Improved Weakly Private Information Retrieval Codes

Chengyuan Qian, Ruida Zhou, Chao Tian, and Tie Liu
Department of Electrical and Computer Engineering, Texas A&M University
{cyqian, ruida, chao.tian,tieliu}@tamu.edu

***Abstract*—We study the problem of weakly private information retrieval (W-PIR), where a user wishes to retrieve a desired message from N non-colluding servers in a way that the privacy leakage regarding the desired message's identity is less than or equal to a threshold. We propose a new code construction which significantly improves upon the best known result in the literature, based on the following critical observation. In previous constructions, for the extreme case of minimum download, the retrieval pattern is to download the message directly from $N-1$ servers; however this causes leakage to all these $N-1$ servers, and a better retrieval pattern for this extreme case is to download the message directly from a single server. The proposed code construction allows a natural transition to such a pattern, and for both the maximal leakage metric and the mutual information leakage metric, significant improvements can be obtained. We provide explicit solutions, in contrast to a previous work by Lin et al., where only numerical solutions were obtained.**

I. Introduction

The study of *private information retrieval* (PIR) systems [1] was motivated by the practical need of protecting privacy during information retrieval. In the canonical PIR setting, a user wishes to retrieve a message from N servers, each keeping a copy of all K messages. The servers are non-colluding, i.e., they cannot communicate with each other. The user wishes to ensure that the servers can infer no information about the identity of the desired message. Since the message is usually quite large, the dominant communication cost is the download from the servers. The highest possible information bits per downloaded bit is referred to as the PIR capacity, which was recently fully characterized by Sun and Jafar [2]. An alternative optimal code (referred to as the TCS code) was later proposed [3], which uses the minimum possible message length and query set. Many variations of the canonical PIR problem has been studied, such as colluding servers [4]–[6], storage constrained [7]–[17], with symmetric privacy requirement [18]–[20], and with side information [21]–[27].

The perfect privacy requirement in the canonical setting can be unnecessarily stringent. A small amount of privacy leakage is likely acceptable in many practical scenarios, e.g., when the user does not mind if the server can infer the identity of the desired message with only relatively low confidence. This setting, where a weaker privacy constraint is placed, is referred to as weakly private information retrieval (W-PIR) [28]–[36]. In exchange for the loss of privacy, a higher retrieval rate can be attained, sometimes with a lower computational complexity [28]. Several different metrics have been proposed to measure the privacy leakage in W-PIR. Differential privacy was used in [29], [30], conditional entropy was used in [31], mutual information in [32], and the maximal leakage metric (see [37]) was adopted in [33], [34]. The W-PIR code proposed in [33] for the maximal leakage metric was obtained by adjusting the code proposed in [3]; similar or identical code constructions were also analyzed in [34] and [35] under different metrics, either theoretically or numerically.

The previously best known W-PIR code under the maximal leakage constraint [33] was obtained by breaking the uniform distribution on the retrieval patterns in the TSC code, which increasingly favors the direct download pattern in the code as the privacy requirement is relaxed; it was shown to be optimal when $N = 2$ in [34]. In this work, we provide a new code construction by making the following critical observation. The direct download pattern in the TSC code essentially downloads the desired message from $N-1$ servers, one symbol from each server. However, this would result in privacy leakage to all these $N-1$ servers, when this pattern is not mixed with other patterns of retrieval. In the extreme case of minimum download, the W-PIR code in [33] can only use this pattern, yet an alternative strategy is to directly download the full message from *one single server*, which only leaks to this single server. Our proposed new code utilizes this observation and allows a natural transition to this retrieval pattern.

The new code can also be viewed as adjusting the probabilities of the retrieval patterns in the TSC code, jointly with the new clean download retrieval pattern. We provide the optimal distributions explicitly under both the maximal leakage metric and the mutual information leakage metric, and show that the new code can achieve significant improvement over existing ones. It should be noted that in [34], the simpler code without the new retrieval pattern was studied only numerically, and no explicit solution was provided.

II. Preliminaries

In this section, we formally introduce the W-PIR problem under the maximum leakage metric and mutual information metric, respectively, and then review the PIR code proposed in [3] that will be instrumental later on.

