Private Information Retrieval With Private Noisy Side Information

Hassan ZivariFard and Rémi A. Chou

***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 under two different privacy metrics. Under the first privacy metric, the client wants to preserve the privacy of its desired file selection and the mapping $\mathcal{M}$. Under the second privacy metric, the client wants to preserve the privacy of its desired file and the mapping $\mathcal{M}$ but is willing to reveal the index of the test channel that is associated to its desired file. For both of these two privacy metrics, we derive the optimal normalized download cost. Our problem setup generalizes PIR with colluding servers, PIR with private noiseless side information, and PIR with private side information under storage constraints.**

***Index Terms*— Private Information Retrieval (PIR), capacity, optimal download cost, colluding servers, noisy side information.**

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 [2], [3]. Aside from its direct applications in data security and privacy, it is closely related to many fundamental problems such as secret sharing [4], [5] and oblivious transfer [6], [7], which is also called symmetric PIR and is a PIR problem where the server wants to keep any non-selected file private from the client. Therefore, PIR is a subject that relates to different areas in computer science.

Manuscript received 6 January 2023; revised 23 August 2023; accepted 13 January 2024. Date of publication 7 February 2024; date of current version 19 March 2024. This work was supported in part by NSF under Grant CCF-2047913. An earlier version of this paper was presented in part at the 2023 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT54713.2023.10206733]. *(Corresponding author: Hassan ZivariFard.)*

Hassan ZivariFard is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: hz2863@columbia.edu).

Rémi A. Chou is with the Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: remi.chou@uta.edu).

Communicated by C. Tian, Associate Editor for Security and Privacy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3363133.

Digital Object Identifier 10.1109/TIT.2024.3363133

The PIR problem was studied in [8] 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 [8], the authors showed that this quantity is $(1+1/N+1/N^2+\cdots+1/N^{K-1})^{-1}$ 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. Reference [9] considered a PIR problem where T of the N servers may collude and some of the servers may not respond. References [10], [11], and [12] studied PIR with N non-colluding servers, where each server stores an MDS-coded version of the K files. References [13] and [14], 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, only PIR with noiseless side information has been studied in the literature, i.e., the client has access to a subset of the files or portions of each file and their corresponding positions in the original files. 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. We assume that there are $D \leq K$ different test channels whose statistics are public knowledge and known by the client and the servers.[1] We denote the mapping between the files and the test channels by $\mathcal{M}$. We study this problem under two different privacy metrics. For the first privacy metric, the client wants to keep the index of the desired file and the entire mapping $\mathcal{M}$ secret from the servers, and this includes the index of the test channel that is associated with the desired file. For the second privacy metric, the client wants to keep the index of the desired file and the mapping $\mathcal{M}$ secret from the servers, but the client is willing to reveal the index of the test channel that is associated with the desired file, i.e., $\mathcal{M}(Z)$. For both privacy metrics, we derive the optimal normalized download cost, and we show that the second privacy metric always leads to a lower normalized download cost.

[1] We assume that the statistics of the test channels, i.e. $C^{(\ell)}$, $\ell \in [D]$, are public information. Note that for each file, X_k^n, for $k \in [K]$, the client has side information about X_k^n which can potentially be $\emptyset$, consequently, no more than K test channels are needed to model the side information available at the client.

B. Motivations

Consider the following motivational example of PIR, e.g., [15]: a stock market investor may want to privately retrieve some of the stock records because showing interest in one specific record could undesirably affect its value. Now, consider the case where an investor has already retrieved some or all of the stock records in the past. The investor could now retrieve a record by leveraging their knowledge of outdated records, which represents side information. As another example, the client could have acquired the noisy side information in several ways. For example, the user could have acquired a noisy version of the files opportunistically from other users in its network, overheard them from a wireless noisy broadcast channel, or downloaded them previously through classical PIR schemes from other servers. Note that the availability of noisy side information encompasses having obtained parts of the files in a noiseless manner. Also, the noise could have been the result of storing the files for a long period of time.

Note that in the stock market example, revealing the mapping between the stock records and the test channels shows how much information the investor has about each stock record, which is not in the interest of the investor since they do not want to affect the value of the stock records. Additionally, if the client has a subset of the files in a noiseless manner as side information [16], [17], then not keeping private the mapping between the test channels and the files may reveal to the servers the indices of the files that are available as side information at the client. These are examples of our first privacy metric. We also consider a privacy metric where the client reveals the index of the test channel associated with the desired file. As discussed in Example 9 and Remark 3, this privacy metric can lead to a lower download cost, when, for example, the desired file is available in the side information in a noiseless manner and the client does not need to download anything.

C. Related Works

As identified in [17], three main models for PIR with side information have been studied in the literature, which are summarized as follows.

- 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 [18], 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 [18], the client has a local cache that can store any function of the K files.
- PIR with 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 [16]. An achievability result for the multiserver case is also derived in [16], and was later shown to be optimal in [19]. 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 [20], [21], and [22] under various scenarios. Also, a multiserver PIR when the client has a noisy version of the desired file is studied in [15]. A more general notion of partial privacy for noiseless side information is introduced in [23], where a subset of the files available as side information is kept private from the servers while the complement of this subset of files is not required to be kept private.
- PIR with private side information, where the joint privacy of the file selection and the side information is required: [16] derived an achievable normalized download cost 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 of the files, that is unknown to the servers, is studied in [24]. Specifically, in [24], the client knows the first r_i bits, for $i \in [M]$, of M randomly selected files, and the identities of these side information files need to be kept private from the servers. 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 [25]. In [17], 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 [26].

