The Capacity of Single-Server Weakly-Private Information Retrieval

Hsuan-Yin Lin[*], **Siddhartha Kumar**[*], **Eirik Rosnes**[*], **Alexandre Graell i Amat**[†*], and **Eitan Yaakobi**[‡]
[*]Simula UiB, N–5006 Bergen, Norway
[†]Department of Electrical Engineering, Chalmers University of Technology, SE–41296 Gothenburg, Sweden
[‡]Department of Computer Science, Technion — Israel Institute of Technology, Haifa, 3200003 Israel

***Abstract*—Weakly-private information retrieval (WPIR) is a variant of the private information retrieval problem in which a user wants to efficiently retrieve a file stored across a set of servers while tolerating some information leakage on the identity of the requested file to the servers. In this paper, we consider WPIR from a single-server database where the information leakage is measured in terms of the mutual information (MI) or maximal leakage (MaxL) privacy metrics. In particular, we establish a connection between the WPIR problem and rate-distortion theory, and fully characterize the optimal tradeoff between the download cost and the allowed information leakage under the MI and MaxL metrics, settling the single-server WPIR capacity.**

I. Introduction

With the ever increasing demand for privacy rights on the Internet, preserving user privacy has become of vital importance for, e.g., big data applications and distributed information systems. The European Union's General Data Protection Regulation is an important example that demonstrates the increasing awareness of this issue.

Private information retrieval (PIR), introduced by Chor *et al.* [1], is one of the fundamental primitives that ensures user privacy. In particular, it enables a user to download a desired file stored in one or several servers without disclosing the identity of the requested file to any of the servers. In recent years, designing efficient PIR schemes from an information-theoretic perspective has attracted significant attention [2], [3]. In this line of research, the efficiency is mainly measured in terms of download rate, defined as the ratio between the requested file size and the expected number of downloaded symbols. The maximum possible download rate is referred to as the *PIR capacity*.

Characterizing the capacity for variants of the PIR problem is of fundamental importance. Sun and Jafar derived the PIR capacity for the classical case where data is replicated across multiple servers [2]. For the case when data is encoded by a maximum distance separable (MDS) code and then stored in multiple servers, the capacity, referred to as the *MDS-PIR capacity*, was given in [3]. The PIR capacity when the storage code is from a particular family of non-MDS codes, introduced in [4], was determined and shown to be equal to the MDS-PIR capacity in [5]. Other PIR capacity results for various scenarios can be found in, e.g., [6]–[9].

Recently, the notion of weakly-private information retrieval (WPIR) was introduced independently by the works [10]–[12]. WPIR can be seen as a generalization of PIR, which guarantees the retrieval of a single file while leaking partial information about the identity of the requested file to the servers. For the scenario with multiple noncolluding servers, it was shown that the download rate as well as the upload cost and access complexity can be improved by trading off the information leakage.

In practice, the servers are typically in control of a single provider, violating the assumption that they can not collude. Motivated by this important and practical observation, in this paper we turn our attention and initiate the study of the more compelling scenario in which data is stored in a single server, i.e., the problem of single-server WPIR is considered. Similar to [11], to measure the information leakage in an information-theoretic sense, we adopt two commonly-used information leakage metrics in the information theory literature, the mutual information (MI) and maximal leakage (MaxL) metrics [13]–[15]. The latter is known as one of the most useful information-theoretic measures of information leakage in the computer security literature. In particular, by developing the similarities between the WPIR problem and rate-distortion theory, we fully characterize the optimal tradeoff between the download cost and a given level of information leakage for the MI and MaxL privacy metrics, and hence determine the single-server WPIR capacity. Moreover, a novel single-server WPIR capacity-achieving scheme for both metrics is constructed by using a time-sharing approach. Due to lack of space, all technical proofs are omitted, and we refer to the extended version of this work [16] for full details as well as the converse proofs.

II. Preliminaries and Problem Statement

A. Notation

We denote by $\mathbb{N}$ the set of all positive integers, $[a] \triangleq \{1, 2, \ldots, a\}$, and $[a : b] \triangleq \{a, a+1, \ldots, b\}$ for $a, b \in \{0\} \cup \mathbb{N}$ and $a \leq b$. The set of nonnegative real numbers is denoted by $\mathbb{R}_+$. Vectors are denoted by bold letters and sets by calligraphic uppercase letters, e.g., $\boldsymbol{x}$ and $\mathcal{X}$, respectively. In general, vectors are represented as row vectors throughout the paper. We use uppercase letters for random variables (RVs)

This work was partially funded by the Swedish Research Council (grant 2016-04253), the Israel Science Foundation (grant #1817/18), and by the Technion Hiroshi Fujiwara Cyber Security Research Center and the Israel National Cyber Directorate.

