Private Information Retrieval When Private Noisy Side Information is Available

Hassan ZivariFard and Rémi A. Chou
Department of Electrical Engineering and Computer Science
Wichita State University, Wichita, KS
E-mails:{hassan.zivarifard, remi.chou}@wichita.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206733

***Abstract*— Consider Private Information Retrieval (PIR), where a client wants to retrieve one file out of K files that are replicated in N different servers and the client selection must remain private when up to T servers may collude. Additionally, suppose that the client has noisy side information about each of the K files, and the side information about a specific file is obtained by passing this file through one of D possible discrete memoryless test channels, where $D \leq K$. While the statistics of the test channels are known by the client and by all the servers, the specific mapping $\mathcal{M}$ between the files and the test channels is unknown to the servers. We study this problem when the client wants to preserve the privacy of its desired file selection and the mapping $\mathcal{M}$. For this problem setup, we derive the optimal download rate. Our problem setup generalizes PIR with private noiseless side information and PIR with private side information under storage constraints.**

I. Introduction

PIR refers to a problem where a client wishes to download, as efficiently as possible, one of the K files that are replicated among a set of distributed servers such that the servers cannot learn anything about the client's file selection [1], [2].

The PIR problem was studied in [3] from an information-theoretic point of view to characterize the maximum number of bits of desired information that can be retrieved privately per bit of downloaded information. In [3], the authors showed that this quantity is $(1+1/N+1/N^2+\cdots+1/N^{K-1})$ when a client wishes to retrieve one of the K files that are distributed in N replicated and non-colluding servers. This problem was subsequently extended to various scenarios. [4] considered a PIR problem where T of the N servers may collude and some of the servers may not respond. [5]–[7] studied PIR with N non-colluding servers, where each server stores an MDS-coded version of the K files. [8], [9], extended the results to symmetric PIR, in which the privacy of both the client and the servers is considered.

A. Overview of the setting studied in this paper

In this paper, we study a PIR problem where the client wants to retrieve one of the K files that are replicated in N servers and T of these servers may collude. As reviewed in the next section, so far, only PIR with noiseless side information, which means that the client has access to a subset of the files or portions of each file and their corresponding positions in the original files, has been studied in the literature. By contrast, in our problem setting, the client has a noisy version of each file which is obtained by passing each file through a discrete memoryless test channel. As depicted in Fig. 1, we assume that there are $D \leq K$ different test channels whose statistics are public knowledge and known by the client and the servers. We denote the mapping between the files and the test channels by $\mathcal{M}$. We study this problem when the client wants to preserve the privacy of both the intended file and the mapping $\mathcal{M}$, and we derive the optimal download rate.

B. Related works

As identified in [10], three main models for PIR with side information have been studied in the literature, which are summarized in the following.

- PIR with side information globally known by all the terminals: The effect of side information on the information-theoretic capacity of the PIR problem was first studied in [11], where the author considers a PIR problem in which a client wishes to privately retrieve one out of K files from N replicated non-colluding servers. Specifically, in [11], the client has a local cache that can store any function of the K files.
- PIR with non-private side information, where the privacy of the side information is not required: The single-server PIR problem where the client has access to a subset of the files and wants to protect only the identity of the desired file, is introduced and solved in [12]. An achievability result for the multiserver case is also derived in [12], and was later shown to be optimal in [13]. Single-server PIR when the client knows M files out of K files, or a linear combination of M files, has further been studied in [14]–[16] under various scenarios. Also, a multiserver PIR when the client has a noisy version of the desired file is studied in [17].
- PIR with private side information, where the joint privacy of the file selection and the side information is required: [12] derived an achievable rate region for N replicated and non-colluding servers. PIR from N replicated and non-colluding servers, where a cache-enabled client possesses side information, in the form of uncoded portions

The authors are with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS. This work was supported in part by NSF grant CCF-2047913. E-mails: {hassan.zivarifard, remi.chou}@wichita.edu.

of the files, that is unknown to the servers, is studied in [18]. Also, PIR from N replicated and non-colluding servers when the client knows M files out of K files as side information, and each server knows the identity of a subset of the side information files, is studied in [19]. In [10], the authors studied the PIR problem where the client wishes to retrieve one of the K files from N replicated servers, when T of the servers may collude, and the client has access to M files in a noiseless manner. This problem is extended to the case where the client wants to retrieve multiple files privately in [20].

Difference between our model and previous models: In this paper, we focus on PIR with private side information. Note that the side information in the PIR problems in [10]–[16], [18]–[22] is always noiseless, in the sense that all the side information available at the client corresponds to subsequences of each file and the client knows the corresponding symbol positions in the original files. By contrast to [10]–[16], [18]–[22], the side information in this paper is noisy, for instance, if the test channels are Binary Symmetric Channels (BSCs), then the client does not know which information bits have been flipped by the BSCs and which ones have not been flipped.

