Optimal Rate-Distortion-Leakage Tradeoff for Single-Server Information Retrieval

Yauhen Yakimenka, Hsuan-Yin Lin, and Eirik Rosnes
Simula UiB
N-5006 Bergen, Norway
Email: {yauhen, lin, eirikrosnes}@simula.no

Jörg Kliewer
New Jersey Institute of Technology
Newark, New Jersey 07102, USA
Email: jkliewer@njit.edu

***Abstract*—Private information retrieval protocols guarantee that a user can *privately* and *losslessly* retrieve a single file from a database stored across multiple servers. In this work, we propose to simultaneously relax the conditions of perfect retrievability and privacy in order to obtain improved download rates in the single server scenario, i.e., all files are stored uncoded on a single server. In particular, we derive the optimal tradeoff between download rate, distortion, and information leakage when the file size is *infinite* and the information leakage is measured in terms of the average success probability for the server of correctly guessing the identity of the requested file. Moreover, we present a novel approach based on linear programming to construct schemes for a finite file size and an arbitrary number of files. When the database contains at most four bits, this approach can be leveraged to find provably optimal schemes.**

I. Introduction

Over the last decade, private information retrieval (PIR) [1] has received significant attention in the information theory community. See, for instance, [2]–[7] and references therein. In PIR, a user can retrieve an arbitrary file stored in a set of databases without disclosing any information (in an information-theoretic sense) about which file she is interested in to the servers storing the databases. For the single server scenario, i.e., all files are stored uncoded on a single server, it is well-known that downloading the entire database is optimal in terms of upload and download cost. Hence, several different approaches have been proposed in order to improve the communication cost. For instance, relaxing the perfect information-theoretical privacy condition by considering computationally-private information retrieval (CPIR), where the privacy requirement relies on an intractability assumption (e.g., the hardness of deciding quadratic residuosity), was proposed already in 1997 [8]. In CPIR, given infinite computational power, the identity of the requested file can be determined precisely. By leveraging the assumption that the user has some prior side information on the content of the database, the download rate can indeed be improved while preserving information-theoretic privacy [9]. In [9], two cases are considered, namely whether or not the privacy of the side information needs to be preserved. Several parallel and follow-up works have appeared recently, see, e.g., [10], [11], and references therein. Alternatively, the download rate can be improved by relaxing the perfect privacy condition, referred to as weakly-private information retrieval (WPIR), as shown in [12]–[15]. In [13], an exact expression for the WPIR capacity in the single server scenario was derived using both mutual information and maximal leakage (MaxL) [16], [17] as privacy metric. Recently, the multi-server WPIR problem under the MaxL metric has also been studied in [18], [19].

In this paper, in addition to relaxing the perfect privacy condition of PIR, we further propose to relax the condition of perfect retrievability in order to obtain improved download rates compared to single-server WPIR. As for information-theoretic PIR, the upload cost is ignored as typically it does not scale with the file size since queries for a small file size can be reused for larger file sizes [2], [3]. From a practical perspective, the single server setup, as opposed to the multi server case, is more realistic as the noncolluding assumption is questionable in many real-world scenarios. Moreover, in several scenarios, for instance, when retrieving video, audio, or image files, allowing for a small level of distortion can be acceptable as long as the retrieved quality is high enough. In general, the range of acceptable distortion is typically limited and decided by the application and the user. In particular, we derive the *optimal* tradeoff between download rate, distortion, and information leakage when the file size is *infinite*, revealing a connection to conditional rate-distortion theory [20], [21]. Here, the information leakage is measured in terms of the average success probability for the server of correctly guessing the identity of the requested file, which can be shown to be equivalent to the MaxL metric. Moreover, we present a novel approach based on linear programming (LP) to construct schemes for a finite file size and an arbitrary number of files. When the database contains at most four bits, this approach is employed to find provably optimal schemes. These schemes can again be used to construct schemes for a larger number of files and for a larger file size. Finally, we compare the proposed approach with a nonconstructive scheme based on random coding that is adapted from [22, Cor. 17].

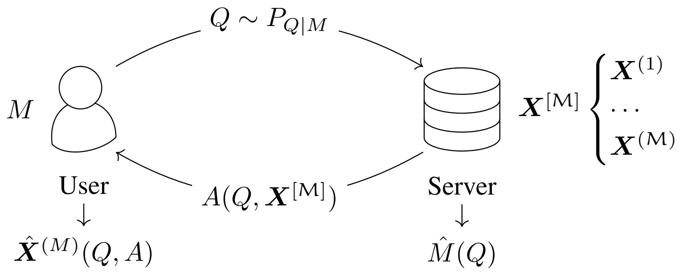
II. Preliminaries and Problem Statement

A. Notation

We denote random variables (RVs) by capital letters, e.g., X, and vectors by bold italic font, e.g., $\boldsymbol{x}$. Analogously, $\boldsymbol{X}$ denotes a random vector. Calligraphic capitals denote sets, e.g., $\mathcal{X}$. We let $[n] \triangleq \{1, 2, \ldots, n\}$ and let $\mathrm{d}_{\mathrm{H}}(x, y)$ denote the Hamming distance, which equals 0 if $x = y$, and 1 otherwise. The set of nonnegative real numbers is denoted by $\mathbb{R}_{\geq 0}$. The probability of the event "$X = x$" is denoted by $\mathbb{P}[X = x]$ and $\mathbb{E}_X[\cdot]$ denotes expectation with respect to X. The binary entropy function is denoted by $\mathsf{H}_{\mathrm{b}}(\cdot)$. P_X denotes the probability distribution of the

This work was suppported in part by US National Science Foundation awards CNS-1815322 and CNS-1526547.