(either scalar or vector), e.g., X or $\boldsymbol{X}$. For a given index set $\mathcal{S}$, we write $X^{\mathcal{S}}$ to represent $\{X^{(m)}: m \in \mathcal{S}\}$. The Hamming weight of a vector $\boldsymbol{x}$ is denoted by $w_{\mathsf{H}}(\boldsymbol{x})$, while its support will be denoted by $\chi(\boldsymbol{x})$. $\mathbb{E}_X[\cdot]$ and $\mathbb{E}_{P_X}[\cdot]$ denote expectation with respect to the RV X and distribution P_X, respectively. $\mathsf{H}(X)$, $\mathsf{H}(P_X)$, or $\mathsf{H}\big(p_1, \ldots, p_{|\mathcal{X}|}\big)$ represents the entropy of X, where $P_X(\cdot) = (p_1, \ldots, p_{|\mathcal{X}|})$ denotes the distribution of the RV X. $\mathsf{I}(X;Y)$ denotes the MI between X and Y. The Galois field with q elements is denoted by $\mathrm{GF}(q)$.

B. System Model

We consider a single server that stores M independent files $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(\mathsf{M})}$, where each file $\boldsymbol{X}^{(m)} = \big(X_1^{(m)}, \ldots, X_\beta^{(m)}\big)$, $m \in [\mathsf{M}]$, is represented as a length-β row vector over $\mathrm{GF}(q)$. Assume that each element of $\boldsymbol{X}^{(m)}$ is chosen independently and uniformly at random from $\mathrm{GF}(q)$. Thus, in q-ary units, we have $\mathsf{H}\big(\boldsymbol{X}^{(m)}\big) = \beta$, $\forall\, m \in [\mathsf{M}]$. A user wishes to efficiently retrieve $\boldsymbol{X}^{(M)}$ by allowing some information leakage to the server, where the requested file index M is assumed to be uniformly distributed over $[\mathsf{M}]$. We give the following definition for a single-server WPIR scheme.

Definition 1. *An M-file WPIR scheme $\mathscr{C}$ for a single server storing M files consists of:*

- *A random strategy $\boldsymbol{S}$, whose alphabet is $\mathcal{S}$.*
- *A query function*

$$\phi\colon \{1, \ldots, \mathsf{M}\} \times \mathcal{S} \to \mathcal{Q}$$

that generates a query $\boldsymbol{Q} = \phi(M, \boldsymbol{S})$ with alphabet $\mathcal{Q}$. The query $\boldsymbol{Q}$ is sent to the server to retrieve the M-th file.

- *An answer function*

$$\varphi\colon \mathcal{Q} \times \mathrm{GF}(q)^{\beta\mathsf{M}} \to \mathcal{A}^{\beta L}$$

that returns the answer $\boldsymbol{A} \triangleq \varphi(\boldsymbol{Q}, \boldsymbol{X}^{[\mathsf{M}]})$ back to the user, with download symbol alphabet $\mathcal{A}$. Here, $L = L(\boldsymbol{Q})$ is the normalized length of the answer, which is a function of the query $\boldsymbol{Q}$.

- *A privacy leakage metric $\rho^{(\cdot)}(P_{\boldsymbol{Q}|M}) \geq 0$, which is defined as a function of the conditional probability mass function (PMF) $P_{\boldsymbol{Q}|M}$, that measures the amount of leaked information of the identity of the requested file to the server by observing the generated query $\boldsymbol{Q}$, where the superscript $(\cdot)$ indicates the used metric.*

This scheme must satisfy the condition of perfect retrievability,

$$\mathsf{H}\big(\boldsymbol{X}^{(M)} \,\big|\, \boldsymbol{A}, \boldsymbol{Q}, M\big) = 0. \tag{1}$$

We remark that a PIR scheme is equivalent to a WPIR scheme for which no information leakage is allowed.

C. Metrics of Information Leakage

Given a single-server M-file WPIR scheme and a fixed distribution P_M, its designed query conditional PMF given the index M of the requested file, $P_{\boldsymbol{Q}|M}$, can be seen as a privacy mechanism (a randomized mapping). The server receives the random outcome $\boldsymbol{Q}$ of the privacy mechanism $P_{\boldsymbol{Q}|M}$, and is curious about the index M of the requested file. The information leakage of a WPIR scheme is then measured with respect to its corresponding privacy mechanism $P_{\boldsymbol{Q}|M}$.

In this paper, we focus on two commonly-used information-theoretic measures, namely MI and MaxL. For the former, the information leakage is quantified by

$$\rho^{(\mathsf{MI})}(P_{\boldsymbol{Q}|M}) \triangleq \mathsf{I}(M;\boldsymbol{Q}). \tag{2}$$

The second privacy metric, MaxL, can be defined based on the min-entropy (MinE) measure discussed in the computer science literature, see, e.g., [13]. Moreover, since we assume that M is uniformly distributed, the MinE information leakage and the MaxL privacy metric can be shown to be equivalent [14], [15]. The latter is given by

$$\mathsf{MaxL}(M;\boldsymbol{Q}) \triangleq \log_2 \sum_{\boldsymbol{q} \in \mathcal{Q}} \max_{m \in [\mathsf{M}]} P_{\boldsymbol{Q}|M}(\boldsymbol{q}|m). \tag{3}$$

Henceforth, we will use the closed-form expression in (3) as the MaxL privacy metric and denote the MaxL privacy metric of a WPIR scheme by

$$\rho^{(\mathsf{MaxL})}(P_{\boldsymbol{Q}|M}) \triangleq \mathsf{MaxL}(M;\boldsymbol{Q}).$$

In the following, the information leakage metric of a WPIR scheme $\mathscr{C}$ is denoted by $\rho^{(\cdot)}(\mathscr{C})$.