Previous works recovered as special case of our model: The problem studied in this paper subsumes the PIR problem [3], the PIR problem with colluding servers [4], the PIR problem with noiseless private side information [12, Theorem 2], the PIR problem with private side information under storage constraints [18], and the PIR problem with colluding servers and noiseless private side information [10] as special cases.

II. Notation

Let $\mathbb{N}_*$ be the set of positive natural numbers, and $\mathbb{R}$ be the set of real numbers. For any $a, b \in \mathbb{N}_*$ such that $a \leq b$, $[a:b]$ denotes the set $\{a, a+1, \ldots, b\}$, $[a]$ denotes the set $\{1, 2, \ldots, a\}$. Random variables are denoted by capital letters and their realizations by lower case letters. Superscripts denote the dimension of a vector, e.g., X^n. For a set of indices $\mathcal{I} \subset \mathbb{N}_*$, $\mathbf{X}_\mathcal{I}$ denotes $(X_i)_{i\in\mathcal{I}}$. $\mathbb{E}_X[\cdot]$ is the expectation with respect to the random variable X. The cardinality of a set is denoted by $|\cdot|$. For a mapping $\mathcal{F}: \mathcal{A} \to \mathcal{B}$, the preimage of $b \in \mathcal{B}$ by $\mathcal{F}$ is denoted as $\mathcal{F}^{-1}(b) \triangleq \{a \in \mathcal{A} : \mathcal{F}(a) = b\}$. For $D \in \mathbb{N}_*$ and a mapping $\mathcal{F}: [D] \to \mathbb{R}$, we represent the domain and co-domain of $\mathcal{F}$ as a matrix of dimension $2 \times D$ as

$$\mathcal{F} = \begin{pmatrix} 1 & 2 & \ldots & D \\ \mathcal{F}(1) & \mathcal{F}(2) & \ldots & \mathcal{F}(D) \end{pmatrix}.$$

III. Problem Statement

Consider a client and N servers, where up to T of these N servers may collude, and each server has a copy of K files of length n. Additionally, consider a set of D test channels, whose transition probabilities are known to the client and the servers, and whose outputs take value in finite alphabets. We assume that the client has noisy side information about all the K files in the sense that each file is passed through one of the D test channels, and the output of this test channel is available at the client but not the servers, as depicted in Fig. 1. The mapping $\boldsymbol{\mathcal{M}}$ between the files and the test channels is not known at the servers. The objective of the client is to retrieve one of the files such that the index of this file and the mapping $\boldsymbol{\mathcal{M}}$ are kept secret from the servers.

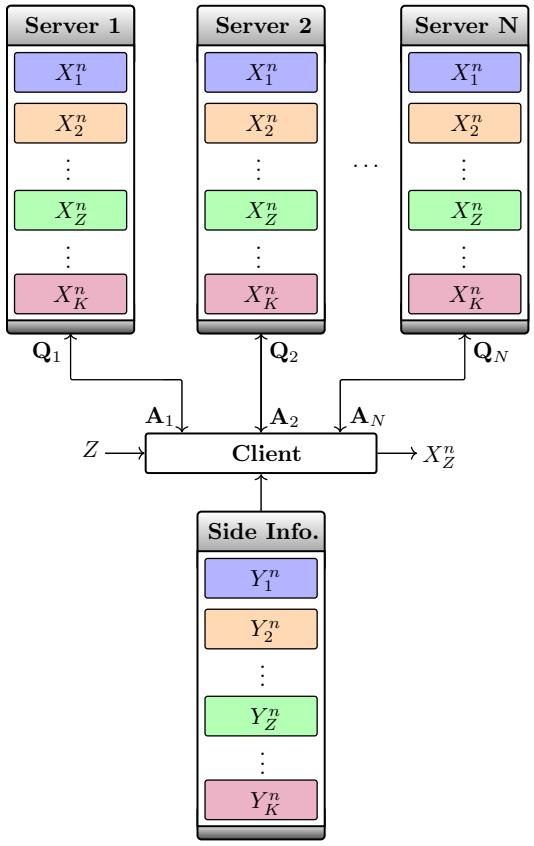

Fig. 1. PIR with private noisy side information and T-colluding servers, where the side information about a specific file is obtained by passing this file through one of D possible discrete memoryless test channels $\left(C^{(i)}\right)_{i\in[D]}$, where $D \leq K$, i.e., for $j \in [K]$, there exists some $i \in [D]$ such that Y_j^n is the output of channel $C^{(i)}$ when X_j^n is the input. Here, $\left(X_i^n\right)_{i\in[K]}$ are the K files that are replicated in N servers, $(\mathbf{Q}_i)_{i\in[N]}$ are the queries for the servers, and $(\mathbf{A}_i)_{i\in[N]}$ are the corresponding answers of the servers. Z is the index of the client's file selection and X_Z^n is the file desired by the client.