A. Information Retrieval Systems

There are a total of N servers, and each server stores an independent copy of K mutually independent messages, denoted as $W_{1:K} := (W_1, W_2, \ldots, W_K)$, where $K \geq 2$ without loss of generality. Each message consists of L symbols, and

This work was supported in part by the National Science Foundation via Grants CCF-20-07067.

each symbol is distributed uniformly in a finite set $\mathcal{X}$, which implies that

$$L := H(W_1) = H(W_2) = \cdots = H(W_K),$$

where the entropy is taken under the logarithm of base $|\mathcal{X}|$. The i-th symbol of the message W_k is denoted as $W_k[i]$, where $i = 1, \ldots, L$. An information retrieval code consists of the following component functions. When a user wishes to retrieve a message W_k, $k \in [1:K]$, the (random) query $Q_n^{[k]}$ sent to server-n is generated according to an encoding function

$$Q_n^{[k]} := \phi_n(k, F), \quad \forall n \in 1:N, \tag{1}$$

by leveraging some private random key $F \in \mathcal{F}$. Let $\mathcal{Q}_n$ be the union of all possible queries $Q_n^{[k]}$ over all k. For each $n \in 1:N$, upon receiving a query $q \in \mathcal{Q}_n$, server-n responds with an answer $A_n^{(q)}$ produced as

$$A_n^{(q)} := \varphi_n(q, W_{1:K}), \tag{2}$$

which is represented by $\ell_n^{(q)}$ symbols in certain coding alphabet $\mathcal{Y}$; to simplify the notation, we assume $\mathcal{X} = \mathcal{Y}$ in this work. We assume $\ell_n^{(q)}$ may vary according to the query but not the messages, and as such the user knows how many symbols are expected in that answer.

For notational simplicity, we denote $A_n^{(Q^{[k]})}$ as $A_n^{[k]}$ and $\ell_n^{(Q^{[k]})}$ as $\ell_n^{[k]}$, both of which are random variables. With the answers from the servers, the user attempts to recover the message $\hat{W}_k$ using the decoding function

$$\hat{W}_k := \psi(A_{1:N}^{[k]}, k, F). \tag{3}$$

A valid information retrieval code must first satisfy $\hat{W}_k = W_k$, i.e., the desired message should be correctly recovered.

We measure the download cost by the normalized (worst-case) *average download cost*,

$$D := \max_{k \in 1:K} \mathbb{E}\left[\frac{1}{L}\sum_{n=1}^{N} \ell_n^{[k]}\right], \tag{4}$$

where $\ell_n^{[k]}$ is the length of the answer in the code and the expectation is taken with respect to the random key F.

B. Maximal Leakage and Mutual Information Leakage

The index of the desired message, denoted as M, is viewed as a random variable following a certain distribution. The identity of the desired message W_M may be leaked to server-n due to the query $Q_n^{[M]}$ sent by the user. We focus on two metrics to study this leakage.

The maximal leakage metric $\mathcal{L}(M \to Q_n^{[M]})$: It was shown in [37] and [33] that

$$\mathcal{L}(M \to Q_n^{[M]}) = \log_2\left(\sum_{q \in \mathcal{Q}_n} \max_{k \in 1:K} \mathbb{P}(Q_n^{[k]} = q)\right), \tag{5}$$

which in fact does not depend on the probability distribution of M. When $\mathcal{L}(M \to Q_n^{[M]})$ is large, $Q_n^{[M]}$ leaks more information of M in the sense that server-n can estimate M more accurately; on the other hand, when $\mathcal{L}(M \to Q_n^{[M]}) = 0$, the retrieval is private in the sense that the distribution of $Q_n^{[k]}$ and $Q_n^{[k']}$ are identical for any $k, k' \in [1:K]$.

A *valid* code for W-PIR with K messages and N servers under a maximum leakage constraint ρ is a collection of functions $(\{\phi_n\}_{n\in[1:N]}, \{\varphi_n\}_{n\in[1:N]}, \psi)$ that can correctly retrieve the desired message, and additionally satisfies the privacy constraints that for each server-n,

$$\mathcal{L}(M \to Q_n^{[M]}) \le \rho, \quad n \in [1:N]. \tag{6}$$

A download cost D is called achievable for ρ, if there exists a valid code such that its download cost is less than or equal to D for privacy constraint ρ. The closure of the collection of such (ρ, D) pairs is called the achievable (ρ, D) region under the maximal leakage constraint, denoted by $\mathcal{G}_{\text{MaxL}}$; the infimum of such achievable download cost for ρ is the download-leakage function, denoted as $D_{\text{MaxL}}(\rho)$.