D. Download Cost and Rate for a Single-Server WPIR Scheme

For WPIR, in contrast to PIR, the download costs for the retrieval of different files do not necessarily need to be the same. Hence, the download cost is defined as the expected download cost over all possible requested files. The download cost of a single-server WPIR scheme $\mathscr{C}$ for the retrieval of the m-th file, denoted by $\mathsf{D}^{(m)}(\mathscr{C})$, is defined as the expected length of the returned answer over all random queries,

$$\mathsf{D}^{(m)}(\mathscr{C}) \triangleq \mathbb{E}_{P_{\boldsymbol{Q}|M=m}}[L(\boldsymbol{Q})],$$

and the overall download cost is measured in terms of the expected download cost over all files, i.e.,

$$\mathsf{D}(\mathscr{C}) \triangleq \mathbb{E}_{P_M}\big[\mathbb{E}_{P_{\boldsymbol{Q}|M}}[L(\boldsymbol{Q})]\big] = \mathbb{E}_{M,\boldsymbol{Q}}[L(\boldsymbol{Q})]. \tag{4}$$

Accordingly, the WPIR rate is defined as $\mathsf{R}(\mathscr{C}) \triangleq \mathsf{D}(\mathscr{C})^{-1}$.

In this paper, our goal is to characterize the optimal trade-off between the download cost and the allowed information leakage with respect to a privacy metric. We start with the following definition of an achievable download-leakage pair.

Definition 2. *Consider a single server that stores M files. A download-leakage pair (D, ϱ) is said to be* achievable *in terms of the information leakage metric $\rho^{(\cdot)}$ if there exists a WPIR scheme $\mathscr{C}$ such that $\mathbb{E}_{M,\boldsymbol{Q}}[L(\boldsymbol{Q})] \leq \mathsf{D}$ and $\rho^{(\cdot)}(\mathscr{C}) \leq \varrho$. The* download-leakage *region is the set of all achievable download-leakage pairs (D, ϱ).*

III. Characterization of the Optimal Download-Leakage Tradeoff

Consider a single-server WPIR scheme, where the leakage is measured by $\rho^{(\mathsf{MI})}$ or $\rho^{(\mathsf{MaxL})}$. The minimum achievable download cost for a given leakage constraint ϱ can be formulated by the optimization problem

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{P_M P_{\boldsymbol{Q}|M}}[L(\boldsymbol{Q})] \\ \text{subject to} \quad & P_{\boldsymbol{Q}|M} \in \mathcal{P}_{\mathsf{ret}}, \qquad (5a) \\ & \rho^{(\cdot)}(P_{\boldsymbol{Q}|M}) \le \varrho, \qquad (5b) \end{aligned}$$

where $\mathcal{P}_{\mathsf{ret}}$ is defined as the set of all PMFs that satisfy (1).

For convenience, since we know that a designed conditional distribution $P_{\boldsymbol{Q}|M}$ of a WPIR scheme always satisfies (5a), throughout this paper we will assume that any $P_{\boldsymbol{Q}|M} \in \mathcal{P}_{\mathsf{ret}}$.

A. The Download-Leakage Function for Single-Server WPIR

To characterize the optimal achievable pairs of download cost and information leakage, we define two functions that describe the boundary of the download-leakage region.

Definition 3. *The download-leakage function* $\mathsf{D}^{(\cdot)}(\varrho)$ *for single-server WPIR is the minimum of all possible download costs* D *for a given information leakage constraint* ϱ *such that* (D, ϱ) *is achievable, i.e.,*

$$\mathsf{D}^{(\cdot)}(\varrho) \triangleq \min_{P_{\boldsymbol{Q}|M}\colon \rho^{(\cdot)}(P_{\boldsymbol{Q}|M}) \le \varrho} \mathbb{E}_{P_M P_{\boldsymbol{Q}|M}}[L(\boldsymbol{Q})].$$

Accordingly, the single-server WPIR capacity is $\mathsf{C}^{(\cdot)}(\varrho) = \left[\mathsf{D}^{(\cdot)}(\varrho)\right]^{-1}$.

It is known that the single-server PIR capacity is $\mathsf{C}_{\mathsf{M}} = \frac{1}{\mathsf{M}}$ [1].