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 [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], and [28] is always noiseless, in the sense that all the side information available at the client corresponds to sub-sequences of each file and the client knows the corresponding symbol positions in the original files. By contrast to [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], and [28], the side information in this paper is noisy, for instance, if the files are binary and 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 cases of our model: Since the side information considered in this paper is generated by passing the files through some Discrete Memoryless Channels (DMCs), our problem setup can recover the previous works if we assume that the test channels are Binary Erasure Channels (BECs). This is because passing a file through a BEC with parameter 0 means that the side information is equal to the input file, and passing a file through a BEC with parameter 1 means that there is no side information. For example, when there is only one BEC with parameter 1, then the problem studied in this paper subsumes the PIR problem [8] and the PIR problem with colluding servers [9]. If we assume there are two BECs with parameters 0 and 1, then the problem studied in this paper subsumes the PIR problem with private noiseless side information [16, Theorem 2], and the PIR problem with colluding servers and private noiseless side information [17] as special cases. If we assume there are $M+1$, where $M \in \mathbb{N}_*$, BECs with parameters $1-r_i$, then the problem studied here subsumes the PIR problem with private side information under

storage constraints [24]. We provide more details about each of these scenarios after we formally define the problem.

D. 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 lowercase letters unless specified otherwise. 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 $\mathcal{S}$ is denoted by $|\mathcal{S}|$. For a mapping $\boldsymbol{\mathcal{M}} : \mathcal{A} \to \mathcal{B}$, the preimage of $b \in \mathcal{B}$ by $\boldsymbol{\mathcal{M}}$ is denoted as $\boldsymbol{\mathcal{M}}^{-1}(b) \triangleq \{a \in \mathcal{A} : \boldsymbol{\mathcal{M}}(a) = b\}$. For $K \in \mathbb{N}_*$ and a mapping $\boldsymbol{\mathcal{M}} : [K] \to \mathbb{R}$, we represent the domain and co-domain of $\boldsymbol{\mathcal{M}}$ as a matrix of dimension $2 \times K$ as

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} 1 & 2 & \ldots & K \\ \boldsymbol{\mathcal{M}}(1) & \boldsymbol{\mathcal{M}}(2) & \ldots & \boldsymbol{\mathcal{M}}(K) \end{pmatrix}.$$

E. Paper Organization

The remainder of this paper is organized as follows. We formally define the problem in Section II. We present our main results in Section III and provide the proofs in Section IV and Section V. Finally, we provide concluding remarks in Section VI.

II. 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.

A. Problem Definitions

Definition 1: Consider $K, N, T, D, n \in \mathbb{N}_*$, $(d_i)_{i\in[D]} \in \mathbb{N}_*^D$ 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 $\left(K, N, T, D, n, (d_i)_{i\in[D]}, \left(C^{(i)}\right)_{i\in[D]}\right)$ consists of,

- N servers, where up to T of these servers may collude;
- K independent random sequences $\mathbf{X}^n_{[K]}$ uniformly distributed over $\mathcal{X}^n$, which represent K files shared at each of the N servers;

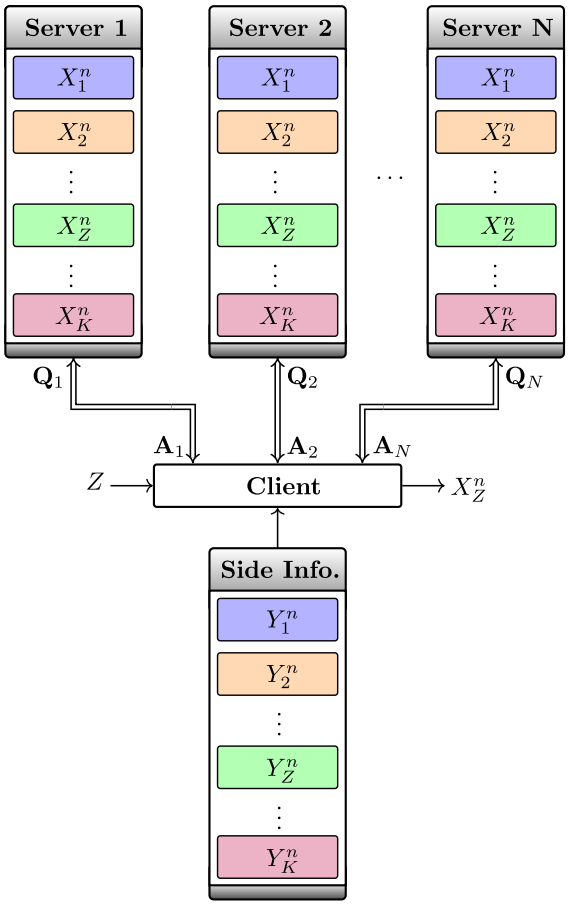

Fig. 1. PIR with private noisy side information and T-colluding servers, where the side information about a specific file is obtained by passing this file through one of D possible DMCs $\left(C^{(i)}\right)_{i\in[D]}$, where $D \leq K$, i.e., for $j \in [K]$, there exists some $i \in [D]$ such that 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 desired file by the client.

- D distinct test channels $\left(C^{(i)}\right)_{i\in[D]}$;
- a mapping $\boldsymbol{\mathcal{M}}$ chosen at random from the set $\mathfrak{M} \triangleq \left\{\boldsymbol{\mathcal{M}} : [K] \to [D] : \forall i \in [D], \left|\boldsymbol{\mathcal{M}}^{-1}(i)\right| = d_i\right\}$; 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 represents the index of the file that the client wishes to retrieve, i.e., the client wants to retrieve the file X_Z^n; when the client has noiseless side information about some of the files, which means that $C^{(1)}$ is a noiseless test channel, then Z is uniformly distributed over $[K]\backslash\boldsymbol{\mathcal{M}}^{-1}(1)$, otherwise Z is uniformly distributed over $[K]$;
- a stochastic query function $\mathcal{F}_i : [K] \times \mathfrak{M} \times \boldsymbol{\mathcal{Y}}^n_{[K]} \to \boldsymbol{\mathcal{Q}}_i$, 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 $\mathbf{Q}_i \triangleq \mathcal{F}_i\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\right)$, 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., $\log|\boldsymbol{\mathcal{Q}}_i| = o(n)$, for $i \in [N]$;[2]
2) then, for all $i \in [N]$, Server i creates the answer $\mathbf{A}_i \triangleq \mathcal{E}_i\left(\mathbf{Q}_i, \mathbf{X}^n_{[K]}\right)$, where $\mathbf{X}^n_{[K]} \triangleq (X^n_i)_{i\in[K]}$, 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^n_Z 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 for the client is,
$$P_{\mathrm{e}} \triangleq \limsup_{n\to\infty} \mathbb{P}\left[\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\right) \neq X^n_Z\right]. \tag{2}$$
$R\left(\mathbf{Q}_{[N]}\right) \triangleq \sum_{i=1}^N R\left(\mathbf{Q}_i\right)$, where $\mathbf{Q}_{[N]} \triangleq (\mathbf{Q}_i)_{i\in[N]}$, is the normalized download cost 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 normalized download cost of the protocol as $R \triangleq \mathbb{E}_{\mathbf{Q}_{[N]}}[R\left(\mathbf{Q}_{[N]}\right)]$.