A. Problem definitions

Definition 1. *Consider $K, n, N, D \in \mathbb{N}_*$, $(d_i)_{i\in[D]} \in \mathbb{N}_*^D$ such that $\sum_{i=1}^D d_i = K$, and D distinct test channels $\left(C^{(i)}\right)_{i\in[D]}$, with $C^{(i)} \triangleq (\mathcal{X}, P^{(i)}_{X|Y}, \mathcal{Y}_i)$, where $\mathcal{X}$ and $\mathcal{Y}_i$, $i \in [K]$, are finite alphabets. Without loss of generality, assume that $H(U|V_i) \leq H(U|V_j)$, for $i, j \in [D]$ such that $i \leq j$, where U is uniformly distributed over $\mathcal{X}$ and V_i and V_j are the outputs of $C^{(i)}$ and $C^{(j)}$, respectively, when U is the input. A PIR protocol with private noisy side information and parameters $\left(K, n, N, D, (d_i)_{i\in[D]}, \left(C^{(i)}\right)_{i\in[D]}\right)$ consists of,*

- *N servers;*
- *K independent random sequences $\mathbf{X}_{[K]}^n$ uniformly distributed over $\mathcal{X}^n$, which represent K files shared at each of the N servers;*
- *D distinct test channels $\left(C^{(i)}\right)_{i\in[D]}$;*

- *a mapping $\boldsymbol{\mathcal{M}}$ chosen at random from the set $\mathfrak{M} \triangleq \{\boldsymbol{\mathcal{M}} : [K] \to [D] : \forall i \in [D], \left|\boldsymbol{\mathcal{M}}^{-1}(i)\right| = d_i\}$; this mapping is only known at the client and not at the servers;*
- *for each file X_i^n, where $i \in [K]$, the client has access to a noisy version of X_i^n, denoted by $Y^n_{i,\boldsymbol{\mathcal{M}}(i)}$, which is the output of the test channel $\mathrm{C}^{(\boldsymbol{\mathcal{M}}(i))}$ when X_i^n is the input;*
- *the random variable Z is uniformly distributed over $[K]$ and represents the index of the file that the client wishes to retrieve, i.e., the client wants to retrieve the file X_Z^n.*
- *a stochastic and one-to-one query function $\mathcal{F}_i : [K] \times \mathfrak{M} \times \mathcal{Y}^n_{[K]} \to \boldsymbol{\mathcal{Q}}_i$, for $i \in [N]$, where $\boldsymbol{\mathcal{Q}}_i$ is a finite alphabet;*
- *for $i \in [N]$, a deterministic answer function $\mathcal{E}_i : \boldsymbol{\mathcal{Q}}_i \times \mathcal{X}^{nK} \to \left[2^{nR(\mathbf{Q}_i)}\right]$;*
- *a decoding function $\mathcal{D} : [K] \times \mathfrak{M} \times \left[2^{n\sum_{i=1}^N R(\mathbf{Q}_i)}\right] \times \mathcal{Y}^{nK} \to \mathcal{X}^n$;*

and operates as follows,

1) *the client creates the queries $\mathbf{Q}_i \triangleq \mathcal{F}_i\big(Z, \boldsymbol{\mathcal{M}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\big)$, where $\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}} \triangleq \left(Y^n_{i,\boldsymbol{\mathcal{M}}(i)}\right)_{i\in[K]}$, and sends it to Server $i \in [N]$; we assume that the queries must be of negligible length compared to the file length n, i.e., $\log|\mathcal{Q}_i| = o(n)$, for $i \in [N]$;*[1]
2) *then, for all $i \in [N]$, Server i creates the answer $\mathbf{A}_i \triangleq \mathcal{E}_i\big(\mathbf{Q}_i, \mathbf{X}^n_{[K]}\big)$, where $\mathbf{X}^n_{[K]} \triangleq (X_i^n)_{i\in[K]}$, and sends it to the client; therefore,*

$$H\left(\mathbf{A}_i \middle| \mathbf{Q}_i, \mathbf{X}^n_{[K]}\right) = 0, \quad \forall i \in [N]; \tag{1}$$

3) *finally, the client computes an estimate of X_Z^n as $\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\right)$, where $\mathbf{A}_{[N]} \triangleq (\mathbf{A}_i)_{i\in[N]}$.*