The mutual information metric: In this setting, the identity of the request of message M is assumed to be uniformly distributed in the set $[1:K]$. Then the mutual information leakage is

$$\text{MI}(M \to Q_n^{[M]}) := I(M; Q_n), \tag{7}$$

where Q_n is the random query induced jointly by the random key F and the random message index M. We can similarly define valid codes under the mutual information leakage constraint under the condition,

$$\text{MI}(M \to Q_n^{[M]}) \le \rho, \quad n \in [1:N]. \tag{8}$$

Similarly the achievable (ρ, D) region under this metric is denoted as $\mathcal{G}_{\text{MI}}$ and the download-leakage function as $D_{\text{MI}}(\rho)$.

C. The TSC Code

The TSC code given in [3] will serve an instrumental role in this work. In this code, the message length $L = N - 1$. A dummy symbol $W_k[0] = 0$ is prepended at the beginning of all messages. In order to better facilitate the new code construction, we give a variation of the original construction, which can be viewed as probabilistic sharing among the cyclic permutations (over the N servers) of the PIR code in [3].

Let the random key F^* be a length-K vector

$$F^* := (F_1^*, F_2^*, \ldots, F_{K-1}^*, U), \tag{9}$$

where $F_1^*, \ldots, F_{K-1}^*, U$ are independent random variables uniformly distributed over the set $[0:N-1]$, i.e.,

$$F^* \in [0:N-1]^K \triangleq \mathcal{F}^*. \tag{10}$$

The query $Q_n^{[k]}$ to server-n is generated by the function $\phi_n^*(k, F^*)$ defined as,

$$\begin{aligned}\phi_n^*(k, F^*) \triangleq (&F_1^*, F_2^*, \ldots, F_{k-1}^*, (U+n)_N,\\ &F_k^*, F_{k+1}^*, \ldots, F_{K-1}^*),\end{aligned} \tag{11}$$

where $(\cdot)_N$ represents the modulo N operation. Note that in [3], the indices n and k start from 0, and the random variable U here is instead $-\sum_{j=1}^{K-1} F_j$. The new random key component U is introduced here, such that the query to the server n is cyclically permuted uniformly at random. As a result, for any server-n, the query $Q_n^{[k]}$ is uniformly distributed on the set $\mathcal{Q}^* \triangleq [0:N-1]^K$, i.e.,

$$\mathbb{P}(Q_n^{[k]} = q) = N^{-K}, \quad \forall n \in [1:N], q \in \mathcal{Q}^*. \tag{12}$$

Upon receiving this query, the server-n returns the answer $A_n^{[k]}$ generated by the function $\varphi^*(q, W_{1:K})$,

$$\begin{aligned}\varphi^*(q, W_{1:K}) &\triangleq W_1[Q_{n,1}^{[k]}] \oplus W_2[Q_{n,2}^{[k]}] \oplus \cdots \oplus W_K[Q_{n,K}^{[k]}] \\ &= W_k[(U+n)_N] \oplus \mathscr{I}, \end{aligned} \tag{13}$$

where $\oplus$ denotes addition in the given finite field, $Q_{n,m}^{[k]}$ represents the m-th symbol of $Q_n^{[k]}$, and $\mathscr{I}$ is the interference signal defined as

$$\begin{aligned}\mathscr{I} = W_1[F_1] \oplus \cdots \oplus W_{k-1}[F_{k-1}] \\ \oplus W_{k+1}[F_k] \oplus \cdots \oplus W_K[F_{K-1}]. \end{aligned} \tag{14}$$

Since there exists an $n \in [1:N]$, $(U+n)_N = 0$ such that $A_n^{[k]} = \mathscr{I}$, the user can retrieve the desired message W_k by subtracting $\mathscr{I}$ from $A_{n'}^{[k]}$ for all $n' \neq n$. Note that with probability $N^{-(K-1)}$ the interference signal $\mathscr{I}$ consists of only dummy symbols and need not to be downloaded, in which case a direct download will be performed by retrieving the desired message from $N-1$ servers, one symbol per server. The download cost is therefore

$$D^* = \frac{N}{N-1}\left(1 - \frac{1}{N^{K-1}}\right) + \frac{1}{N^{K-1}} = \frac{1-N^{-K}}{1-N^{-1}}, \tag{15}$$

matching the capacity result given in [2]. An example of the code (with adjusted probabilities for W-PIR) is given in Section IV; more details can be found in [3].