Definition 4. *The leakage-download function* $\rho^{(\cdot)}(\mathsf{D})$ *for single-server WPIR is the minimum of all possible information leakages* ϱ *for a given download cost constraint* D *such that* (D, ϱ) *is achievable.*

Lemma 1. *The MI download-leakage function*

$$\mathsf{D}^{(\mathsf{MI})}(\varrho) = \min_{P_{\boldsymbol{Q}|M}\colon \mathsf{I}(P_{\boldsymbol{Q}|M}) \le \varrho} \mathbb{E}_{P_M P_{\boldsymbol{Q}|M}}[L(\boldsymbol{Q})]$$

is convex in ϱ*, while the MaxL download-leakage function*

$$\mathsf{D}^{(\mathsf{MaxL})}(\varrho) = \min_{P_{\boldsymbol{Q}|M}\colon \mathsf{MaxL}(P_{\boldsymbol{Q}|M}) \le \varrho} \mathbb{E}_{P_M P_{\boldsymbol{Q}|M}}[L(\boldsymbol{Q})]$$

is not a convex function, but $\mathsf{D}^{(\mathsf{MaxL})}(\log_2(\varrho))$ *is convex in* ϱ.

The convexity of $\mathsf{D}^{(\mathsf{MI})}$ can help to describe the download-leakage region if some achievable pairs are known. This observation can be summarized in the following corollary.

Corollary 1. *Assume that both pairs* $(\mathsf{D}_1, \varrho_1)$ *and* $(\mathsf{D}_2, \varrho_2)$ *are achievable. Then, for any* $\lambda \in [0,1]$*, the pair* $\big(\mathsf{D}_\lambda = (1-\lambda)\mathsf{D}_1 + \lambda \mathsf{D}_2, \varrho_\lambda = (1-\lambda)\varrho_1 + \lambda\varrho_2\big)$ *is achievable under MI leakage, while the pair* $\big(\mathsf{D}_\lambda = (1-\lambda)\mathsf{D}_1 + \lambda\mathsf{D}_2, \varrho_\lambda = \log_2\left[(1-\lambda)2^{\varrho_1} + \lambda 2^{\varrho_2}\right]\big)$ *is achievable for MaxL.*

B. Connection to Rate-Distortion Theory

Consider an information source sequence with independent and identically distributed components according to P_X and a *distortion measure* $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ between the source sequence $\boldsymbol{x}$ and the reconstructed sequence $\hat{\boldsymbol{x}}$. The optimal rate-distortion region is characterized by the *rate-distortion function*, defined as the minimum achievable compression rate $\mathsf{I}(X;\hat{X})$ under a given constraint on the average distortion $\mathbb{E}_{P_X P_{\hat{X}|X}}\big[d(X,\hat{X})\big]$, where $\hat{X}$ represents the reconstructed source.

To relate the leakage-download function to the rate-distortion function, the desired file index m needs to be added as an argument to the answer-length function L, without changing the download cost in (4). Then, the leakage and the download cost play similar roles as the compression rate and the average distortion, respectively. This can be achieved by letting $L(m, \boldsymbol{q}) \triangleq L(\boldsymbol{q})$ for all $m \in [\mathsf{M}]$ for which $P_{\boldsymbol{Q}|M}(\boldsymbol{q}|m) > 0$, and $L(m,\boldsymbol{q}) \triangleq \infty$ otherwise (i.e., an infinite length for a given m and query realization $\boldsymbol{q}$ indicates that $\boldsymbol{q}$ is never sent when requesting the m-th file). Below, we will equivalently use either $L(\boldsymbol{q})$ or $L(m,\boldsymbol{q})$ (as defined above).

IV. Partition WPIR Scheme

In [11], a WPIR scheme based on partitioning was proposed. The scheme is formally described as follows. The files are first partitioned into η equally-sized partitions, each consisting of M_η files, where $\mathsf{M}_\eta = \mathsf{M}/\eta \in \mathbb{N}$. Assume that the requested file $\boldsymbol{X}^{(m)}$ belongs to the j-th partition, where $j \in [\eta]$. Then, the query $\boldsymbol{Q}$ is constructed as

$$\boldsymbol{Q} = \big(\tilde{\boldsymbol{Q}}, j\big) \in \tilde{\mathcal{Q}} \times [\eta], \qquad (6)$$

where $\tilde{\boldsymbol{Q}}$ is the query of an existing M_η-file WPIR scheme.

The following theorem states the achievable download-leakage pairs of the partition scheme.

Theorem 1. *Consider a single server that stores* M *files and let* $\mathsf{M}_\eta = \mathsf{M}/\eta \in \mathbb{N}$*,* $\eta \in \mathbb{N}$*. Assume that an* M_η*-file WPIR scheme* $\tilde{\mathscr{C}}$ *with achievable download-leakage pair* $\big(\tilde{\mathsf{D}}, \tilde{\varrho}\big)$ *exists. Then, the download-leakage pair*

$$\big(\mathsf{D}(\mathscr{C}), \rho^{(\cdot)}(\mathscr{C})\big) = \big(\tilde{\mathsf{D}}, \tilde{\varrho} + \log_2 \eta\big) \qquad (7)$$

is achievable by the M*-file partition scheme* $\mathscr{C}$ *constructed from* $\tilde{\mathscr{C}}$ *as described in* (6)*.*

We refer to the partition scheme that uses a PIR scheme as the underlying subscheme and the query generation in (6) as a *basic scheme* and denote it by $\mathscr{C}^{\mathsf{basic}}$ (it achieves the pair in (7) with $\tilde{\varrho} = 0$). It can be seen that for the single-server setting, the basic scheme simply retrieves all the files in the partition that includes the requested file. This idea will be extended to our capacity-achieving scheme presented in Section VI, where for any subset $\mathcal{M} \subseteq [\mathsf{M}]$ that includes the requested file, all files in $\mathcal{M}$ are downloaded.