Example 1 (When $K = D = 2$, $T = 1$, and $d_1 = d_2 = 1$): Let X^n_1 and X^n_2 be the two files at the server and $Y^n_{i,\boldsymbol{\mathcal{M}}(i)}$ be the side information about X^n_i, $i \in \{1, 2\}$, 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^n_i. Note that $\boldsymbol{\mathcal{M}}$ can take two values (with the notation introduced in Section I-D):
$$\mathbf{M}_1 \triangleq \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{M}_2 \triangleq \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$
When $Z = 1$, since there are two different possibilities for the side information about X^n_1, that are $Y^n_{1,1}$ and $Y^n_{1,2}$, we define,
$$P^{(1)}_{\mathrm{e}}\left(Z = 1, \mathbf{M}_1\right) \triangleq \mathbb{P}\left[\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, Y^n_{1,1}, Y^n_{2,2}\right) \neq X^n_1 \middle| Z = 1, \boldsymbol{\mathcal{M}} = \mathbf{M}_1\right],$$
$$P^{(2)}_{\mathrm{e}}\left(Z = 1, \mathbf{M}_2\right) \triangleq \mathbb{P}\left[\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, Y^n_{1,2}, Y^n_{2,1}\right) \neq X^n_1 \middle| Z = 1, \boldsymbol{\mathcal{M}} = \mathbf{M}_2\right].$$
Similarly, when $Z = 2$, since there are two different possibilities for the side information about X^n_2 at the server, that are $Y^n_{2,1}$ and $Y^n_{2,2}$, we define,
$$P^{(3)}_{\mathrm{e}}\left(Z = 2, \mathbf{M}_1\right) \triangleq \mathbb{P}\left[\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, Y^n_{1,1}, Y^n_{2,2}\right) \neq X^n_2 \middle| Z = 2, \boldsymbol{\mathcal{M}} = \mathbf{M}_1\right],$$
$$P^{(4)}_{\mathrm{e}}\left(Z = 2, \mathbf{M}_2\right) \triangleq \mathbb{P}\left[\mathcal{D}\left(Z, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, Y^n_{1,2}, Y^n_{2,1}\right) \neq X^n_2 \middle| Z = 2, \boldsymbol{\mathcal{M}} = \mathbf{M}_2\right].$$

[2]When $D = 1$ and the test channel is a 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 [9] and PIR with noiseless side information in [17], respectively, it is shown in [9] and [17] that there is no loss of generality by making this assumption. In general, allowing the query cost to be non-negligible with the file length n is a different problem. However, similar to [16, Remark 1] and [24], 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 $(Z, \boldsymbol{\mathcal{M}})$.

Therefore, the probability of error in (2) is equal to,
$$\mathbb{P}[Z = 1, \boldsymbol{\mathcal{M}} = \mathbf{M}_1] P^{(1)}_{\mathrm{e}}\left(Z = 1, \mathbf{M}_1\right) + \mathbb{P}[Z = 1, \boldsymbol{\mathcal{M}} = \mathbf{M}_2] P^{(2)}_{\mathrm{e}}\left(Z = 1, \mathbf{M}_2\right) + \mathbb{P}[Z = 2, \boldsymbol{\mathcal{M}} = \mathbf{M}_1] P^{(3)}_{\mathrm{e}}\left(Z = 2, \mathbf{M}_1\right) + \mathbb{P}[Z = 2, \boldsymbol{\mathcal{M}} = \mathbf{M}_2] P^{(4)}_{\mathrm{e}}\left(Z = 2, \mathbf{M}_2\right).$$

We consider two privacy metrics to study the problem defined above. For the first metric, we keep the index of the desired file Z and the mapping $\boldsymbol{\mathcal{M}}$ private from the servers, whereas, for the second metric, we allow the index $\boldsymbol{\mathcal{M}}(Z)$ to be revealed to the server through the queries. As discussed in Section II-B, these two privacy metrics recover, as special cases, several PIR settings previously studied in the literature.

Definition 2 ($\mathrm{C}_{\text{PIR-PNSI}}$ Optimal Normalized Download Cost): An expected normalized download cost $R \in \mathbb{R}_+$ is achievable with private noisy side information and undisclosed side information statistics of the desired file, when up to T servers may collude, if there exist PIR protocols such that, for any set $\mathcal{T} \subseteq [N]$ such that $|\mathcal{T}| = T$,
$$P_{\mathrm{e}} = 0, \tag{3a}$$
$$I\left(\mathbf{Q}_{\mathcal{T}}, \mathbf{A}_{\mathcal{T}}, \mathbf{X}^n_{[K]}; Z, \boldsymbol{\mathcal{M}}\right) = 0. \tag{3b}$$
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 infimum of all achievable normalized download costs is referred to as the PIR optimal normalized download cost with private noisy side information and undisclosed side information statistics of the desired file, and is denoted by $\mathrm{C}_{\text{PIR-PNSI}}$.