III. Main Results

We summarize the performance of the new code in the following two theorems.

Theorem 1. *For W-PIR under the maximal leakage constraint,*

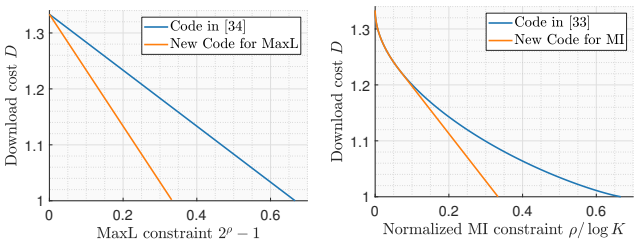
$$\begin{aligned}&D_{MaxL}(\rho) \\ &\leq 1 + \left(1 - N\frac{2^\rho - 1}{K-1}\right)_+ \left(\frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right), \end{aligned} \tag{16}$$

where $(x)_+ := \max(x, 0)$.

The code construction for Theorem 1 is given in the next section, which is obtained by probabilistic sharing of the TSC code with the new retrieval pattern. This result is presented in terms of the download-leakage function for the maximal leakage setting. For the mutual information metric setting, a few additional quantities are required to parametrize the solution. First define a sequence $\boldsymbol{x} = (x_1, x_, \ldots, x_{K-1})$, which can be shown to be greater than or equal to 1 component-wise, using the following recursion backwards from $K-1, K-2, \ldots, 1$:

$$\begin{aligned}\log\frac{(K-i)x_{K-i}+i}{K} &= \sum_{j=0}^{i-1}(1-N)^j \log\frac{(K-1)x_{K-1}+1}{K} \\ &\quad - \sum_{j=1}^{i-1}(1-N)^j \log x_{K-i+j}. \end{aligned} \tag{17}$$

Since the RHS only depends on $x_{K-1}, x_{K-2}, \ldots, x_{K-i+1}$, the sequence is well defined, when $x_{K-1} \in [1, \infty]$ is specified. With this sequence defined, we further define the following



(a) Maximal leakage metric. (b) Mutual information metric.

Fig. 1: Numerical comparisons between the proposed code and previously best known code under different privacy metrics when $N = 3, K = 2$.

probability vector $\boldsymbol{p} = (p_0, p_1, \ldots, p_{K-1})$:

$$p_0(\boldsymbol{x}) = \left(N + N\sum_{w=1}^{K-1}\binom{K-1}{w}(N-1)^w \prod_{j=1}^{w}\frac{1}{x_j}\right)^{-1}, \tag{18}$$

$$p_w(\boldsymbol{x}) = p_0(\boldsymbol{x})\prod_{j=1}^{w}\frac{1}{x_j}, \quad w \in 1:K-1. \tag{19}$$

It can be verified that $\boldsymbol{p}$ induces a probability distribution with the appropriate combinatorial coefficients taken into account.

Define the region $\hat{\mathcal{G}}_{MI}$ to be the nonnegative (ρ, D) pairs satisfying the following conditions

$$\begin{aligned}\rho \geq \frac{1}{K}\sum_{w=0}^{K}\binom{K}{w}(N-1)^w \\ \Big\{wp_{w-1}\log p_{w-1} + (K-w)p_w \log p_w \\ - [wp_{w-1} + (K-w)p_w]\log\frac{wp_{w-1}+(K-w)p_w}{K}\Big\} \end{aligned} \tag{20}$$

$$D \geq \frac{N}{N-1}(1-p_0), \tag{21}$$

for some $x_{K-1} \in [1, \infty]$; for simplicity, p_{-1} and p_K are defined as zero. We then have the following theorem.

Theorem 2. *For the mutual information leakage metric,* $conv(\hat{\mathcal{G}}_{MI} \cup \{(\frac{\log K}{N}, 1)\}) \subseteq \mathcal{G}_{MI}$, *where* $conv(\cdot)$ *is the convex hull operation.*

The performance of the proposed code is illustrated in Fig. 1. In both cases, the new extreme point of minimum download provides a new anchor point for the (ρ, D) tradeoff. In essence, the new achievable regions can be obtained by proper probabilistic sharing of the existing code with the clean retrieval pattern for this new extreme point (see details in the next section), in the correct coordinate for the two metrics. The sharing structure is more sophisticated for the mutual information leakage case, and it can be seen that when ρ is below a threshold, the new clean download pattern is in fact not effective, meaning it is not utilized during retrieval.