Therefore, the probability of error at the client is,

$$P_e \triangleq \limsup_{n\to\infty} \mathbb{P}\left[\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\right) \neq X_Z^n\right]. \tag{2}$$

$R\left(\mathbf{Q}_{[N]}\right) \triangleq \sum_{i=1}^N R\left(\mathbf{Q}_i\right)$, where $\mathbf{Q}_{[N]} \triangleq (\mathbf{Q}_i)_{i\in[N]}$, is the rate of the PIR protocol and is random with respect to $\mathbf{Q}_{[N]}$, which makes the protocol a variable length coding scheme. We also define the expected rate of the protocol as $R \triangleq \mathbb{E}_{\mathbf{Q}_{[N]}}[R\left(\mathbf{Q}_{[N]}\right)]$.

We keep the index of the desired file Z and the mapping $\boldsymbol{\mathcal{M}}$ private from the servers.

Definition 2 ($\mathrm{C}_{\text{PIR-PNSI}}$ capacity)**.** *An expected rate $R \in \mathbb{R}_+$ is achievable with private noisy side information, when up to T servers may collude, if there exists PIR protocols such that, for any set $\mathcal{T} \subseteq [N]$ such that $|\mathcal{T}| = T$,*

$$P_e = 0, \tag{3a}$$

$$I\left(\mathbf{Q}_\mathcal{T}, \mathbf{A}_\mathcal{T}, \mathbf{X}^n_{[K]}; Z, \boldsymbol{\mathcal{M}}\right) = 0. \tag{3b}$$

The privacy metric (3b) *means that the client file choice Z and mapping $\boldsymbol{\mathcal{M}}$ must be kept secret from any T colluding servers. The supremum of all achievable rates is referred to as the PIR capacity with private noisy side information, and is denoted by $\mathrm{C}_{\text{PIR-PNSI}}$.*

B. Examples

In Example 1, Example 2, and Example 3, we show that our problem setup recovers the problem setup for PIR with colluding servers [4], PIR with colluding servers and noiseless side information [10], [12], and PIR with private side information under storage constraints [18]. Then, we illustrate our definitions when $K = D = 2$ and $T = 1$ in Example 4.

Example 1 (PIR with colluding servers)**.** *When $D = 1$ and the test channel is a BEC with parameter $\epsilon = 1$, then the client has no side information about the files. In this case, Definition 1 reduces to PIR without side information as introduced in [4], and the privacy constraint in Definition 2 is equivalent to the privacy constraint in [4].*

Example 2 (PIR with private noiseless side information)**.** *When $D = 2$ and the test channels are BECs with parameters $\epsilon_1 = 0$, and $\epsilon_2 = 1$, the client has access to d_1 files in a noiseless manner as side information. This case corresponds to PIR with side information as introduced in [12, Theorem 2] for non-colluding servers and in [10, Theorem 1] for colluding servers.*

Example 3 (PIR with private side information under storage constraints)**.** *Suppose that $T = 1$, $D = M + 1$, for $M \in \mathbb{N}_*$ and $M \leq K$, the test channels are BECs with parameters $\epsilon_D = 1$, $\epsilon_i = 1 - r_i$, for $i \in [M]$, with $r_1 \geq r_2 \geq \cdots \geq r_M$, and $d_i = 1$, for $i \in [M]$. This problem setup, under the privacy constraint in Definition 2, is related to the problem studied in [18]. The difference with [18] is that the positions of the erasures are known at the servers in [18], whereas in our setting, the positions of the erasures are random and unknown at the servers. Therefore, the optimal download rate for our problem setup in this example might be higher than the download rate in [18]. However, we will show in the next section that the same download rate as in [18] is achievable.*

Example 4 (When $K = D = 2$, $T = 1$, and $d_1 = d_2 = 1$)**.** *Let X_1^n and X_2^n be the two files at the server and $Y^n_{i,\boldsymbol{\mathcal{M}}(i)}$ be the side information about X_i^n, $i \in \{1, 2\}$, available at the client but unavailable at the server, where $Y^n_{i,\boldsymbol{\mathcal{M}}(i)}$ is the output of the test channel $C^{\boldsymbol{\mathcal{M}}(i)}$ when the input is X_i^n. Note that $\boldsymbol{\mathcal{M}}$ can take two values (with the notation introduced in Section II):*

$$\mathbf{M}_1 \triangleq \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{M}_2 \triangleq \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

[1] When $D = 1$ and the test channel is a Binary Erasure Channel (BEC) with parameter $\epsilon_1 = 1$, or when $D = 2$ and the test channels are BECs with parameters $\epsilon_1 = 0$ and $\epsilon_2 = 1$, which correspond to PIR without side information in [4] and PIR with side information in [10], respectively, it is shown in [4], [10] that there is no loss of generality by making this assumption. In general, allowing the query rate to be non-negligible with the file length n is a different problem. However, similar to [12, Remark 1] and [18], this assumption can also be removed in our converse proofs when the queries $\mathbf{Q}_i$, for $i \in [N]$, are only allowed to depend on $(Z, \boldsymbol{\mathcal{M}})$.