Definition 3 ($\mathrm{C}^*_{\text{PIR-PNSI}}$ Optimal Normalized Download Cost): An expected normalized download cost $R \in \mathbb{R}_+$ is achievable with private noisy side information and disclosed side information statistics of the desired file, when up to T servers may collude, if there exist PIR protocols such that, for any set $\mathcal{T} \subseteq [N]$ such that $|\mathcal{T}| = T$,
$$P_{\mathrm{e}} = 0, \tag{4a}$$
$$I\left(\mathbf{Q}_{\mathcal{T}}, \mathbf{A}_{\mathcal{T}}, \mathbf{X}^n_{[K]}; Z, \boldsymbol{\mathcal{M}} \middle| \boldsymbol{\mathcal{M}}(Z)\right) = 0. \tag{4b}$$
The privacy metric (4b) means that the client file choice Z and the mapping $\boldsymbol{\mathcal{M}}$ must be kept secret from any T colluding servers, but the noise statistics of the side information of the desired file available at the client, i.e., $\boldsymbol{\mathcal{M}}(Z)$, may be revealed to the servers.[3] This contrasts with privacy metric (3b) where the noise statistics of the side information of the desired file available at the client must be kept secret from the servers. The infimum of all achievable normalized download costs is referred to as the PIR optimal normalized download cost with private noisy side information and disclosed side information statistics of the desired file, and is denoted by $\mathrm{C}^*_{\text{PIR-PNSI}}$.

Note that the privacy constraint in Definition 2 implies the privacy constraint in Definition 3, i.e., (3b)⇒(4b), and we will show that revealing the index of the test channel

[3]For this privacy metric, we assume that Z is uniformly distributed over $[K]$ because if the client has access to the desired file in a noiseless manner, then according to (4b), the client can reveal this to the servers and as a result the normalized download cost will be equal to zero.

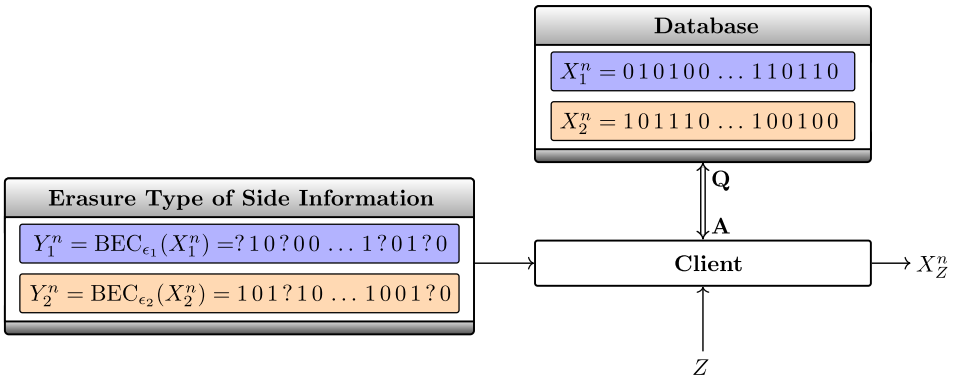

Fig. 2. Example with $(K, N, T, D) = (2, 1, 1, 2)$ when the test channels are BECs.

associated with the desired file X_Z^n can result in a strictly lower normalized download cost. When $D = 1$, the mapping $\boldsymbol{\mathcal{M}}$ is deterministic and therefore the secrecy constraints in (3b) and (4b) are both equal to

$$I\big(\mathbf{Q}_{\mathcal{T}}, \mathbf{A}_{\mathcal{T}}, \mathbf{X}_{[K]}^n; Z\big) = 0. \tag{5}$$

Remark 1 (Rate definition): Note that in all the previous PIR settings, as reviewed in the introduction, the normalized download cost is not a random variable. However, in our setting, since the privacy condition in Definition 3 depends on the index of the desired file, the normalized download cost is potentially a random variable. Specifically, we allow the query $\mathbf{Q}_{[N]}$ to depend on $\boldsymbol{\mathcal{M}}(Z)$, which leads to the normalized download cost $R\left(\mathbf{Q}_{[N]}\right)$ to be a random variable.

B. Examples

In Example 2, Example 3, and Example 4, we show that our problem setup recovers the problem setup for PIR with colluding servers [9], PIR with colluding servers and noiseless side information [16], [17], and PIR with private side information under storage constraints [24].

Example 2 (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 [9], and the privacy constraints in Definition 2 and Definition 3 reduce to (5), which is equivalent to the privacy constraint in [9].

Example 3 (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 [16, Theorem 2] for non-colluding servers and in [17, Theorem 1] for colluding servers, with the privacy constraint in Definition 2.

Example 4 (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 [24]. The difference with [24] is that the positions of the erasures are known at the servers in [24], whereas in our setting, the positions of the erasures are random and unknown at the servers. Therefore, the optimal normalized download cost for our problem setup in this example might be higher than the normalized download cost in [24]. However, we will show in the next section that the same normalized download cost as in [24] is achievable.

III. Main Results

The novel element of the achievability scheme in this paper is the redundancy removal based on the noisy side information. Before we present our main results, we provide a toy example to illustrate the main ideas of the achievability scheme.

A. Example With $(K, N, T, D) = (2, 1, 1, 2)$

In this example, as illustrated in Fig. 2, we assume that the files are binary sequences, and the test channels are BECs with parameters ϵ_1 and ϵ_2, where $\epsilon_1 < \epsilon_2$. Therefore, the mapping $\boldsymbol{\mathcal{M}}$ is a random mapping that maps the first file to one of the test channels and maps the other file to the other test channel. Hence, $\boldsymbol{\mathcal{M}}$ is a random permutation of the set $\{1, 2\}$. Here, as seen in Fig. 3, we first generate the source codes of the files in the database, assuming that the side information of all the files at the client is according to a BEC with parameter ϵ_1, by using the Slepian-Wolf encoder [29] [30, Section 10.4]. We refer to these source codes as $\mathbf{SC}_1$. Similarly, we generate the source codes of the files in the database, assuming that the side information of all the files at the client is according to a BEC with parameter $\epsilon_2 - \epsilon_1$, by using a second Slepian-Wolf encoder that is nested with the first Slepian-Wolf encoder, and refer to these source codes as $\mathbf{SC}_2$. For the privacy metric defined in Definition 2, i.e., when the client does not reveal the index of the test channel associated with the desired file, the client first downloads $\mathbf{SC}_1$, which results in retrieving the file that is associated with the BEC with parameter ϵ_1 and also gaining some information about the other file. The normalized download cost for downloading $\mathbf{SC}_1$ is $2\epsilon_1$ [8]. According to the Slepian-Wolf Theorem, e.g., [29] and [30], the probability of error for retrieving the file that is associated with the BEC with parameter ϵ_1 at the client goes to zero as n goes to infinity since the source coding rate is ϵ_1. Next, since the client has noiseless access to the file that is associated with the BEC with parameter ϵ_1, and therefore to the source coded version of this file, it then suffices to download $\mathbf{SC}_2(1) \oplus \mathbf{SC}_2(2)$,