IV. A New Code Construction

We first give an example to illustrate the proposed code based on probabilistic sharing, then present the general construction.

TABLE I: Proposed code for $N = 3, K = 2$

(a) Retrieval of W_1

$\mathbb{P}_F(F)$	F	**Server** 1		**Server** 2		**Server** 3	
		$Q_1^{[1]}$	A_1	$Q_2^{[1]}$	A_2	$Q_3^{[1]}$	A_3
p_0'	1	$\#_1$	a_1, a_2	$\underline{00}$	$\varnothing$	$\underline{00}$	$\varnothing$
p_0'	2	$\underline{00}$	$\varnothing$	$\#_1$	a_1, a_2	$\underline{00}$	$\varnothing$
p_0'	3	$\underline{00}$	$\varnothing$	$\underline{00}$	$\varnothing$	$\#_1$	a_1, a_2
p_0	00	10	a_1	20	a_2	$\underline{00}$	$\varnothing$
p_0	10	20	a_2	$\underline{00}$	$\varnothing$	10	a_1
p_0	20	$\underline{00}$	$\varnothing$	10	a_1	20	a_2
p_1	01	11	$a_1 \oplus b_1$	21	$a_2 \oplus b_1$	01	b_1
p_1	11	21	$a_2 \oplus b_1$	01	b_1	11	$a_1 \oplus b_1$
p_1	21	01	b_1	11	$a_1 \oplus b_1$	21	$a_2 \oplus b_1$
p_1	02	12	$a_1 \oplus b_2$	22	$a_2 \oplus b_2$	02	b_2
p_1	12	22	$a_2 \oplus b_2$	02	b_2	12	$a_1 \oplus b_2$
p_1	22	02	b_2	12	$a_1 \oplus b_2$	22	$a_2 \oplus b_2$

(b) Retrieval of W_2

$\mathbb{P}_F(F)$	F	**Server** 1		**Server** 2		**Server** 3	
		$Q_1^{[2]}$	A_1	$Q_2^{[2]}$	A_2	$Q_3^{[2]}$	A_3
p_0'	1	$\#_2$	b_1, b_2	$\underline{00}$	$\varnothing$	$\underline{00}$	$\varnothing$
p_0'	2	$\underline{00}$	$\varnothing$	$\#_2$	b_1, b_2	$\underline{00}$	$\varnothing$
p_0'	3	$\underline{00}$	$\varnothing$	$\underline{00}$	$\varnothing$	$\#_2$	b_1, b_2
p_0	00	01	b_1	02	b_2	$\underline{00}$	$\varnothing$
p_0	01	02	b_2	$\underline{00}$	$\varnothing$	01	b_1
p_0	02	$\underline{00}$	$\varnothing$	01	b_1	02	b_2
p_1	10	11	$a_1 \oplus b_1$	12	$a_1 \oplus b_2$	10	a_1
p_1	11	12	$a_1 \oplus b_2$	10	a_1	11	$a_1 \oplus b_1$
p_1	12	10	a_1	11	$a_1 \oplus b_1$	12	$a_1 \oplus b_2$
p_1	20	21	$a_2 \oplus b_1$	22	$a_2 \oplus b_2$	20	a_2
p_1	21	22	$a_2 \oplus b_2$	20	a_2	21	$a_2 \oplus b_1$
p_1	22	20	a_2	21	$a_2 \oplus b_1$	22	$a_2 \oplus b_2$

A. An Illustrative Example

Consider the case with $K = 2$ messages and $N = 3$ servers. The message length is $L = N - 1 = 2$, and we write $W_1 = (a_1, a_2)$, $W_2 = (b_1, b_2)$. The queries and answers are given in Table I. The queries in the top three rows of the two tables directly request the full message from a single server denoted by $\#_1$ and $\#_2$, while the remaining nine rows are essentially the TSC code with different probabilities for the queries, assigned according to their interference signals. Note that the interference signal is controlled by the first-$(K-1)$ entries $F^*_{1:K-1}$ of the random key F^*. For any $F \in \mathcal{F}^*$, denote $|F|$ as the *size* of the interference corresponding to random key F, which is also the hamming weight of $F^*_{1:K-1}$. In this example $|F|$ can only be 0 or 1.