V. The Capacity of Single-Server WPIR

The main result of this work is the characterization of the optimal tradeoff between the download cost and the information leakage for single-server WPIR for an arbitrary number

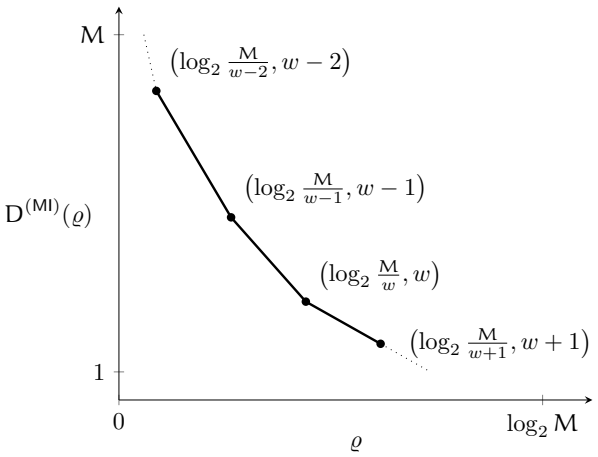

Fig. 1. An illustration of the function $\mathsf{D}^{(\mathsf{MI})}(\varrho)$, which is defined by many linear functions.

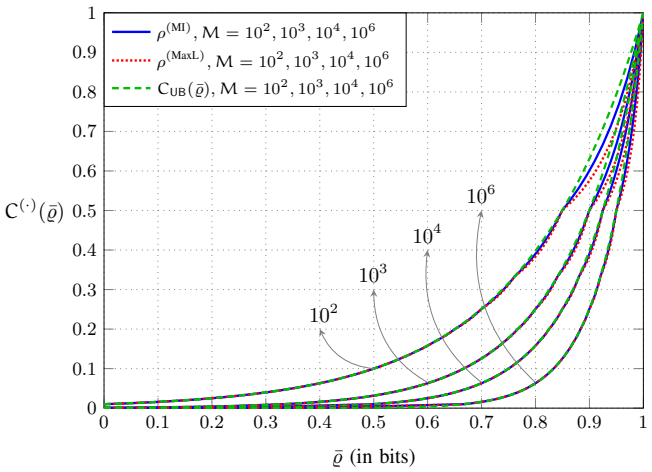

Fig. 2. The capacity $\mathsf{C}^{(\cdot)}(\bar{\varrho})$ and its upper bound $\mathsf{C}_{\mathsf{UB}}(\bar{\varrho})$ for a number of files $\mathsf{M} = 10^2, 10^3, 10^4$, and 10^6 with privacy metrics $\rho^{(\mathsf{MI})}$ and $\rho^{(\mathsf{MaxL})}$.

of files for the MI and MaxL privacy metrics. The optimal download-leakage region for the MI privacy metric is stated in the following theorem. For the sake of illustration, we consider the normalized leakage-download function $\bar{\rho}^{(\cdot)} \triangleq \frac{\rho^{(\cdot)}}{\log_2 \mathsf{M}}$.

Theorem 2. *For a single server that stores M files, the single-server WPIR capacity for the MI leakage metric $\rho^{(\mathsf{MI})}$ is*

$$\mathsf{C}^{(\mathsf{MI})}(\bar{\varrho}) = \left[w + \frac{\log_2 \frac{\mathsf{M}}{w}}{\log_2 \frac{w}{w-1}} - \frac{\bar{\varrho} \log_2 \mathsf{M}}{\log_2 \frac{w}{w-1}} \right]^{-1},$$

$$\text{for } 1 - \frac{\log_2 w}{\log_2 \mathsf{M}} \leq \bar{\varrho} \leq 1 - \frac{\log_2 (w-1)}{\log_2 \mathsf{M}},\ w \in [2:\mathsf{M}].$$

Theorem 3. *For a single server that stores M files, the single-server WPIR capacity for the MaxL metric $\rho^{(\mathsf{MaxL})}$ is*

$$\mathsf{C}^{(\mathsf{MaxL})}(\bar{\varrho}) = \left[(2w-1) - \frac{2^{\bar{\varrho} \log_2 \mathsf{M}}}{\frac{\mathsf{M}}{w-1} - \frac{\mathsf{M}}{w}} \right]^{-1},$$

$$\text{for } 1 - \frac{\log_2 w}{\log_2 \mathsf{M}} \leq \bar{\varrho} \leq 1 - \frac{\log_2 (w-1)}{\log_2 \mathsf{M}},\ w \in [2:\mathsf{M}].$$

It is worthwhile noting that the download-leakage function $\left[\mathsf{C}^{(\cdot)}(\cdot)\right]^{-1}$ is a piecewise continuous function. For the MI privacy metric, $\left[\mathsf{C}^{(\mathsf{MI})}(\varrho)\right]^{-1}$ is a piecewise linear function as demonstrated in Fig. 1 (without normalization). Note also that, when $\mathsf{M}_\eta = \mathsf{M}/\eta \in \mathbb{N}$, the basic scheme $\mathscr{C}^{\mathsf{basic}}$ achieves the capacity for both the MI and MaxL privacy metrics.