where $\mathbf{SC}_2(i)$, for $i \in \{1,2\}$, is the source code of File i, and $\oplus$ denote the modulo 2 addition. Therefore, the normalized download cost of this operation is $\epsilon_2 - \epsilon_1$ [16, Theorem 2], [17, Therorem 1]. The probability of error for decoding the file that is associated with the BEC with parameter ϵ_2 at the client goes to zero as n goes to infinity since the source coding rate for this file is $\epsilon_1 + (\epsilon_2 - \epsilon_1) = \epsilon_2$. The total normalized download cost for this privacy metric is $2\epsilon_1+\epsilon_2-\epsilon_1 = \epsilon_1+\epsilon_2$.

Consider now the privacy metric defined in Definition 3, i.e., when the client is willing to reveal the index of the test channel associated with the desired file. When $Z = 1$, the client only downloads $\mathbf{SC}_1$, therefore, the normalized download cost is $2\epsilon_1$. When $Z = 2$, the client downloads $\mathbf{SC}_1$ and retrieves the file that is associated with the BEC with parameter ϵ_1, then it retrieves the desired file by downloading $\mathbf{SC}_2(1)\oplus\mathbf{SC}_2(2)$. Therefore, the normalized download cost is $2\epsilon_1+\epsilon_2-\epsilon_1 = \epsilon_1+\epsilon_2$. Since Z is a random variable with uniform distribution, the average normalized download cost is $\epsilon_1+\frac{1}{2}(\epsilon_1+\epsilon_2)$. Therefore, the normalized download cost when the client is willing to reveal the index of the test channel associated with the desired file, i.e., Definition 3, is $\frac{1}{2}(\epsilon_2 - \epsilon_1)$ less than the normalized download cost when the client does not reveal the index of the test channel associated with the desired file, i.e., Definition 2. However, this reduced normalized download cost comes at a price, since the privacy metric in Definition 2 is stronger than the privacy metric in Definition 3.

B. Main Results

We now state our main results 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 optimal normalized download cost of PIR with private noisy side information and undisclosed side information statistics of the desired file is

$$\begin{aligned}\mathrm{C}_{\text{PIR-PNSI}} &= \sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \times \\ &\qquad \left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{d_\ell-1}\right) \\ &= \sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \Psi^{-1}\left(\frac{T}{N}, d_\ell\right),\end{aligned} \tag{6}$$

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 $i \le j$, and $d_{[i:j]} \triangleq 0$, when $i > j$.

Proof: The achievability proof is based on a multilevel nested random binning scheme, that allows the client to use the side information efficiently, and the achievability schemes in [9] and [17]. Without loss of generality, we assume that the channels are ordered according to their noise level, i.e., we assume $H(U|V_i) \le H(U|V_j)$, for $i,j \in [D]$ such that $i \le j$, 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. In our scheme, for each test channel $C^{(\ell)}$, $\ell \in [D]$, the servers store the source coded version of all the files in a new database, denoted by $\mathbf{SC}_\ell$. These databases are obtained by applying a multilevel nested random binning scheme that performs source coding with side information. The client first downloads $\mathbf{SC}_1$, which results in retrieving the d_1 files that are associated with the first test channel and obtaining some information about the other files. To download $\mathbf{SC}_2$ the client uses the d_1 files that have been retrieved from the previous level as noiseless side information similar to [17]. Downloading $\mathbf{SC}_2$ results in retrieving the d_2 files that are associated with the second test channel and obtaining some information about the other files. The client continues this process to download all the $(\mathbf{SC}_1,\ldots,\mathbf{SC}_D)$ new databases and, in each step, uses all the files retrieved from the previous steps as noiseless side information. When the client is willing to reveal the index of the test channel that is associated with the desired file, which is denoted by i, it only downloads $(\mathbf{SC}_1,\ldots,\mathbf{SC}_i)$. Our converse proof shows that this scheme is optimal. The details of the achievability proof are available in Section IV-B. The converse proof is presented in Section IV-A. □

Remark 2 (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 optimal normalized download cost of PIR with private noisy side information and undisclosed side information statistics of the desired file is

$$\begin{aligned}\mathrm{C}_{\text{PIR-PNSI}} = \sum_{\ell=1}^{D} \epsilon_\ell \left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \times \\ \left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{d_\ell-1}\right).\end{aligned}$$

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 optimal normalized download cost result in Corollary 1 reduces to [9, 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 $D = 2$, $T = 1$, $\epsilon_1 = 0$, 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 $d_2 = K-d_1$ files, then the optimal normalized download cost result in Corollary 1 reduces to [16, 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 (Erasure Side Information With $D = 1$): In Corollary 1, if we set $D = 1$ and the test channel to be a

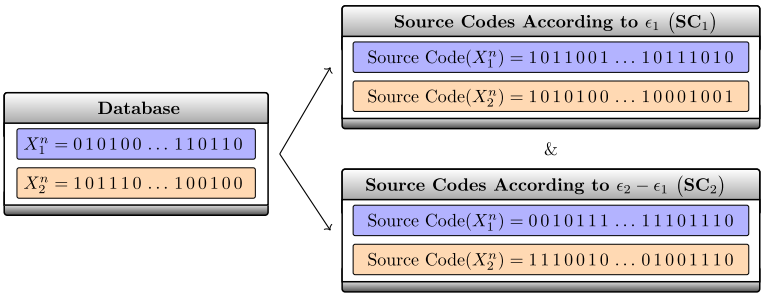

Fig. 3. Source codes of the files when the side information available at the client is according to the BEC with parameter ϵ_1, i.e., $\mathbf{SC}_1$, and source codes of the files when the side information available at the client is according to the BEC with parameter $\epsilon_2 - \epsilon_1$, i.e., $\mathbf{SC}_2$. Note that the source codes considered are nested.