We have omitted the dummy symbols a_0 and b_0 for conciseness. The random key F has a total of 12 possible realizations, with the probability parametrized by (p_0', p_0, p_1), where p_0' is the probability of direct download from a single given server, p_0 is that of the interference having hamming weight 0, p_1 that of the interference having haming weight 1.

B. General Code Construction

For general W-PIR with parameter (N, K), we set $L = N - 1$. The random key F is generated from set $\mathcal{F}$ with a probability distribution $\mathbb{P}_F(F)$, where $\mathcal{F} = \mathcal{F}^* \cup [1:N] = [0:N-1]^K \cup [1:N]$, and

$$\mathbb{P}_F(F) = \begin{cases} p_0', & \forall F \in [1:N] \\ p_w, & \forall F \in \mathcal{F}^*, |F| = w,\ w \in [0:K-1] \end{cases}, \tag{22}$$

which needs to satisfy

$$Np_0' + N\sum_{w=0}^{K-1} \binom{K-1}{w} (N-1)^w p_w = 1. \tag{23}$$

The query $Q_n^{[k]}$ to server-n is produced as:

$$Q_n^{[k]} = \begin{cases} \#_k, & F = n \\ \underline{0_K}, & \forall F \in [1:N],\ F \neq n \\ \phi_n^*(k, F), & \forall F \in \mathcal{F}^* \end{cases}, \tag{24}$$

where $\underline{0_K}$ is the length-K all-zero vector. The answer $A_n^{[k]}$ from server-n is generated as

$$A_n^{[k]} = \begin{cases} W_k, & q = \#_k \\ \varphi^*(q, W_{1:K}), & \forall q \in \mathcal{Q}^* \end{cases}. \tag{25}$$

The correctness of the code is obvious, and the download cost D can be simply computed as

$$p_d \triangleq N(p_0' + p_0) \tag{26}$$

$$D = p_d + \frac{N}{N-1}(1 - p_d), \tag{27}$$

where p_d is the overall probability of using a direct download. We defer the analysis of the privacy for the two metrics to the next subsection.

V. Code Optimization and Performance Analysis

We have provided the new code construction in a general form in the previous section, however, without optimizing the probability distribution. In this section, we optimize the probability distributions for the two leakage constraints, respectively.

A. Optimizing for Maximal Leakage

Since p_d is directly related to D in the proposed code, setting p_d is equivalent to specifying a target download cost D in this code. Therefore, the constrained minimization problem can be written as follows:

$$\text{Minimize:} \quad \mathcal{L}(M \to Q_n^{[M]}), \tag{28}$$

$$\text{Variables:} \quad \boldsymbol{p} = (p_0, p_1, \ldots, p_{K-1}), \tag{29}$$

$$\text{Subject to:} \quad -p_w \leq 0, \forall w \in 0:K-1, \tag{30}$$

$$Np_0 - p_d \leq 0, \tag{31}$$

$$N\sum_{w=0}^{K-1} \binom{K-1}{w} (N-1)^w p_w + (p_d - Np_0) - 1 = 0. \tag{32}$$

This optimization problem can indeed be solved (see a proof of a similar nature in [33]), for which the solution is

$$p_0' = \min\left(\frac{1}{N}, \frac{2^\rho - 1}{K-1}\right), \tag{33}$$

$$p_w = \frac{1 - Np_0'}{N^K}, \quad w \in 0 : K-1, \tag{34}$$

and the leakage can thus be found as

$$\begin{aligned}
\mathcal{L}_n(M \to Q_n^{[M]}) &= \log_2 \sum_{q_n \in \mathcal{Q}_n} \max_{j\in[1:K]} \mathbb{P}_{Q_n^{[j]}}(q_n) \\
&= \log_2 \left[\left(\sum_{q_n = \underline{0}_K} + \sum_{\substack{q_n \in \mathcal{Q}^*, \\ q_n \neq \underline{0}_K}} + \sum_{\substack{\#_k: \\ k\in[1:K]}} \right) \max_{j\in[1:K]} \mathbb{P}_{Q_n^{[j]}}(q_n) \right] \\
&= \log_2 \Bigg\{ \left[(N-1)p_0' + \frac{1-Np_0'}{N^K} \right] \\
&\qquad + (N^K - 1)\frac{1-Np_0'}{N^K} + Kp_0' \Bigg\} \\
&= \log_2[1 + (K-1)p_0'] \\
&= \min\left(\rho, \log_2[1 + (K-1)/N]\right), \qquad (35)
\end{aligned}$$

which is exactly that given in Theorem 1.