When $Z=1$, since there are two different possibilities for the side information about X_1^n, that are $Y_{1,1}^n$ and $Y_{1,2}^n$, we define,

$$P_e^{(1)}\big(Z=1,\mathbf{M}_1\big) \triangleq \mathbb{P}\Big[\mathcal{D}\left(Z,\boldsymbol{\mathcal{M}},\mathbf{A}_{[N]},Y_{1,1}^n,Y_{2,2}^n\right)\neq X_1^n\Big|Z=1,\boldsymbol{\mathcal{M}}=\mathbf{M}_1\Big],$$
$$P_e^{(2)}\big(Z=1,\mathbf{M}_2\big) \triangleq \mathbb{P}\Big[\mathcal{D}\left(Z,\boldsymbol{\mathcal{M}},\mathbf{A}_{[N]},Y_{1,2}^n,Y_{2,1}^n\right)\neq X_1^n\Big|Z=1,\boldsymbol{\mathcal{M}}=\mathbf{M}_2\Big].$$

Similarly, when $Z=2$, since there are two different possibilities for the side information about X_2^n at the server, that are $Y_{2,1}^n$ and $Y_{2,2}^n$, we define,

$$P_e^{(3)}\big(Z=2,\mathbf{M}_1\big) \triangleq \mathbb{P}\Big[\mathcal{D}\left(Z,\boldsymbol{\mathcal{M}},\mathbf{A}_{[N]},Y_{1,1}^n,Y_{2,2}^n\right)\neq X_2^n\Big|Z=2,\boldsymbol{\mathcal{M}}=\mathbf{M}_1\Big],$$
$$P_e^{(4)}\big(Z=2,\mathbf{M}_2\big) \triangleq \mathbb{P}\Big[\mathcal{D}\left(Z,\boldsymbol{\mathcal{M}},\mathbf{A}_{[N]},Y_{1,2}^n,Y_{2,1}^n\right)\neq X_2^n\Big|Z=2,\boldsymbol{\mathcal{M}}=\mathbf{M}_2\Big].$$

Therefore, the probability of error in (2)*, by taking* $\limsup$ *when $n\to\infty$, is equal to*

$$\begin{aligned}&\mathbb{P}[Z=1]\mathbb{P}\big[\boldsymbol{\mathcal{M}}=\mathbf{M}_1\big]P_e^{(1)}\big(Z=1,\mathbf{M}_1\big)\\&+\mathbb{P}[Z=1]\mathbb{P}\big[\boldsymbol{\mathcal{M}}=\mathbf{M}_2\big]P_e^{(2)}\left(Z=1,\mathbf{M}_2\right)\\&+\mathbb{P}[Z=2]\mathbb{P}\big[\boldsymbol{\mathcal{M}}=\mathbf{M}_1\big]P_e^{(3)}\big(Z=2,\mathbf{M}_1\big)\\&+\mathbb{P}[Z=2]\mathbb{P}\big[\boldsymbol{\mathcal{M}}=\mathbf{M}_2\big]P_e^{(4)}\left(Z=2,\mathbf{M}_2\right).\end{aligned}$$

IV. Main Results

In this section, we provide the main results of the paper and present some examples that recover and extend known results.

Theorem 1. *Consider K files that are replicated in N servers where up to T of them may collude. Then, the capacity of PIR with private noisy side information is*

$$\mathrm{C}_{\text{PIR-PNSI}}=\sum_{\ell=1}^{D}H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right),$$

where $\Psi^{-1}(A,B)\triangleq\big(1+A+A^2+\cdots+A^{B-1}\big)$, and for $i,j\in\mathbb{N}_*$, $d_{[i:j]}\triangleq\sum_{t=i}^{j}d_t$, when $i\leq j$, and $d_{[i:j]}\triangleq 0$, when $i>j$.