978-1-5386-8209-8/21/$31.00 ©2021 IEEE



$Q \sim P_{Q|M}$
M
$\boldsymbol{X}^{[\mathsf{M}]}$
$\boldsymbol{X}^{(1)}$
$\dots$
$\boldsymbol{X}^{(\mathsf{M})}$
User
$\downarrow$
$\hat{\boldsymbol{X}}^{(M)}(Q, A)$
$A(Q, \boldsymbol{X}^{[\mathsf{M}]})$
Server
$\downarrow$
$\hat{M}(Q)$

Fig. 1. System model.

RV X, and if it is clear from the context, we sometimes drop the subscript, i.e., $P_X(\cdot) = P(\cdot)$. We write $X \sim \mathcal{U}(\mathcal{X})$ to denote that X is uniformly distributed over the set $\mathcal{X}$ and also write $X \sim P(\cdot)$ to denote that X is distributed according to $P(\cdot)$. If Y is a deterministic function of RVs $X_1, \dots, X_n$, we conventionally use Y to denote both the function and the random variable, i.e., $Y = Y(X_1, \dots, X_n)$.

Values of a discrete RV X can be encoded into variable-length binary codewords by an *optimal* lossless source code (e.g., a Huffman code). Throughout this paper, for every message x, we denote by $\ell(x)$ the codeword length in bits of x for a source code, and by $\ell^*(x)$ the length of x for an optimal code. Finally, we denote by $\mathfrak{R}_X(\mathsf{D})$ the information rate-distortion function of the source X under the distortion constraint $\mathbb{E}_{X,\hat{X}}[\mathrm{d}(X, \hat{X})] \leq \mathsf{D}$ (cf. [23, Sec. 10.2]), where $\mathrm{d}(\cdot, \cdot)$ denotes a distortion function.

B. System Model

We consider the case of a single server storing M files $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(\mathsf{M})}$, each of β symbols from $\mathcal{X}$, where $\boldsymbol{X}^{(m)} = (X_1^{(m)}, \dots, X_\beta^{(m)})$, for $m \in [\mathsf{M}]$. We assume the files are independent and identically distributed over $\mathcal{U}\left(\mathcal{X}^\beta\right)$. As a shorthand, we denote all the files together as $\boldsymbol{X}^{[\mathsf{M}]} = (\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(\mathsf{M})}) \sim \mathcal{U}\left(\mathcal{X}^{\mathsf{M}\beta}\right)$. The user wants to obtain the file with index $M \sim \mathcal{U}([\mathsf{M}])$ while keeping M to a certain extent private. Accordingly, the user generates a randomized query $Q \in \mathcal{Q}$, for some set $\mathcal{Q}$, according to a conditional distribution $P_{Q|M}(q|m)$ and sends it to the server. The conditional probabilities $P(q|m)$ are considered to be public. Based on the query Q and the files $\boldsymbol{X}^{[\mathsf{M}]}$, the server produces the response $A = A(Q, \boldsymbol{X}^{[\mathsf{M}]}) \in \mathcal{A}$, for some set $\mathcal{A}$, and sends it back to the user. Finally, the user produces an estimate of the desired file $\hat{\boldsymbol{X}}^{(M)} = \hat{\boldsymbol{X}}^{(M)}(Q, A)$, where $\hat{\boldsymbol{X}}^{(M)} = (\hat{X}_1^{(M)}, \dots, \hat{X}_\beta^{(M)}) \in \hat{\mathcal{X}}^\beta$, for some set $\hat{\mathcal{X}}$. The server produces its own guess $\hat{M} = \hat{M}(Q) \in [\mathsf{M}]$ of the index M. The overall system model is depicted in Fig. 1.

We call $\{\mathcal{X}, \hat{\mathcal{X}}, \mathcal{Q}, \mathcal{A}, \mathsf{M}, \beta, \{P(q|m)\}, A(\cdot), \{\hat{\boldsymbol{X}}^{(m)}(\cdot)\}\}$ a *lossy weakly-private information retrieval (LWPIR) scheme*. The function $\hat{M}(\cdot)$ is not part of the scheme as it can be chosen by the server freely. In the rest of the paper, we assume $\mathcal{X} = \hat{\mathcal{X}}$ is finite and therefore, without loss of generality, $\mathcal{A}$ is also finite. The user wants to retrieve the file $\boldsymbol{X}^{(M)}$ while leaking only partial information about the index M. Clearly, some information is always leaked, e.g., the very fact that the user is interested in some part of the database. The server is assumed to be *honest-but-curious*, i.e., it serves the user's requests correctly but tries to learn from them what the user is interested in.

A straightforward approach is to download all the files, yet the user wants to minimize the size of the downloaded data. On the other hand, some degree of imprecision with the data is often allowed, which can potentially improve other parameters. This results in a trifold tradeoff between: 1) the download rate, 2) the distortion between the stored data and the reconstructed data by the user, and 3) the amount of information leaked about what the user wants to download.