B. Optimizing for the Mutual Information Leakage

The optimization problem is very similar to that in the maximal leakage case except that the objective function is the mutual information leakage. We first analytically solve for the optimal probability distribution with the new retrieval pattern excluded, i.e. $p_d = Np_0$, and then prove Theorem 2 using the fact that the $D_{\text{MI}}(\rho)$ is convex [34]. The optimization problem is thus formulated as

$$\begin{aligned}
\text{Minimize:} \quad & I(Q_n; M), && (36)\\
\text{Variables:} \quad & (p_1, p_2, \ldots, p_{K-1}), && (37)\\
\text{Subject to:} \quad & -p_w \le 0, \forall w \in 1 : K-1, && (38)\\
& N\sum_{w=0}^{K-1} \binom{K-1}{w} (N-1)^w p_w - 1 = 0. && (39)
\end{aligned}$$

The download cost D here is directly related to $p_d = Np_0$, and setting a positive p_0 value is equivalent to specifying D.

The following proposition shows that the vector $\boldsymbol{p}$ gives an optimal solution, when the new retrieval pattern is not used.

Proposition 1. *For each $x_{K-1} \in [1, \infty)$, the vector $\boldsymbol{p}$ given by (17)-(19) is optimal for the optimization problem given above.*

Proof. We first write the Lagrangian of the problem,

$$\begin{aligned}
\mathscr{L} = \hat{I}(\boldsymbol{p}) &- \sum_{w=1}^{K-1} \lambda_w p_w \\
&+ \nu \left[\sum_{w=0}^{K-1} \binom{K-1}{w} (N-1)^w N p_w - 1 \right], \qquad (40)
\end{aligned}$$

where $\hat{I}(\boldsymbol{p})$ is defined as

$$\begin{aligned}
\hat{I}(\boldsymbol{p}) &= \frac{1}{K} \sum_{w=0}^{K} \binom{K}{w} (N-1)^w \\
&\Big\{ w p_{w-1} \log p_{w-1} + (K-w) p_w \log p_w \\
&- [w p_{w-1} + (K-w) p_w] \log \frac{w p_{w-1} + (K-w) p_w}{K} \Big\}. \qquad (41)
\end{aligned}$$

The KKT condition can be explicitly derived as follows: the partial derivatives of $\mathscr{L}$ w.r.t p_w are,

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial p_w} &= \binom{K-1}{w} (N-1)^w \big[-y_w - (N-1) y_{w+1} \\
&+ (N-1)\log x_{w+1} + N\nu \big] - \lambda_w, \quad w \in 1 : K-2, \qquad (42)
\end{aligned}$$

$$\frac{\partial \mathscr{L}}{\partial p_{K-1}} = (N-1)^{K-1} \big[-y_{K-1} + N\nu \big] - \lambda_{K-1}, \tag{43}$$

where we have introduced the two new sets of variables:

$$x_w \triangleq p_{w-1}/p_w, \tag{44}$$

$$y_w \triangleq \log \frac{w x_w + K - w}{K}. \tag{45}$$

It is straightforward to verify that x_w and y_w satisfying (17), with y_w's properly eliminated, which along with the following dual variable assignments

$$\lambda_w = 0, \ w \in 1 : K-1, \tag{46}$$

$$\nu = y_{K-1}/N, \tag{47}$$

render the partial derivatives zeros for all $w \in 1 : K-1$, and moreover satisfy all complementary slackness requirement. These assignments are thus a solution to the primal optimization problem. The mutual information leakage $I(Q_n; M)$ and the download cost D with this solution is exactly as the right hand sides of (20) and (21). □

It is straightforward to show that the new strategy (clean download from any single server) gives the extreme point $(\rho, D) = (\frac{\log K}{N}, 1)$. Together with the convexity of $D_{\text{MI}}(\rho)$ (see [34]), Theorem 2 is now obvious. In fact, for any optimized TSC code with $\boldsymbol{p} = \tilde{\boldsymbol{p}}$, probabilistic sharing with the new strategy results in (ρ, D) operating points on the straight line connecting the them, and the resultant code has a positive p_0' and $\boldsymbol{p} = (1 - Np_0')\tilde{\boldsymbol{p}}$.