Next, we consider the asymptotic capacity of single-server WPIR, i.e., the capacity as the number of files M tends to infinity. An upper bound on the single-server WPIR capacity for any number of files is given in the following theorem.

Theorem 4. *For a single server that stores M files, the single-server WPIR capacity under both the MI metric $\rho^{(\mathsf{MI})}$ and the MaxL metric $\rho^{(\mathsf{MaxL})}$ is upperbounded as*

$$\mathsf{C}^{(\cdot)}(\bar{\varrho}) \leq \mathsf{C}_{\mathsf{UB}}(\bar{\varrho}) \triangleq \frac{1}{\mathsf{M}^{1-\bar{\varrho}}}, \quad 0 \leq \bar{\varrho} \leq 1.$$

In Fig. 2, the capacity $\mathsf{C}^{(\cdot)}(\bar{\varrho})$ and the upper bound $\mathsf{C}_{\mathsf{UB}}(\bar{\varrho})$ are plotted for $\mathsf{M} = 10^2, 10^3, 10^4$, and 10^6, which illustrates the asymptotic behavior of $\mathsf{C}^{(\cdot)}(\bar{\varrho})$ as M tends to infinity. Observe that from Theorems 2 and 3, we have $\mathsf{C}^{(\cdot)}(1) = 1$ for either a finite or infinite number of files M. Hence, by Theorem 4 it can be shown that as M tends to infinity, the asymptotic capacity is equal to

$$\mathsf{C}^{(\cdot)}_{\infty}(\bar{\varrho}) = \begin{cases} 0 & \text{if } 0 \leq \bar{\varrho} < 1, \\ 1 & \text{if } \bar{\varrho} = 1. \end{cases}$$

This indicates that the asymptotic capacity is still equal to zero, unless the server exactly knows the index of the requested file.

VI. Achievability of Theorems 2 and 3

Throughout this section, for simplicity, we set $\beta = 1$, i.e., $\mathsf{H}(\boldsymbol{X}^{(m)}) = 1, \forall\, m \in [\mathsf{M}]$. In fact, our proposed single-server WPIR capacity-achieving scheme works for an arbitrary file size β, which indicates that subpacketization does not improve the performance of single-server WPIR.

A. Motivating Example: $\mathsf{M} = 3$ Files

Before describing the achievable scheme in detail for the general case of M files, we present an example for $\mathsf{M} = 3$. We start by considering three achievable download-leakage pairs $(\mathsf{D}_1, \varrho_1) = (1, \log_2 3)$, $(\mathsf{D}_2, \varrho_2) = (2, \log_2 \frac{3}{2})$, and $(\mathsf{D}_3, \varrho_3) = (3, \log_2 \frac{3}{3} = 0)$. In terms of the MI or MaxL privacy metrics, one can check that the pair $(\mathsf{D}_w, \varrho_w)$ is achieved by the conditional distribution $P_{\boldsymbol{Q}_w|M}$, $w \in [3]$, listed in Table I. Moreover, it is clear that the retrievability condition in (1) is satisfied for any convex combination of the three PMFs.

Now, construct two conditional query distributions as follows,

$$P_{\boldsymbol{Q}_{\lambda_1}|M} = (1-\lambda_1) P_{\boldsymbol{Q}_2|M} + \lambda_1 P_{\boldsymbol{Q}_1|M}, \tag{8}$$

$$P_{\boldsymbol{Q}_{\lambda_2}|M} = (1-\lambda_2) P_{\boldsymbol{Q}_3|M} + \lambda_2 P_{\boldsymbol{Q}_2|M}, \tag{9}$$

TABLE I
THE CONDITIONAL PMFS $P_{\boldsymbol{Q}_w|M}$, $w \in [3]$.

$\mathcal{Q}_1$	$P_{\boldsymbol{Q}_1\|M}(\boldsymbol{q} \mid 1)$	$P_{\boldsymbol{Q}_1\|M}(\boldsymbol{q} \mid 2)$	$P_{\boldsymbol{Q}_1\|M}(\boldsymbol{q} \mid 3)$	$\boldsymbol{A}$	$P_{\boldsymbol{Q}_1}(\boldsymbol{q})$
$(1,0,0)$	1	0	0	$X_1^{(1)}$	$\frac{1}{3}$
$(0,1,0)$	0	1	0	$X_1^{(2)}$	$\frac{1}{3}$
$(0,0,1)$	0	0	1	$X_1^{(3)}$	$\frac{1}{3}$