C. Download Rate, Distortion, and Information Leakage

In order to transmit the response A, we encode it with a lossless source code, which in general depends on the query. We define the *download rate* as

$$\begin{aligned} R &\triangleq \frac{\mathbb{E}_{A,Q}[\ell(A) \,|\, Q]}{\beta} \\ &= \frac{1}{\beta} \sum_{q \in \mathcal{Q}} P_Q(q)\, \mathbb{E}_{A|Q=q}[\ell(A) \,|\, Q = q]. \end{aligned}$$

The *distortion* of the user's reconstruction is defined as

$$D = \frac{1}{\beta} \sum_{i=1}^{\beta} \mathbb{E}_{M,Q,\boldsymbol{X}^{(M)}} \left[\mathrm{d}\left(X_i^{(M)}, \hat{X}_i^{(M)} \right) \right] = \mathbb{E}_M\big[D^{(M)}\big],$$

where $\mathrm{d}\colon \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ is a per-symbol distortion function, which is chosen based on the particular type of data considered, and

$$D^{(m)} \triangleq \frac{1}{\beta} \sum_{i=1}^{\beta} \mathbb{E}_{Q,\boldsymbol{X}^{(m)}} \left[\mathrm{d}\left(X_i^{(m)}, \hat{X}_i^{(m)} \right) \right], \quad \forall\, m \in [\mathsf{M}].$$

Note that a scheme might have distortions for some files that are worse than the average distortion D. However, it can be shown that any LWPIR scheme can be transformed into a scheme with equal distortions for all the files.

The only source of undesirable leakage is the query Q, and we denote and define the *leakage* of a given $P_{Q|M}$ as the probability of the maximum-likelihood (ML) guess as

$$L(P_{Q|M}) \triangleq \frac{1}{\mathsf{M}} \sum_{q \in \mathcal{Q}} \max_{m \in [\mathsf{M}]} P_{Q|M}(q|m).$$

It is clear that we have ${}^1\!/\mathsf{M} \leq L \leq 1$ under this privacy metric. More precisely, $L = {}^1\!/\mathsf{M}$ corresponds to the "no-leakage" case where the server cannot do anything better than randomly guess the index M. On the other hand, $L = 1$ corresponds to the "no-privacy" case where the server always guesses M correctly. We remark here that the robust MaxL metric introduced in [16], [17] is given by $\log_2\left[\mathsf{M} \cdot L(P_{Q|M})\right]$. Hence, our leakage measure is equivalent to MaxL and has a clear operational meaning.

If $\mathrm{d}_{\mathrm{H}}(x, y)$ is used as per-symbol distortion, then the distortion D is the average number of incorrectly reconstructed symbols of $\boldsymbol{X}^{(M)}$, and the best estimate is the per-symbol ML estimate $\hat{\boldsymbol{X}}^{(m)} = (\hat{X}_1^{(m)}, \dots, \hat{X}_\beta^{(m)})$, where

$$\hat{X}_i^{(m)}(q, a) \triangleq \operatorname*{arg\,max}_{y \in \mathcal{X}} \mathbb{P}\Big[A = a \,\Big|\, Q = q, X_i^{(m)} = y\Big].$$

Our goal is to characterize the minimum download rate (over all LWPIR schemes) under a given distortion constraint $D \leq \mathsf{D}$ and a given leakage level $L \leq \mathsf{L}$, for either an *infinite* or a *finite* file size β. In the finite setting of M files of length β, we denote such a minimum rate by $R^*(\mathsf{D}, \mathsf{L}; \mathsf{M}, \beta)$. For the infinite setting, we define $R^*(\mathsf{D}, \mathsf{L}; \mathsf{M}) \triangleq \lim_{\beta \to \infty} R^*(\mathsf{D}, \mathsf{L}; \mathsf{M}, \beta)$. For notational convenience, we sometimes omit the argument M if it is contextually unambiguous. Similarly to the accustomed characteristic of the PIR problem, we define the single-server LWPIR capacity as the reciprocal of the minimum download rate $R^*(\mathsf{D},\mathsf{L};\mathsf{M})$. Note that from a source coding perspective, a particular $Q=q$ fixes the lossy compressor $A_q(\boldsymbol{X}^{[\mathsf{M}]}) \triangleq A(q,\boldsymbol{X}^{[\mathsf{M}]})$. The rate of an LWPIR scheme is thus equal to the average compression rate (averaging both over Q and $\boldsymbol{X}^{[\mathsf{M}]}$).

III. Main Results

A. *Example 1: Binary Data With* $\mathsf{M}=3$ *Files and* $|\mathcal{Q}|=4$

We first present a motivating example to obtain a rate-distortion tradeoff for a fixed $P_{Q|M}$ (and thus a fixed leakage). In this example, we assume $X \sim \mathcal{U}(\{0,1\})$ (uniform binary source). It is known that $\mathfrak{R}_X(\mathsf{D}) = 1-\mathsf{H}_\mathrm{b}(\mathsf{D})$, $0 \leq \mathsf{D} \leq 1/2$.