Proof. The achievability proof, which is outlined here, is based on source coding with side information and the achievability schemes in [4], [10]. We use a nested random binning scheme and assign D nested random bin indices to each file X_i^n, for $i\in[K]$. Specifically, the ℓ^{th} random bin indices of all the files, referred to as ℓ^{th} database, enable lossless reconstruction of the d_ℓ files that are associated with the test channel $C^{(\ell)}$. Therefore, by downloading the database ℓ, the client, in addition to being able to reconstruct the d_ℓ files that are associated with the test channel $C^{(\ell)}$, also receives the ℓ^{th} random bin indices of all the other files. The achievability scheme consists in successively downloading each of the D databases, in ascending order, by using the same coding scheme and query structure as [10], for each database. The details of the proof are available in Section V-A. The converse proof is omitted due to the space limitation and is available in [23]. □

Remark 1 (Index of random variables)**.** *Since all the files are generated according to the same distribution, namely, the uniform distribution over $\mathcal{X}^n$, the index 1 of X_1 and $Y_{1,\ell}$ in Theorem 1 can be replaced with any other index $i\in[K]$.*

Corollary 1. *Consider K files, that are replicated in N servers where up to T of them may collude. Additionally, the test channels are BECs with parameters $(\epsilon_i)_{i\in[D]}\in[0,1]^D$ such that $\epsilon_i<\epsilon_j$, for $i,j\in\mathbb{N}_*$ and $i<j$. Then, the capacity of PIR with private noisy side information is*

$$\mathrm{C}_{\text{PIR-PNSI}}=\sum_{\ell=1}^{D}\epsilon_\ell\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right).$$

Example 5 (No side information)**.** *In Corollary 1, if we set $D=1$, and $\epsilon_1=1$, which means that the client has no side information and $d_D=K$, then the capacity result in Corollary 1 reduces to [4, Theorem 1], i.e.,*

$$\mathrm{C}_{\text{PIR-PNSI}}=\left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{K-1}\right).$$

Example 6 (Private noiseless side information)**.** *In Corollary 1, if we set $D=2$ and $T=1$, $\epsilon_1=0$, which means that the client knows d_1 files as side information in a noiseless manner, and $\epsilon_2=1$, which means that there is no side information about d_2 files, then the capacity result in Corollary 1 reduces to [12, Theorem 2], i.e.,*

$$\mathrm{C}_{\text{PIR-PNSI}}=\left(1+\frac{1}{N}+\left(\frac{1}{N}\right)^2+\cdots+\left(\frac{1}{N}\right)^{K-d_1-1}\right).$$

Example 7 (PIR with private side information under storage constraints)**.** *Set $T=1$, $D=M+1$, for $M\in\mathbb{N}_*$ and $M<K$. If we set $d_i=1$, for $i\in[M]$, $d_{M+1}=K-M$, and $\epsilon_i=1-r_i$, with $r_1\geq r_2\geq\cdots\geq r_M$, and $\epsilon_D=1$ then the capacity in Corollary 1 reduces to*

$$\mathrm{C}_{\text{PIR-PNSI}}=\frac{1-r_1}{N^{K-1}}+\frac{1-r_2}{N^{K-2}}+\frac{1-r_3}{N^{K-3}}+\cdots+\frac{1-r_{M-1}}{N^{K-M+1}}+\frac{1-r_M}{N^{K-M}}+1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^{K-M-1}}.$$

Note that, this result is stronger than that of [18, Theorem 1], since in [18] it is assumed that the client knows the first r_i bits, for $i\in[M]$, of M randomly selected files, whereas, in our setting, the r_i bits of side information for file i are chosen at random and we obtain the same capacity.

V. Proof of Theorem 1

A. Achievability proof

A high level description of the achievability scheme is provided after Theorem 1. Assume that each file is of length $n=N^K$ with symbols in a sufficient large finite field $\mathbb{F}_q$. Fix $\delta>0$.

1) Random Binning: For every file x_i^n, $i \in [K]$, assign D random bin indices as follows. For $\ell \in [D]$, randomly and independently assign a bin index $j_\ell^{(i)} \in \mathcal{J}_\ell \triangleq \left[q_\ell^n\right]$ to x_i^n, where $q_\ell \triangleq q^{R_\ell}$, with $R_\ell > 0$, to be defined later, and $R_0 \triangleq 0$.

We refer to $\mathbf{M}_\ell \triangleq \left(j_\ell^{(1)}, \ldots, j_\ell^{(K)}\right)_{\ell \in [D]}$ as the database ℓ. The query is constructed to retrieve each one of the $\mathbf{M}_\ell$ databases in ascending order.