$\mathcal{Q}_2$	$P_{\boldsymbol{Q}_2\|M}(\boldsymbol{q} \mid 1)$	$P_{\boldsymbol{Q}_2\|M}(\boldsymbol{q} \mid 2)$	$P_{\boldsymbol{Q}_2\|M}(\boldsymbol{q} \mid 3)$	$\boldsymbol{A}$	$P_{\boldsymbol{Q}_2}(\boldsymbol{q})$
$(1,1,0)$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\{X_1^{(1)}, X_1^{(2)}\}$	$\frac{1}{3}$
$(1,0,1)$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\{X_1^{(1)}, X_1^{(3)}\}$	$\frac{1}{3}$
$(0,1,1)$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\{X_1^{(2)}, X_1^{(3)}\}$	$\frac{1}{3}$

$\mathcal{Q}_3$	$P_{\boldsymbol{Q}_3\|M}(\boldsymbol{q} \mid 1)$	$P_{\boldsymbol{Q}_3\|M}(\boldsymbol{q} \mid 2)$	$P_{\boldsymbol{Q}_3\|M}(\boldsymbol{q} \mid 3)$	$\boldsymbol{A}$	$P_{\boldsymbol{Q}_3}(\boldsymbol{q})$
$(1,1,1)$	1	1	1	$\{X_1^{(1)}, X_1^{(2)}, X_1^{(3)}\}$	1

where $0 \le \lambda_1, \lambda_2 \le 1$.

Using (4), (2), and (3), respectively, and the conditional PMFs listed in Table I (or, alternatively, Corollary 1), it follows that the WPIR scheme defined by (8)–(9) achieves the download cost

$$\mathsf{D}(\mathscr{C}) = \begin{cases} (1-\lambda_1)\mathsf{D}_2 + \lambda_1 \mathsf{D}_1 = 2 - \lambda_1, & 0 \le \lambda_1 \le 1, \\ (1-\lambda_2)\mathsf{D}_3 + \lambda_2 \mathsf{D}_2 = 3 - \lambda_2, & 0 \le \lambda_2 \le 1, \end{cases}$$

the MI leakage

$$\rho^{(\mathsf{MI})} = \begin{cases} (1-\lambda_1)\log_2 \frac{3}{2} + \lambda_1 \log_2 \frac{3}{1}, & 0 \le \lambda_1 \le 1, \\ (1-\lambda_2)\log_2 \frac{3}{3} + \lambda_2 \log_2 \frac{3}{2}, & 0 \le \lambda_2 \le 1, \end{cases}$$

and the MaxL

$$\rho^{(\mathsf{MaxL})} = \begin{cases} \log_2\big((1-\lambda_1)\frac{3}{2} + \lambda_1 \frac{3}{1}\big), & 0 \le \lambda_1 \le 1, \\ \log_2\big((1-\lambda_2)\frac{3}{3} + \lambda_2 \frac{3}{2}\big), & 0 \le \lambda_2 \le 1. \end{cases}$$

In terms of the MI or MaxL privacy metrics, it can be verified that the download cost corresponds to the single-server WPIR capacity for $\mathsf{M} = 3$.

B. Arbitrary Number of Files M

We describe the achievable scheme for the general case of M files. From Corollary 1, it follows that it is sufficient to show that the download-leakage pairs

$$(\mathsf{D}_w, \varrho_w) = \left(w, \log_2 \frac{\mathsf{M}}{w}\right), \quad w \in [\mathsf{M}],$$

are achievable.

1) Query Generation: Consider M random queries $\boldsymbol{Q}_w$, $w \in [\mathsf{M}]$, whose alphabet is $\mathcal{Q}_w \triangleq \{\boldsymbol{q} = (q_1, \ldots, q_{\mathsf{M}}) \in \{0,1\}^{\mathsf{M}} : w_{\mathsf{H}}(\boldsymbol{q}) = w\}$. Each query $\boldsymbol{q} \in \mathcal{Q}_w$ sent to the server is generated by the conditional PMF

$$P_{\boldsymbol{Q}_w|M}(\boldsymbol{q}|m) = \begin{cases} \frac{1}{\binom{\mathsf{M}-1}{w-1}} & \text{if } m \in \chi(\boldsymbol{q}), \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid query design, since for each $m \in [\mathsf{M}]$, we have $\sum_{\boldsymbol{q}\in\mathcal{Q}_w} P_{\boldsymbol{Q}_w|M}(\boldsymbol{q}|m) = \sum_{\boldsymbol{q}\in\mathcal{Q}_w :\, m\in\chi(\boldsymbol{q})} P_{\boldsymbol{Q}_w|M}(\boldsymbol{q}|m) = 1$.

2) Answer Construction: The answer function φ maps the query $\boldsymbol{q} \in \mathcal{Q}_w$ onto $\boldsymbol{A} = \varphi(\boldsymbol{q}, \boldsymbol{X}^{[\mathsf{M}]}) = X_1^{\chi(\boldsymbol{q})}$. The answer length is $L(\boldsymbol{q}) = w$.