Consider the following $P_1(q|m)$:

q	q_1	q_2	q_3	q_4
$P_1(q\|1)$	$1/4$	$1/4$	0	$1/2$
$P_1(q\|2)$	$1/4$	0	$1/4$	$1/2$
$P_1(q\|3)$	0	$1/4$	$1/4$	$1/2$

Clearly, this gives a leakage of $L(P_1(q|m)) = 1/3 \cdot (1/4+1/4+1/4+1/2) = 5/12$. Let us focus on the case of infinite β. By using a time-sharing approach, a scheme can be constructed as follows. During a fraction $\alpha = 1/2$ of the time, $R = 2\mathfrak{R}_X(\mathsf{D})$ and $L = 1/2$ can be achieved by requesting the server to randomly compress any two of the three files, and in the remaining time, $R = 3\mathfrak{R}_X(\mathsf{D})$ and $L = 1/3$ can be achieved by requesting the server to compress all the files, which gives a scheme with $R = 1/2\bigl[2\mathfrak{R}_X(\mathsf{D})\bigr] + 1/2\bigl[3\mathfrak{R}_X(\mathsf{D})\bigr]$ and $L = 5/12$. However, as shown below, this simple scheme is only suboptimal.

B. *Minimum Download Rate (Infinite* β*)*

In this subsection, we derive the minimum download rate $R^*(\mathsf{D},\mathsf{L};\mathsf{M})$.[1] We express it as a solution of an optimization problem, where the objective function is a weighted sum of rate-distortion functions.

Theorem 1: The minimum download rate $R^*(\mathsf{D},\mathsf{L};\mathsf{M})$ of LWPIR with M files is the minimum value of the optimization problem

$$\min_{P_{Q|M}} \min_{\{\mathsf{D}_q^{(m)}\}} \quad \sum_{q\in\mathcal{Q}} P_Q(q) \sum_{m\in[\mathsf{M}]} \mathfrak{R}_X(\mathsf{D}_q^{(m)}) \tag{1a}$$

$$\text{s.t.} \quad \frac{1}{\mathsf{M}} \sum_{q\in\mathcal{Q}} \max_{m\in[\mathsf{M}]} P_{Q|M}(q|m) \leq \mathsf{L}, \tag{1b}$$

$$\frac{1}{\mathsf{M}} \sum_{q\in\mathcal{Q}} \sum_{m\in[\mathsf{M}]} P_{Q|M}(q|m)\mathsf{D}_q^{(m)} \leq \mathsf{D}, \tag{1c}$$

where $|\mathcal{Q}| \leq \mathsf{M}+3$.

The detailed proof can be found in the extended version. Here, we sketch the main steps of the proof as follows. First, using the optimal solutions of (1), it can be seen that the user can compress each file $\boldsymbol{X}^{(m)}$ with a given distortion $\mathsf{D}_q^{(m)}$, and thus the corresponding rate $\mathfrak{R}_X\bigl(\mathsf{D}_q^{(m)}\bigr)$ is achievable as $\beta \to \infty$. Using a time-sharing approach, we can see that the right-hand side of (1) is achievable. The converse part is shown by combining the standard converse proof for the rate-distortion function and the approaches of conditional rate-distortion theory [21]. Finally, the upper bound on the size of $\mathcal{Q}$ is proved by applying Carathéodory's theorem [23, Thm. 15.3.5].

We remark here that for the special case where $\mathsf{D}=0$ and the so-called *normal distortion measure* is used [24], it can be shown that the optimal LWPIR scheme from Theorem 1 is the WPIR scheme presented in [13, Sec. V].

In fact, given an arbitrary conditional distribution $P_{Q|M}$, the inner minimization over the variables $\{\mathsf{D}_q^{(m)}\}$ in Theorem 1 can be solved, as stated in the following corollary.

Corollary 1: Assume that the rate-distortion function $\mathfrak{R}_X(\cdot)$ is differentiable in D. Then,

$$R^*(\mathsf{D},\mathsf{L};\mathsf{M}) = \min_{\substack{P_{Q|M}:\\ L(P_{Q|M})\leq\mathsf{L}}} \sum_{q\in\mathcal{Q}} P_Q(q) \sum_{m\in[\mathsf{M}]} \mathfrak{R}_X\bigl(\mathsf{D}_q^{(m)*}\bigr),$$

where the values $\mathsf{D}_q^{(m)*}$, $m\in[\mathsf{M}]$, $q\in\mathcal{Q}$, satisfy

$$\begin{aligned} &\frac{P_Q(q)}{P(m,q)} \frac{\mathrm{d}\mathfrak{R}_X}{\mathrm{d}\mathsf{D}}\Big|_{\mathsf{D}=\mathsf{D}_q^{(m)*}} = \lambda, \\ &\qquad \forall\, m\in[\mathsf{M}], q\in\mathcal{Q}, \text{such that } \mathsf{D}_q^{(m)*} > 0, \\ &\frac{P_Q(q)}{P(m,q)} \frac{\mathrm{d}\mathfrak{R}_X}{\mathrm{d}\mathsf{D}}\Big|_{\mathsf{D}=\mathsf{D}_q^{(m)*}} \geq \lambda, \\ &\qquad \forall\, m\in[\mathsf{M}], q\in\mathcal{Q}, \text{such that } \mathsf{D}_q^{(m)*} = 0. \end{aligned} \tag{2}$$

The Lagrange multiplier λ must be chosen such that

$$\sum_{m\in[\mathsf{M}]} \sum_{q\in\mathcal{Q}} P(m,q)\mathsf{D}_q^{(m)*} = \mathsf{D}. \tag{3}$$

Proof: The rate-distortion function $\mathfrak{R}_X(\cdot)$ is nonincreasing, convex, and continuous (see, e.g., [25, Ch. 3]). Fix a feasible $P_{Q|M}$ in the optimization problem (1), and thus P_Q is also fixed. Then, the objective function $\sum_{q\in\mathcal{Q}} P_Q(q) \sum_{m\in[\mathsf{M}]} \mathfrak{R}_X(\mathsf{D}_q^{(m)})$ is a nonnegative weighted sum of convex functions, and therefore it is convex (cf. [26, Sec. 3.2.1]). The result then follows immediately from the Karush–Kuhn–Tucker optimality conditions for convex minimization problems [26, Sec. 5.5.3]. ■

The following corollary gives expressions for the minimum download rate in two special cases.