BEC with parameter ϵ, then the result in Corollary 1 reduces to

$$\mathrm{C}_{\text{PIR-PNSI}} = \epsilon \left(1 + \frac{T}{N} + \left(\frac{T}{N}\right)^2 + \cdots + \left(\frac{T}{N}\right)^{K-1}\right). \quad (7)$$

The optimal result in Example 7 is equal to the optimal normalized download cost of the PIR problem when the side information is known by all the terminals.

Example 8 (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 optimal normalized download cost 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 [24, Theorem 1], since in [24] it is assumed that the client knows the *first* r_i bits, for $i \in [M]$, of M randomly selected files, however, our result in Corollary 1 reduces to the same optimal download cost as the optimal download cost derived in [24] by removing the constraint that the client knows the *first* r_i bits of M files and assumes that the client knows *any randomly chosen* r_i bits of M files.

Theorem 2: Consider K files, N replicated servers, where up to T of them may collude, and D test channels $\left(C^{(i)}\right)_{i\in[D]}$ as in Definition 1. Then, the optimal normalized download cost of PIR with private noisy side information and disclosed side information statistics of the desired file is

$$\begin{aligned}
\mathrm{C}^*_{\text{PIR-PNSI}} &= \mathbb{E}_U[R(U)] \\
&= \frac{1}{K}\sum_{\ell=1}^{D} H\left(X_1|Y_{1,\ell}\right)\left[d_{[\ell+1:D]} \sum_{j=d_{[\ell+1:D]}}^{-1+d_\ell+d_{[\ell+1:D]}} \left(\frac{T}{N}\right)^j + d_\ell \sum_{j=0}^{-1+d_{[\ell:D]}} \left(\frac{T}{N}\right)^j\right] \\
&= \frac{1}{K}\sum_{\ell=1}^{D} H\left(X_1|Y_{1,\ell}\right)\left[d_{[\ell+1:D]}\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \Psi^{-1}\left(\frac{T}{N}, d_\ell\right) + d_\ell \Psi^{-1}\left(\frac{T}{N}, d_{[\ell:D]}\right)\right], \quad (8a)
\end{aligned}$$

where

$$R(U) \triangleq \sum_{\ell=1}^{U-1} H\left(X_1|Y_{1,\ell}\right)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \Psi^{-1}\left(\frac{T}{N}, d_\ell\right) + H\left(X_1|Y_{1,U}\right)\Psi^{-1}\left(\frac{T}{N}, d_{[U:D]}\right), \quad (8b)$$

with U distributed according to $\mathbb{P}[U = u] \triangleq \frac{d_u}{K}$, for $u \in [D]$, $\Psi^{-1}(A, B) \triangleq \left(1 + A + A^2 + \cdots + A^{B-1}\right)$, 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: Similar to the achievability scheme of Theorem 1, the achievability scheme of Theorem 2 is based on source coding with side information, and the achievability schemes in [9] and [17]. Specifically, we use the same achievability scheme as in Theorem 1 by using $\boldsymbol{\mathcal{M}}(Z)$, instead of D, nested random bin indices for each file. The details of the proof are available in Section V-B. The converse proof is presented in Section V-A. □

The optimal results in Theorem 1 and Theorem 2 show that the optimal normalized download cost grows linearly with $H\left(X_1|Y_{1,\ell}\right)$, for $\ell \in [D]$, which quantifies how noisy the side information is. This confirms the intuition that the noisier the side information is, the higher the normalized download cost will become. Note that, the same remark as Remark 2 also applies to Theorem 2.

Corollary 2 (Binary Erasure Test Channels): Consider K files and N replicated servers, where up to T of them may collude. Additionally, assume that 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 optimal normalized download cost of PIR with private noisy side information and disclosed side information statistics of the desired file is

$$\begin{aligned}
\mathrm{C}^*_{\text{PIR-PNSI}} &= \mathbb{E}_U[R(U)] \\
&= \frac{1}{K}\sum_{\ell=1}^{D} \epsilon_\ell \left[d_{[\ell+1:D]}\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \Psi^{-1}\left(\frac{T}{N}, d_\ell\right) + d_\ell \Psi^{-1}\left(\frac{T}{N}, d_{[\ell:D]}\right)\right], \quad (9a) \\
R(U) &\triangleq \sum_{\ell=1}^{U-1} \epsilon_\ell \left(\frac{T}{N}\right)^{d_{[\ell+1:D]}} \Psi^{-1}\left(\frac{T}{N}, d_\ell\right) + \epsilon_U \Psi^{-1}\left(\frac{T}{N}, d_{[U:D]}\right). \quad (9b)
\end{aligned}$$

Corollary 3 (Binary Symmetric Channels): When the test channels $C^{(\ell)}$, for $\ell \in [D]$, are BSCs with parameters $0 \leq$

$p_1 < p_2 < \cdots < p_D \leq \frac{1}{2}$, then the optimal normalized download cost of the PIR problem with private noisy side information and disclosed side information statistics of the desired file is

$$\begin{aligned}
&\mathrm{C}^*_{\text{PIR-PNSI}} \\
&= \frac{1}{K}\sum_{\ell=1}^{D} H(p_\ell)\left[d_{[\ell+1:D]}\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right)+\right. \\
&\qquad \left. d_\ell\Psi^{-1}\left(\frac{T}{N},d_{[\ell:D]}\right)\right], \qquad (10)
\end{aligned}$$

where $H(p_\ell) \triangleq -p_\ell\log(p_\ell)-(1-p_\ell)\log(1-p_\ell)$.