In Fig. 1, the tangent point gives the threshold beyond which the new download pattern becomes effective in the sharing solution. It can be shown after some algebra that this occurs at $x_1 = (K-1)/(K^{\frac{N-2}{N-1}} - 1)$.

VI. Conclusion

We studied the the problem of weakly private information retrieval, and proposed a new code construction based on a simple yet critical observation on the minimum download extreme case. The optimizing query pattern probability distributions are provided for the maximal leakage metric and the mutual information leakage metrics, resulting in strict improvements in both case. The new inner bounds do not yet match the known outer bounds in the literature, and we are currently working on reducing this gap.

REFERENCES

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

[2] 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.

[3] 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, Nov. 2019.

[4] 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, Feb. 2019.

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

[6] R. Zhou, C. Tian, H. Sun, and J. S. Plank, "Two-level private information retrieval," in *2021 IEEE International Symposium on Information Theory (ISIT)*, Melbourne, Victoria, Australia, Jul. 2021, pp. 1919–1924.

[7] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4904–4916, Aug. 2020.

[8] T. Guo, R. Zhou, and C. Tian, "New results on the storage-retrieval tradeoff in private information retrieval systems," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 403–414, Mar. 2021.

[9] C. Tian, "On the storage cost of private information retrieval," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7539–7549, Dec. 2020.

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

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

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

[13] 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, Nov. 2018.

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

[15] 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, Feb. 2018.

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

[17] J. Zhu, Q. Yan, C. Qi, and X. Tang, "A new capacity-achieving private information retrieval scheme with (almost) optimal file length for coded servers," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1248–1260, 2019.

[18] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2999–3012, Mar. 2020.

[19] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[20] Z. Wang, K. Banawan, and S. Ulukus, "Private set intersection: A multi-message symmetric private information retrieval perspective," *IEEE Transactions on Information Theory*, in press.

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

[22] 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, May 2019.

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

[24] Z. Chen, Z. Wang, and S. A. Jafar, "The capacity of T-private information retrieval with private side information," *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4761–4773, Aug. 2020.

[25] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2023–2031, Apr. 2020.

[26] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information: the general cases," in *2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, Jun. 2020, pp. 1083–1088.

[27] Z. Wang and S. Ulukus, "Symmetric private information retrieval with user-side common randomness," in *2021 IEEE International Symposium on Information Theory (ISIT)*, Melbourne, Victoria, Australia, Jul. 2021, pp. 2119–2124.

[28] D. Asonov and J. C. Freytag, "Repudiative information retrieval," in *2002 ACM Workshop on Privacy in the Electronic Society*, Washington, DC, USA, Nov. 2002, pp. 32–40.

[29] R. R. Toledo, G. Danezis, and I. Goldberg, "Lower-cost ϵ-private information retrieval," in *2016 Privacy Enhancing Technologies Symposium (PETS)*, Darmstadt, Germany, Jul. 2016, pp. 184–201.

[30] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *2019 IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1262–1266.

[31] Z. Jia, "On the capacity of weakly-private information retrieval," Master's thesis, University of California, Irvine, CA, 2019.

[32] H.-Y. Lin, S. Kumar, E. Rosnes, A. G. i. Amat, and E. Yaakobi, "Weakly-private information retrieval," in *2019 IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jun. 2019, pp. 1257–1261.

[33] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric," in *2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, Jun. 2020, pp. 1089–1094.

[34] H.-Y. Lin, S. Member, S. Kumar, E. Rosnes, A. Graell Amat, and E. Yaakobi, "Multi-server weakly-private information retrieval," *IEEE Transactions on Information Theory*, vol. 68, no. 2, pp. 1197–1219, 2022.

[35] I. Samy, M. Attia, R. Tandon, and L. Lazos, "Asymmetric leaky private information retrieval," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5352–5369, Aug. 2021.

[36] H.-Y. Lin, S. Kumar, E. Rosnes, A. G. i. Amat, and E. Yaakobi, "The capacity of single-server weakly-private information retrieval," in *2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, Jun. 2020, pp. 1053–1058.

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