Corollary 2: The minimum download rate $R^*(\mathsf{D},\mathsf{L};\mathsf{M})$ of LWPIR with M files in the "no-leakage" and "no-privacy" special cases are $R^*(\mathsf{D},\mathsf{L}=1/\mathsf{M};\mathsf{M}) = \mathsf{M}\mathfrak{R}_X(\mathsf{D})$ and $R^*(\mathsf{D},\mathsf{L}=1;\mathsf{M}) = \mathfrak{R}_X(\mathsf{D})$, respectively.

C. *Example 1 (Continued)*

We now present a scheme by using Corollary 1 to obtain the optimal rate-distortion tradeoff for Example 1 (i.e., for the given $P_1(q|m)$). Note that $\mathrm{d}\mathfrak{R}_X/\mathrm{d}\mathsf{D} = \log_2(\mathsf{D}/(1-\mathsf{D}))$.

From Corollary 1, we get the optimal solution of $\{\mathsf{D}_q^{(m)*}\}_{m\in[3],q\in\mathcal{Q}}$ as follows:

q	q_1	q_2	q_3	q_4
$\mathsf{D}_q^{(1)*}$	D_1^*	D_1^*	0	D_2^*
$\mathsf{D}_q^{(2)*}$	D_1^*	0	D_1^*	D_2^*
$\mathsf{D}_q^{(3)*}$	0	D_1^*	D_1^*	D_2^*

[1] The results for minimum download rate automatically give analogous results for capacity.

Furthermore, from (2) and (3), D_1^* and D_2^* should satisfy

$$\frac{2}{1}\frac{\mathrm{d}\mathfrak{R}_X}{\mathrm{d}\mathsf{D}}\big(\mathsf{D}_1^*\big) = \frac{3}{1}\frac{\mathrm{d}\mathfrak{R}_X}{\mathrm{d}\mathsf{D}}\big(\mathsf{D}_2^*\big),$$
$$\frac{1}{2}\mathsf{D}_1^* + \frac{1}{2}\mathsf{D}_2^* = \mathsf{D},$$

for D_1^*, $\mathsf{D}_2^* > 0$. From this, the optimal solution D_1^* is the (unique) root of $\big({}^1\!/_{(2\mathsf{D}-\mathsf{D}_1^*)}-1\big)^{3/2}-{}^1\!/_{\mathsf{D}_1^*}+1$ and $\mathsf{D}_2^* = 2\mathsf{D}-\mathsf{D}_1^*$.

D. Equivalent Linear Programming Formulation (Finite β)

In this subsection, we show how to transform the problem of finding $R^*(\mathsf{D},\mathsf{L};\mathsf{M},\beta)$ into an LP. The solution of the LP provably provides the value of $R^*(\mathsf{D},\mathsf{L};\mathsf{M},\beta)$ and a corresponding scheme.

First, we concentrate on response functions. For a fixed value of the query $Q = q$, it is a deterministic function

$$A_q(\boldsymbol{x}) : \mathcal{X}^{\mathsf{M}\beta} \to \mathcal{A}$$

of $\mathsf{M}\beta$ symbols. A fixed response function $A_q(\cdot)$ defines a random variable $A_q(\boldsymbol{X}^{[\mathsf{M}]})$ with rate

$$R_q \triangleq \frac{\mathbb{E}_{\boldsymbol{X}^{[\mathsf{M}]}}\left[\ell^*\left(A_q(\boldsymbol{X}^{[\mathsf{M}]})\right)\right]}{\beta}$$

and per-file distortion

$$D_q^{(m)} \triangleq \frac{1}{\beta}\sum_{i=1}^{\beta}\mathbb{E}_{\boldsymbol{X}^{(m)}}\Big[\mathrm{d}\big(X_i^{(m)},\hat{X}_i^{(m)}\big) \mid Q = q\Big], \forall\, m \in [\mathsf{M}].$$

An important observation is that for a fixed $Q = q$, the values R_q and $D_q^{(m)}$, $m \in [\mathsf{M}]$, are only determined by the function $A_q(\cdot)$ and do not depend on the way the user generates other queries. Thus, we can express the rate and distortion of an LWPIR scheme as

$$R = \mathbb{E}_Q\left[R_Q\right] = \frac{1}{\mathsf{M}}\sum_{m\in[\mathsf{M}]}\sum_{q\in\mathcal{Q}}P(q|m)R_q,$$
$$D = \mathbb{E}_{M,Q}\left[D_Q^{(M)}\right] = \frac{1}{\mathsf{M}}\sum_{m\in[\mathsf{M}]}\sum_{q\in\mathcal{Q}}P(q|m)D_q^{(m)},$$