Example 9 (Private Noiseless Side Information): In Corollary 2, if we set $D=2$, $\epsilon_1=0$, which means that the client knows d_1 files as side information in a noiseless manner, $\epsilon_2=1$, which means that there is no side information about d_2 files, and $\boldsymbol{\mathcal{M}}(Z)=1$, which means that the intended file is included in the noiseless side information, then the normalized download cost $R(U)$ in (9b) is equal to zero. When $\boldsymbol{\mathcal{M}}(Z)=2$, which means that the intended file is not included in the noiseless side information, then the optimal normalized download cost $R(U)$ in (9b) reduces to,

$$R(U)=\left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{d_2-1}\right),$$

which is the result in [16, Theorem 2], with $T=1$, and [17, Theorem 1]. Additionally, on average over the file choice, the expected normalized download cost is $\mathrm{C}^*_{\text{PIR-PNSI}}=\mathbb{E}_U[R(U)]$.

Example 10 (When $D=1$): When $D=1$, $U=1$, the first term on the Right Hand Side (RHS) of (8b) is equal to zero, and the optimal normalized download cost in (8a) reduces to Theorem 1 when $D=1$, that is,

$$\begin{aligned}
&\mathrm{C}^*_{\text{PIR-PNSI}}=\mathrm{C}_{\text{PIR-PNSI}} \\
&=H\big(X_1|Y_{1,1}\big)\left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{K-1}\right).
\end{aligned}$$

Remark 3 (Comparing the results in Theorem 1 and Theorem 2): We rewrite the result in Theorem 1 for any $U\in[D]$ as follows,

$$\begin{aligned}
&\mathrm{C}_{\text{PIR-PNSI}}=\sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right) \\
&=\sum_{\ell=1}^{U-1} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right) \\
&\quad+\sum_{\ell=U}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right) \\
&\overset{(a)}{=} R(U)-H(X_1|Y_{1,U})\Psi^{-1}\left(\frac{T}{N},d_{[U:D]}\right)+ \\
&\qquad\sum_{\ell=U}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right) \\
&\overset{(b)}{=} R(U)+\sum_{\ell=U}^{D}\Big(H\big(X_1|Y_{1,\ell}\big)-H\big(X_1|Y_{1,U}\big)\Big)\times \\
&\qquad\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N},d_\ell\right) \\
&\overset{(c)}{\geq} R(U), \qquad (11)
\end{aligned}$$

where

- (a) follows from (8b);
- (b) follows by expanding $\Psi^{-1}\left(\frac{T}{N},d_{[U:D]}\right)$;
- (c) follows since $H\big(X_1|Y_{1,U}\big)\leq H\big(X_1|Y_{1,\ell}\big)$, for $\ell\in[U:D]$.

Therefore, the optimal normalized download cost in (8a), which is the average of $R(U)$, with respect to U, is always smaller than or equal to the optimal normalized download cost in Theorem 1, i.e., $\mathrm{C}^*_{\text{PIR-PNSI}}\leq\mathrm{C}_{\text{PIR-PNSI}}$. This shows that revealing the index of the test channel that is associated with the desired file reduces the normalized download cost.

Hence, the optimal normalized download cost in Theorem 2 is smaller than the optimal normalized download cost in Theorem 1, and the difference between these two quantities increases as the index $\boldsymbol{\mathcal{M}}(Z)$ of the test channel that is associated with the desired file decreases.

IV. Proof of Theorem 1

A. Converse Proof

Define $\mathbf{Z}\triangleq(Z,\bar{\mathbf{Z}})$, where $\bar{\mathbf{Z}}\triangleq\big(\bar{Z}_1,\bar{Z}_2,\ldots,\bar{Z}_{K-1}\big)$, and

$$\begin{cases}
\bar{\mathbf{Z}}_{[1+d_{[i-1]}:d_{[i]}]}\triangleq\boldsymbol{\mathcal{M}}^{-1}(i) & \text{if}\quad i<\boldsymbol{\mathcal{M}}(Z)\\
\bar{\mathbf{Z}}_{[1+d_{[i-1]}:-1+d_{[i]}]}\triangleq\boldsymbol{\mathcal{M}}^{-1}(i)\backslash\{Z\} & \text{if}\quad i=\boldsymbol{\mathcal{M}}(Z)\\
\bar{\mathbf{Z}}_{[d_{[i-1]}:-1+d_{[i]}]}\triangleq\boldsymbol{\mathcal{M}}^{-1}(i) & \text{if}\quad i>\boldsymbol{\mathcal{M}}(Z)
\end{cases}, \qquad (12a)$$

where $d_{[i]}\triangleq\sum_{j=1}^{i}d_i$, $\bar{\mathbf{Z}}_{[i:j]}\triangleq(\bar{Z}_i,\bar{Z}_{i+1},\ldots,\bar{Z}_j)$, and, by convention, for $a,b\in\mathbb{N}_*$ and $a>b$ define $\bar{\mathbf{Z}}_{[a:b]}\triangleq\emptyset$. Then, we index all the files as depicted in Fig. 4 such that the mapping $\boldsymbol{\mathcal{M}}$ can be described as (13), shown at the bottom of the page, (with the notation introduced in Section I-D). From (12a) and (13), the side information available at the client is (14), shown at the bottom of the page, in which $\mathbf{Y}^n_{\bar{\mathbf{Z}}_{[i:j]},\ell}\triangleq\big(Y^n_{\bar{Z}_i,\ell},Y^n_{\bar{Z}_{i+1},\ell},\ldots,Y^n_{\bar{Z}_j,\ell}\big)$.