2) Query Structure Construction: The client constructs the query in D different levels. In the first level, we apply to the database $\mathbf{M}_1$ the same query structure as in [4], which consists of K sublevels. In the level $\ell \in [2:D]$, we apply to the database $\mathbf{M}_\ell$ the same query structure as in [10], which also consists of K sublevels. Specifically, as in [10], the k_ℓ^{th} sublevel consists of sums of k_ℓ symbols, which are called k-sums. There are $\binom{K}{k_\ell}$ different types of k-sums and $(N-T)^{k_\ell-1}T^{K-k_\ell}$ different instances of each type in the k_ℓ^{th} sublevel. Hence, the total number of symbols that will be downloaded from each server is $\sum_{k_\ell=1}^{K}\binom{K}{k_\ell}(N-T)^{k_\ell-1}T^{K-k_\ell}$.

3) Query Specialization: For $\ell \in [D]$, we do the query structure construction and query specialization without considering availability of any side information as in [10], and denote this scheme by Π_ℓ. Then, we do query redundancy removal based on availability of noiseless side information similar to [10]. Specifically, after each level $\ell \in [D]$, the client is able to recover the d_ℓ files that are associated with the ℓ^{th} test channel, and therefore considering the files that it has decoded in the previous levels, the client knows $\mathbf{X}_{[d_{[\ell]}]}^n$ and therefore it knows $\left(j_{\ell+1}^{(1)}, \ldots, j_{\ell+1}^{(d_{[\ell]})}\right)$, which is used as noiseless side information to recover $\left(j_{\ell+1}^{(d_{[\ell]}+1)}, \ldots, j_{\ell+1}^{(K)}\right)$ in level $\ell+1$. For level $\ell = 1$, the client does not have any noiseless side information and cannot perform query redundancy removal but, for level $\ell \in [2:D]$, since it has recovered $\sum_{t=1}^{\ell-1} d_t$ files, the client can perform query redundancy removal. For each $\ell \in [D]$ and for each server, let $p_{\ell,1}$ denote the number of symbols downloaded with Π_ℓ. Out of these $p_{\ell,1}$ symbols, we denote by $p_{\ell,2} < p_{\ell,1}$ the number of symbols that the client already knows by decoding some of the files in the previous levels. For $\ell \in [D]$, let $\mathbf{U}_{\ell,j} \in \mathbb{F}_{q_\ell}^{p_{\ell,1}}$ denote the symbols downloaded from the j^{th} server with Π_ℓ. For each server, use a systematic $(2p_{\ell,1}-p_{\ell,2}, p_{\ell,1})$ Maximum Distance Separable (MDS) code [24], with generator matrix $\mathbf{G}_{(2p_{\ell,1}-p_{\ell,2})\times p_{\ell,1}} = [\mathbf{V}_{p_{\ell,1}\times(p_{\ell,1}-p_{\ell,2})}|\mathbf{I}_{p_{\ell,1}\times p_{\ell,1}}]^\intercal$ to encode the $p_{\ell,1}$ symbols into $2p_{\ell,1}-p_{\ell,2}$ symbols, of which $p_{\ell,1}$ are systematic, and $p_{\ell,1}-p_{\ell,2}$ are parity symbols, such that it is sufficient to download $\mathbf{V}^\intercal_{p_{\ell,1}\times(p_{\ell,1}-p_{\ell,2})}\mathbf{U}_{\ell,j}$. For level $\ell = 1$, since the client does not have any noiseless side information about $\mathbf{M}_1$, $p_{1,2} = 0$.

4) Decoding: For $\ell \in [D]$, after reconstructing $(j_i^{(t)})_{i\in[\ell]}$ from 3), for $t \in \boldsymbol{\mathcal{M}}^{-1}(\ell)$, given $\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}$, the client declares $\hat{X}_t^n$ to be an estimate of the sequence X_t^n if it is a unique sequence that is typical with $Y_{t,\ell}^n$ in the bin $(j_i^{(t)})_{i\in[\ell]}$. According to the Slepian-Wolf theorem, e.g. [25], the decoding is successful, i.e., $\mathbb{P}\big[\hat{X}_t^n \neq X_t^n\big] \xrightarrow[n\to\infty]{} 0$, if

$$\sum_{i=1}^{\ell} R_i = H\big(X_t|Y_{t,\ell}\big) + \delta. \tag{4}$$

5) Rate Calculation: Similar to [10], for the scheme Π_ℓ, the total number of downloaded symbols from each server is $p_{\ell,1} = \sum_{k_\ell=1}^{K}\binom{K}{k_\ell}(N-T)^{k_\ell-1}T^{K-k_\ell}$, $\ell \in [D]$ and out of these $p_{\ell,1}$ symbols $p_{\ell,2} = \sum_{k_\ell=1}^{d_{[\ell-1]}}\binom{d_{[\ell-1]}}{k_\ell}(N-T)^{k_\ell-1}T^{K-k_\ell}$ symbols are already known at the client, where $d_{[\ell-1]} \triangleq \sum_{i=1}^{\ell-1} d_i$ and $d_{[0]} = 0$. Then, similar to [10, Eq. (22), (25)], we have

$$p_{\ell,1} = \frac{N^K - T^K}{N-T}, \tag{5a}$$

$$p_{\ell,2} = \frac{T^{K-d_{[\ell-1]}}\big(N^{d_{[\ell-1]}} - T^{d_{[\ell-1]}}\big)}{N-T}. \tag{5b}$$