and the leakage of the ML estimate by the server as

$$L = \frac{1}{\mathsf{M}}\sum_{q\in\mathcal{Q}}\max_{m\in[\mathsf{M}]}P(q|m).$$

Now, assume all the possible response functions $\{A_q \mid q \in \mathcal{Q}\}$ defined over $\mathcal{X}^{\mathsf{M}\beta}$ are given, and their corresponding R_q, $D_q^{(m)}$, $m \in [\mathsf{M}]$, are pre-calculated. Then, the problem of rate minimization can be formulated as an LP in the decision variables $\mathcal{P} \triangleq \{P(q|m) \mid m \in [\mathsf{M}], q \in \mathcal{Q}\}$ for each pair of target values of distortion D and leakage L, as

$$\begin{aligned}
\min_{\mathcal{P}}\quad & \frac{1}{\mathsf{M}}\sum_{m\in[\mathsf{M}]}\sum_{q\in\mathcal{Q}}P(q|m)R_q \\
\text{s.t.}\quad & \sum_{q\in\mathcal{Q}}P(q|m) = 1, && m \in [\mathsf{M}], \\
& \frac{1}{\mathsf{M}}\sum_{m\in[\mathsf{M}]}\sum_{q\in\mathcal{Q}}P(q|m)D_q^{(m)} \leq \mathsf{D}, \\
& \frac{1}{\mathsf{M}}\sum_{q\in\mathcal{Q}}\xi_q \leq \mathsf{L}, \\
& 0 \leq P(q|m) \leq \xi_q, && q \in \mathcal{Q},\ m \in [\mathsf{M}].
\end{aligned}$$

Here, the auxiliary variables ξ_q are introduced to model the behavior of the $\max$ function.

To obtain an optimal solution, one needs in principle to consider all possible $(|\mathcal{X}|^{\mathsf{M}\beta})^{|\mathcal{X}|^{\mathsf{M}\beta}}$ response functions, which grows super-exponentially in M and β. However, many of them are in fact equivalent up to a permutation of $\mathcal{A}$. Additionally, many functions can be filtered out as they are inferior to other functions, i.e., they have higher rate and higher per-file distortion for all files, than some other response function. Alternatively, one can restrict the response functions to a subclass, for which the values of R_q and $D_q^{(m)}$ are relatively easy to calculate. Using these functions in the LP, we can find provably optimal solutions in this restricted case. The schemes found can be used as suboptimal constructive schemes for the original (unrestricted) problem.

As a remark, we note that since the LWPIR problem can be reformulated as an LP, the optimal tradeoff curve $R^*(\mathsf{D},\mathsf{L};\mathsf{M},\beta)$ is a piecewise-linear, convex, decreasing function of D when L is fixed (or vice versa) and $\mathcal{X}$ is finite. This follows since the number of response functions is finite for a finite $\mathcal{X}$.

E. Optimal Tradeoff for $\mathsf{M} = 2$, $\beta = 2$, and $\mathcal{X} = \{0, 1\}$

To illustrate the LP method in Section III-D, we present more details for the case of $\mathsf{M} = 2$ files and $\beta = 2$ bits. Since the input to each A_q is 4 bits in total, there are not more than 2^4 different elements in its image and thus the size of $\mathcal{A}$ can be limited to 2^4. Therefore, the number of different functions from $\{0,1\}^4$ to $\mathcal{A}$ is $(2^4)^{2^4}$. Discarding all the equivalent ones, we have roughly 10^{10}, which further drops to 3457 after the filtering. An LP of this size can be solved easily, e.g., by using Gurobi [27]. Moreover, the LP can be solved symbolically which gives a closed-form expression for the optimal tradeoff.

Theorem 2 (Optimal tradeoff for $\mathsf{M} = 2$, $\beta = 2$, $\mathcal{X} = \{0, 1\}$): For $\mathsf{M} = 2$ files each of $\beta = 2$ bits, the minimum rate is

$$R^*(\mathsf{D},\mathsf{L};\beta) = \begin{cases}
-\frac{11}{2}\mathsf{D} + 3 - 2\mathsf{L}, & \text{if } \mathsf{D} \in \left[0, \frac{1-L}{4}\right], \\
-5\mathsf{D} + \frac{23-15\mathsf{L}}{8}, & \text{if } \mathsf{D} \in \left[\frac{1-\mathsf{L}}{4}, \frac{3(1-\mathsf{L})}{8}\right], \\
-4\mathsf{D} + \frac{5-3\mathsf{L}}{2}, & \text{if } \mathsf{D} \in \left[\frac{3(1-\mathsf{L})}{8}, \frac{5(1-\mathsf{L})}{8}\right], \\
-\frac{8}{3}\mathsf{D} + \frac{5-2\mathsf{L}}{3}, & \text{if } \mathsf{D} \in \left[\frac{5(1-\mathsf{L})}{8}, 1 - \mathsf{L}\right], \\
-2\mathsf{D} + 1, & \text{if } \mathsf{D} \in \left[1 - \mathsf{L}, \frac{1}{2}\right].
\end{cases}$$

Using the same approach, we can find closed-form expressions for all the binary cases with $\mathsf{M}\beta \leq 4$. However, efficient filtering of potential response functions in general remains an open question for future research.

IV. Finite-Size LWPIR Schemes

In this section, we present some schemes for $\mathcal{X} = \{0, 1\}$.