3) Download Cost and Information Leakage: Clearly, the download cost is equal to $\mathbb{E}_{P_M P_{\boldsymbol{Q}_w|M}}[L(\boldsymbol{Q}_w)] = w$. The MI leakage is

$$\rho^{(\mathsf{MI})}(P_{\boldsymbol{Q}_w|M}) = \mathsf{I}(M;\boldsymbol{Q}_w) = \log_2 \mathsf{M} - \log_2 w = \log_2 \frac{\mathsf{M}}{w}$$

and the MaxL is

$$\rho^{(\mathsf{MaxL})}(P_{\boldsymbol{Q}_w|M}) = \log_2 \sum_{\boldsymbol{q}\in\mathcal{Q}_w} \max_{m\in[\mathsf{M}]} \frac{1}{\binom{\mathsf{M}-1}{w-1}} = \log_2 \frac{\mathsf{M}}{w}.$$

Notice that the presented capacity-achieving WPIR scheme can be seen as a generalization of the basic WPIR scheme $\mathscr{C}^{\mathsf{basic}}$. If $\mathsf{M}/\eta = w \in \mathbb{N}$, it can be shown that the download-leakage pair $(w, \log_2 \eta) = (w, \log_2 \frac{\mathsf{M}}{w})$ is also achievable by $\mathscr{C}^{\mathsf{basic}}$ for both the MI and MaxL privacy metrics.

VII. CONVERSE OF THEOREMS 2 AND 3

The converse proofs for the MI and MaxL privacy metrics are fully elaborated in the extended version [16, Secs. VII and VIII]. Here, we briefly outline the main steps of the converse proof for MI leakage.

Consider a single-server WPIR scheme where the leakage at the server is measured by $\mathsf{I}(M;\boldsymbol{Q})$. A general converse for Theorem 2 can be derived from the download-leakage function of a given leakage constraint ϱ, or equivalently, from the leakage-download function of a given download cost constraint D. Similar to (5), the leakage-download function can be formulated by the convex minimization problem

$$\min_{P_{\boldsymbol{Q}|M}:\ \mathbb{E}[L(M,\boldsymbol{Q})]\le\mathsf{D}} \mathsf{I}(M;\boldsymbol{Q}). \tag{10}$$

The proof consists of two parts as described below.

Part 1: With some technical efforts, we can prove that (10) is bounded from below by

$$\rho^{(\mathsf{MI})}(\mathsf{D}) \triangleq \min_{P_{\boldsymbol{U}|M}:\ \mathbb{E}[L(M,\boldsymbol{U})]\le\mathsf{D}} \mathsf{I}(M;\boldsymbol{U}), \tag{11}$$

where $\boldsymbol{U}$ is a length-M binary random vector and the length function $L(m, \boldsymbol{u})$ is constructed for any $\boldsymbol{u} \in \{0,1\}^{\mathsf{M}}$ by

$$L(m, \boldsymbol{u}) = \begin{cases} w_{\mathsf{H}}(\boldsymbol{u}) & \text{if } m \in \chi(\boldsymbol{u}), \\ \infty & \text{otherwise.} \end{cases}$$

Part 2: Since (11) can be related to the rate-distortion function with a certain distortion measure, we use a useful result from the rate-distortion theory (see [16, Lem. 3]) to prove a lower bound on (11). It can be shown that the pair $(\bar{\varrho}, \mathsf{D}^{(\mathsf{MI})}(\bar{\varrho}))$ of Theorem 2 lies on the lower bound of (11).

VIII. CONCLUSION

We characterized the capacity of single-server WPIR with an arbitrary number of files for the MI and MaxL privacy metrics. We showed that, interestingly, the optimal download rate is higher than the single-server PIR capacity. Moreover, we showed that if the server can not determine the identity of the requested file, then the capacity tends to zero as the number of files goes to infinity.

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] 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.

[3] 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.

[4] 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.

[5] H.-Y. Lin, S. Kumar, E. Rosnes, and A. Graell i Amat, "Asymmetry helps: Improved private information retrieval protocols for distributed storage," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 25–29, 2018, pp. 1–5.

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

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

[8] 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.

[9] L. Holzbaur, R. Freij-Hollanti, J. Li, and C. Hollanti, "Towards the capacity of private information retrieval from coded and colluding servers," Mar. 2019, arXiv:1903.12552v5 [cs.IT].

[10] R. R. Toledo, G. Danezis, and I. Goldberg, "Lower-cost ϵ-private information retrieval," in *Proc. Privacy Enhancing Technol. Symp. (PETS)*, Darmstadt, Germany, Jul. 19–22, 2016, pp. 184–201.

[11] 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.

[12] 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.

[13] 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.

[14] G. Barthe and B. Köpf, "Information-theoretic bounds for differentially private mechanisms," in *Proc. 24th Comput. Secur. Found. Symp. (CSF)*, Cernay-la-Ville, France, Jun. 27–29, 2011, pp. 191–204.

[15] 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.

[16] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "The capacity of single-server weakly-private information retrieval," Jan. 2020, arXiv:2001.08727v1 [cs.IT].