Therefore, the transmission rate for the level ℓ is,

$$\begin{aligned}
R^{(\ell)} &= \frac{R_\ell N(p_{\ell,1}-p_{\ell,2})}{n} \\
&\overset{(a)}{=} \frac{R_\ell N(p_{\ell,1}-p_{\ell,2})}{N^K} \\
&\overset{(b)}{=} \frac{R_\ell\left(1-\left(\frac{T}{N}\right)^{K-d_{[\ell-1]}}\right)}{\left(1-\frac{T}{N}\right)} \\
&\overset{(c)}{=} \Big(H\big(X_1|Y_{1,\ell}\big) - H\big(X_1|Y_{1,\ell-1}\big)\Big)\sum_{i=0}^{K-d_{[\ell-1]}-1}\left(\frac{T}{N}\right)^i,
\end{aligned}$$

where

(a) follows since $n = N^K$;
(b) follows from (5);
(c) follows from (4).

Therefore, the total transmission rate is,

$$\begin{aligned}
R &= \sum_{\ell=1}^{D} R^{(\ell)} \\
&= \sum_{\ell=1}^{D}\Big(H\big(X_1|Y_{1,\ell}\big) - H\big(X_1|Y_{1,\ell-1}\big)\Big)\sum_{i=0}^{K-d_{[\ell-1]}-1}\left(\frac{T}{N}\right)^i \\
&= \sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{K-d_{[\ell]}}\sum_{i=0}^{d_\ell-1}\left(\frac{T}{N}\right)^i.
\end{aligned}$$

6) Privacy Analysis: Note that for all the D levels, the client does not use any side information to construct the queries. Indeed, the systematic MDS codes of all the levels in the query redundancy removal do not depend on the side information that the client obtain after each level. The decoding in 4) starts when the client collects all the answers from the servers for all the D levels. Thus, the side information is used only when the client collects all the answers from the servers for all the D levels. Therefore, privacy is inherited from the privacy of the schemes in [10] and [4].

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 36th Annu. Symp. Found. Comput. Sci.*, 1995, pp. 41–50.

[2] ——, "Private information retrieval," *Journal of the ACM*, vol. 45, no. 6, pp. 965–981, Nov. 1998.

[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] ——, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 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] J. Li, D. Karpuk, and C. Hollanti, "Towards practical private information retrieval from MDS array codes," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3415–3425, Jun. 2020.

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

[9] Q. Wang and M. Skoglund, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3183–3197, May 2019.

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

[11] R. Tandon, "The capacity of cache aided private information retrieval," in *Proc. 55th Annu. Allerton Conf. on Commun., Control, and Comput. (Allerton)*, Monticello, IL, Oct. 2017, pp. 1078–1082.

[12] 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.

[13] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," in *Proc. 54th Annual Conf. on Info. Sciences and Systems (CISS)*, 2020, pp. 1–6.

[14] A. Heidarzadeh, B. Garcia, S. Kadhe, S. El Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. on Commun., Control, and Comput. (Allerton)*, Monticello, IL, Oct. 2018, pp. 180–187.

[15] A. Heidarzadeh, S. Kadhe, S. El Rouayheb, and A. Sprintson, "Single-server multi-message individually-private information retrieval with side information," in *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Paris, France, Jun. 2019, pp. 1042–1046.

[16] 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.

[17] B. Herren, A. Arafa, and K. Banawan, "Download cost of private updating," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.

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

[19] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8222–8231, Dec. 2019.

[20] M. Jafari Siavoshani, S. P. Shariatpanahi, and M. A. Maddah-Ali, "Private information retrieval for a multi-message scenario with private side information," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3235–3244, May 2021.

[21] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3215–3232, May 2019.

[22] ——, "Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1126–1139, Jun. 2018.

[23] H. ZivariFard and R. Chou, "Private information retrieval with private noisy side information," *submitted to IEEE Trans. Inf. Theory*, Jan. 2023.

[24] S. Lin and D. J. Costello, *Error Control Coding*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[25] A. El Gamal and Y.-H. Kim, *Network Information Theory*, 1st ed. Cambridge, U.K: Cambridge University Press, 2012.