A. LWPIR Schemes From Small Optimal Schemes

We can further use the optimal schemes obtained for $\mathsf{M}\beta \leq 4$ with the LP method (see Section III-E) in order to construct schemes for larger M and β. First, the longer files can be split into smaller blocks and then a small scheme is run on the corresponding blocks. The resulting scheme has the same rate, distortion, and leakage as the small scheme. Second, a large set of files can be split into subsets and a small scheme is run on

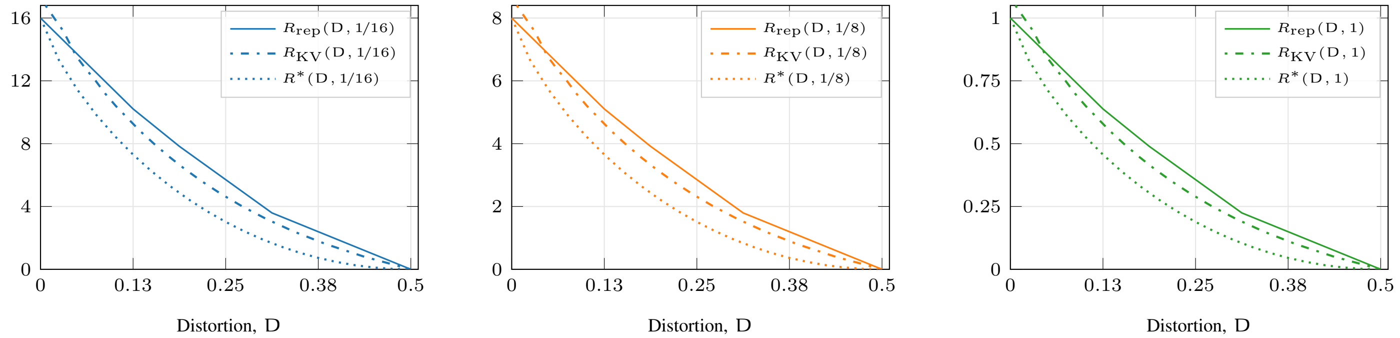


Fig. 2. Download rate versus distortion of two finite-size LWPIR schemes for $\mathsf{M} = 16$, $\beta = 20$ bits, and leakage $\mathsf{L} \in \{1/16, 1/8, 1\}$. For comparison, the corresponding asymptotic curves $R^*(\mathsf{D}, \mathsf{L})$, obtained from Theorem 1 and Corollary 2, are also depicted.

each subset of the files. The construction gives a higher rate but smaller leakage, while distortion does not change.

We note that in both approaches the user needs to generate the query only once and can reuse it for all instances of the small scheme. Additionally, all the answers obtained from the small schemes can be re-coded together, thus potentially decreasing the overall rate. Finally, the two aforementioned constructions can be combined.

For example, consider an optimal scheme $\mathcal{S}$ for $\mathsf{M} = 2$ files of $\beta = 2$ bits each with leakage $\mathsf{L} = 1/2$, distortion $\mathsf{D} = 5/16$, and rate $\mathsf{R} = 1/2$. The scheme uses only one response function (i.e., $|\mathcal{Q}| = 1$), which outputs $\mathtt{0}$ if $\boldsymbol{X}^{[2]} \in \{1100, 1010, 0110, 1110, 1111\}$, and $\mathtt{1}$ otherwise. Note that $\mathbb{P}[A = \mathtt{0}] = 1 - \mathbb{P}[A = \mathtt{1}] = 5/16$. The response $\mathtt{0}$ is decoded to the estimates $(\hat{\boldsymbol{X}}^{(1)}, \hat{\boldsymbol{X}}^{(2)}) = 1110$, and $\mathtt{1}$ to 0001. If we need a scheme $\mathcal{S}'$ for $\mathsf{M}' = 2$, $\beta' = 20$, and $\mathsf{L}' = 1/2$, we split 20-bit files into 2-bit blocks, and repeat the small scheme 10 times. The server's response is a $\mathtt{0}/\mathtt{1}$-string of length 10. We can encode this string with a Huffman code, which has an average length of 8.99 bits. The rate of the new scheme is $\mathsf{R}' = 8.99/20 = 0.45$ and the distortion is $\mathsf{D}' = \mathsf{D} = 5/16$. Further, we construct a scheme $\mathcal{S}''$ for $\mathsf{M}'' = 16$, $\beta'' = 20$, and $\mathsf{L}'' = 1/16$ as follows. Files are split as $\boldsymbol{X}^{[16]} = ((\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}), (\boldsymbol{X}^{(3)}, \boldsymbol{X}^{(4)}), \ldots, (\boldsymbol{X}^{(15)}, \boldsymbol{X}^{(16)}))$, and the scheme $\mathcal{S}'$ runs on each of these pairs of files. If the user wants file $\boldsymbol{X}^{(4)}$, she keeps the answer corresponding to $(\boldsymbol{X}^{(3)}, \boldsymbol{X}^{(4)})$ and obtains $\hat{\boldsymbol{X}}^{(4)}$ with distortion $\mathsf{D}'' = \mathsf{D}'$. The rate is $\mathsf{R}'' = 8\mathsf{R}' = 3.60$ and the leakage is $\mathsf{L}'' = 1/8\mathsf{L}' = 1/16$.

In a similar manner, we constructed other schemes for $\mathsf{M} = 16$ and $\beta = 20$ from the optimal schemes obtained with the LP method for $\mathsf{M}\beta \leq 4$. The corresponding rate-distortion-leakage curves are labeled by $R_{\text{rep}}(\mathsf{D}, \mathsf{L})$ in Fig. 2. For comparison, we also plot the corresponding asymptotic curves $R^*(\mathsf{D}, \mathsf{L})$ (i.e., when $\beta \to \infty$), obtained from Corollary 2 (for $\mathsf{L} = 1$ and $1/16$) or by numerically solving the optimization problem in Theorem 1 with Gurobi [27] (for $\mathsf{L} = 1/8$). The asymptotic curves serve as lower bounds on the finite-size curves.

