2018.

[22] R. Tandon, "The capacity of cache aided private information retrieval," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 1078–1082.

[23] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3215–3232, 2018.

[24] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, 2018, pp. 1–5.

[25] A. Heidarzadeh, S. Kadhe, S. El Rouayheb, and A. Sprintson, "Single-server multi-message individually-private information retrieval with side information," *2019 IEEE International Symposium on Information Theory (ISIT)*, Jul 2019. [Online]. Available: http://dx.doi.org/10.1109/ISIT.2019.8849283

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

[27] 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.

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

[29] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti, "Private information retrieval from coded storage systems with colluding, Byzantine, and unresponsive servers," *IEEE Transactions on information theory*, vol. 65, no. 6, pp. 3898–3906, 2019.

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

[31] Z. Jia, H. Sun, and S. A. Jafar, "Cross subspace alignment and the asymptotic capacity of x-secure t-private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5783–5798, 2019.

[32] 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, 2019.

[33] N. Raviv and I. Tamot, "Private information retrieval is graph based replication systems," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1739–1743.

[34] K. Banawan and S. Ulukus, "Private information retrieval from non-replicated databases," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1272–1276.

[35] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The capacity of private information retrieval from heterogeneous uncoded caching databases," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3407–3416, 2020.

[36] T. Guo, R. Zhou, and C. Tian, "New results on the storage-retrieval tradeoff in private information retrieval systems," *arXiv preprint arXiv:2008.00960*, 2020.

[37] J. Cheng, N. Liu, and W. Kang, "The capacity of symmetric private information retrieval under arbitrary collusion and eavesdropping patterns," *arXiv preprint arXiv:2010.08249*, 2020.

[38] S. Vithana, K. Banawan, and S. Ulukus, "Semantic private information retrieval," *arXiv preprint arXiv:2003.13667*, 2020.

[39] X. Yao, N. Liu, and W. Kang, "The capacity of private information retrieval under arbitrary collusion patterns," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1041–1046.

[40] I. Samy, M. A. Attia, R. Tandon, and L. Lazos, "Latent-variable private information retrieval," *the IEEE International Symposium on Information Theory (ISIT)*, 2020.

[41] ——, "On the capacity of latent-variable private information retrieval," Link:https://www.dropbox.com/s/3mkiptszeuav6am/TheCapacityofLatentVariablePIR.pdf?dl=0.