B. Achievable LWPIR Rates From Lossy Compressors

Another way to construct LWPIR schemes is from lossy compressors, as described below.

Consider the following (not necessarily optimal) LWPIR scheme for M files of β bits each. The user chooses uniformly at random a subset $\mathcal{I} \subset [\mathsf{M}]$ of cardinality $|\mathcal{I}| = \mathsf{N}$ such that the index of interest M is in $\mathcal{I}$, and sends $\mathcal{I}$ to the server. The server concatenates the files indexed by $\mathcal{I}$ into a block of $\mathsf{N}\beta$ bits, compresses it with some pre-agreed lossy compressor, and sends it back to the user. The user reconstructs (with distortion) the compressed files and keeps only the desired one, $\hat{\boldsymbol{X}}^{(M)}$. The remaining $\mathsf{N} - 1$ files have been requested only to trick the server and are thus discarded. From the server's perspective, M is uniformly distributed over $\mathcal{I}$, and thus the leakage is $1/\mathsf{N}$. The rate and the distortion follow from the properties of the chosen lossy compressor. This approach is general and works for both finite and infinite β. We remark that by time-sharing schemes with different values of N, we can construct LWPIR schemes for arbitrary leakage levels (not only reciprocals of integers).

Now, we only need a lossy compressor. In the context of finite-length information theory, the authors in [22] derived an achievable rate of lossy compression for any finite block length β and thus, proved existence of a corresponding lossy compressor. The special case of the source $\mathcal{X} \sim \mathcal{U}(\{0, 1\})$ is addressed in [22, Cor. 17]. However, the fidelity criterion used in [22, Cor. 17] is the excess-distortion probability, while most of the works in rate-distortion theory, as well as our work, focus on the average distortion. Since for a nonnegative RV Z, it holds that $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}[Z > z]\, \mathrm{d}z$, we can derive results for the average distortion criterion from the results in [22].

In Fig. 2, we plot achievable rate curves (labeled by $R_{\text{KV}}(\mathsf{D}, \mathsf{L})$) corresponding to LWPIR schemes constructed as described above using the lossy compressor from [22, Cor. 17]. Although the achievable rates are in general, except for very low distortion values, lower than those of the schemes from Section IV-A (labeled by $R_{\text{rep}}(\mathsf{D}, \mathsf{L})$), they come at the price of being nonconstructive, since the achievable rates presented in [22, Cor. 17] are based on random coding arguments.

V. Conclusion

We proposed to simultaneously relax the conditions of perfect retrievability and privacy of standard PIR, referred to as LWPIR, in order to obtain improved download rates in the single server scenario. In particular, the optimal rate-distortion-leakage tradeoff was established for an arbitrary number of asymptotically large files. Moreover, we presented an approach based on LP to construct schemes for a finite file size and an arbitrary number of files. When the database contains at most four bits, the approach allows to obtain provably optimal LWPIR schemes.

References

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 36th Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Milwaukee, WI, USA, Oct. 23–25, 1995, pp. 41–50.

[2] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 14–19, 2015, pp. 2842–2846.

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

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

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

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

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

[8] E. Kushilevitz and R. Ostrovsky, "Replication is not needed: Single database, computationally-private information retrieval," in *Proc. 38th Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Miami Beach, FL, USA, Oct. 20–22, 1997, pp. 364–373.

[9] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.

[10] S. Li and M. Gastpar, "Single-server multi-user private information retrieval with side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 17–22, 2018, pp. 1954–1958.

[11] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "The role of coded side information in single-server private information retrieval," *IEEE Trans. Inf. Theory*, vol. 67, no. 1, pp. 25–44, Jan. 2021.

[12] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "Weakly-private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1257–1261.

[13] ——, "The capacity of single-server weakly-private information retrieval," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 415–427, Mar. 2021.

[14] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1262–1266.

[15] I. Samy, M. A. Attia, R. Tandon, and L. Lazos, "Asymmetric leaky private information retrieval," Jun. 2020, arXiv:2006.03048v1 [cs.IT]. [Online]. Available: https://arxiv.org/abs/2006.03048

[16] G. Smith, "On the foundations of quantitative information flow," in *Proc. 12th Int. Conf. Found. Softw. Sci. Comput. Struct. (FoSSaCS)*, York, U.K., Mar. 22–29, 2009, pp. 288–302.

[17] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[18] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "Multi-server weakly-private information retrieval," Jul. 2020, arXiv:2007.10174v2 [cs.IT]. [Online]. Available: https://arxiv.org/abs/2007.10174

[19] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, USA, Jun. 21–26, 2020, pp. 1089–1094.

[20] R. M. Gray, "Conditional rate-distortion theory," Stanford Electronics Laboratories, Stanford, CA, USA, Tech. Rep. 6502-2, Oct. 1972.

[21] ——, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 480–489, Jul. 1973.

[22] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, Jun. 2012.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[24] R. W. Yeung, *Information Theory and Network Coding*. Boston, MA, USA: Springer, 2008.

[25] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge University Press, 2011.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[27] Gurobi Optimization, LLC, "Gurobi optimizer reference manual," 2021. [Online]. Available: http://www.gurobi.com