$$\boldsymbol{\mathcal{M}}=\begin{pmatrix}\bar{\mathbf{Z}}_{[1:d_1]} & \ldots & \left(Z,\bar{\mathbf{Z}}_{[1+d_{[i-1]}:-1+d_{[i]}]}\right) & \ldots & \bar{\mathbf{Z}}_{[d_{[D-1]}:-1+d_{[D]}]}\\ (1,\ldots,1) & \ldots & (i,i,\ldots,i) & \ldots & (D,\ldots,D)\end{pmatrix}, \qquad (13)$$

$$\begin{aligned}
\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\triangleq\left(\mathbf{Y}^n_{\boldsymbol{\mathcal{M}}^{-1}(i),i}\right)_{i\in[D]}=\Bigg(&\mathbf{Y}^n_{\bar{\mathbf{Z}}_{[1:d_1]},1},\mathbf{Y}^n_{\bar{\mathbf{Z}}_{\left[1+d_1:d_{[2]}\right]},2},\ldots,\mathbf{Y}^n_{\bar{\mathbf{Z}}_{\left[1+d_{[i-2]}:d_{[i-1]}\right]},i-1},Y^n_{Z,i},\\
&\mathbf{Y}^n_{\bar{\mathbf{Z}}_{\left[1+d_{[i-1]}:-1+d_{[i]}\right]},i},\mathbf{Y}^n_{\bar{\mathbf{Z}}_{\left[d_{[i]}:-1+d_{[i+1]}\right]},i+1},\mathbf{Y}^n_{\bar{\mathbf{Z}}_{\left[d_{[i+1]}:-1+d_{[i+2]}\right]},i+2},\ldots,\mathbf{Y}^n_{\bar{\mathbf{Z}}_{\left[d_{[D-1]}:-1+d_{[D]}\right]},D}\Bigg).
\end{aligned} \qquad (14)$$

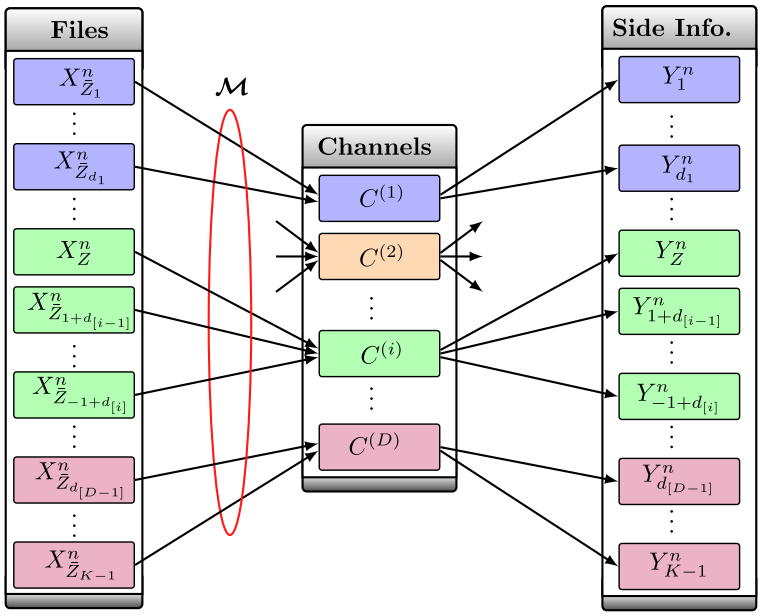

Fig. 4. Indexing the files based on the mapping $\mathcal{M}$.

Example 11: To illustrate the definition of $\bar{\mathbf{Z}} \triangleq (\bar{Z}_1, \ldots, \bar{Z}_{K-1})$ in (12a), consider a setting where $d_i = 1$, for $i \in [D]$, which means that $D = K$. In this case, $\sum_{t=1}^{i-1} d_t = i-1$ and $\sum_{t=1}^{i} d_t = i$, therefore, the definition of $\bar{Z}_i$, for $i \in [K-1]$, in (12a) reduces to

$$\begin{cases} \bar{Z}_i \triangleq \mathcal{M}^{-1}(i) & \text{if} \quad i < \mathcal{M}(Z) \\ \bar{Z}_{i-1} \triangleq \mathcal{M}^{-1}(i) & \text{if} \quad i > \mathcal{M}(Z) \end{cases}. \tag{15}$$

For example, let $K = D = 3$, $Z = 2$, and the side information at the client be $\left(Y^n_{1,3}, Y^n_{2,1}, Y^n_{3,2}\right)$, therefore $\mathcal{M}(Z) = 1$, $\bar{Z}_1 \triangleq \mathcal{M}^{-1}(2) = 3$, and $\bar{Z}_2 \triangleq \mathcal{M}^{-1}(3) = 1$.

The following equations and lemmas are essential for the converse proof. From the dependency graph in Fig. 5 we have

$$I\left(Z, \mathbf{Q}_{[N]}; \mathbf{X}^n_{[K]} \middle| \mathbf{Y}^n_{[K],\mathcal{M}}, \mathcal{M}\right) = 0. \tag{16}$$

Considering the probability of error in (2), by Fano's inequality [31, Section 2.11], we also have

$$\max_{z\in[K]} \max_{\mathbf{M}\in\mathfrak{M}} H\big(X^n_Z | \mathbf{Q}_{[N]}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\mathcal{M}}, Z = z, \mathcal{M} = \mathbf{M}\big) = o(n). \tag{17}$$

Lemma 1: For all $\mathbf{M} \in \mathfrak{M}$, $z \in [K]$, $\mathcal{T}' \subseteq [N]$, and $\mathcal{T} \subseteq [N]$ such that $|\mathcal{T}| = T$, we have

$$I\left(\mathbf{A}_\mathcal{T}; \mathbf{Q}_{[N]\setminus\mathcal{T}} \middle| \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, \mathbf{X}^n_{\mathcal{T}'}, Z = z, \mathcal{M} = \mathbf{M}\right) = 0. \tag{18}$$

Proof: We have,

$$\begin{aligned} &I\left(\mathbf{A}_\mathcal{T}; \mathbf{Q}_{[N]\setminus\mathcal{T}} \middle| \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, \mathbf{X}^n_{\mathcal{T}'}, Z = z, \mathcal{M} = \mathbf{M}\right) \\ &\overset{(a)}{\leq} I\left(\mathbf{A}_\mathcal{T}, \mathbf{X}^n_{[K]}; \mathbf{Q}_{[N]\setminus\mathcal{T}} \middle| \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, \mathbf{X}^n_{\mathcal{T}'}, Z = z, \mathcal{M} = \mathbf{M}\right) \\ &\overset{(b)}{=} I\left(\mathbf{X}^n_{[K]}; \mathbf{Q}_{[N]\setminus\mathcal{T}} \middle| \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, \mathbf{X}^n_{\mathcal{T}'}, Z = z, \mathcal{M} = \mathbf{M}\right) \end{aligned}$$

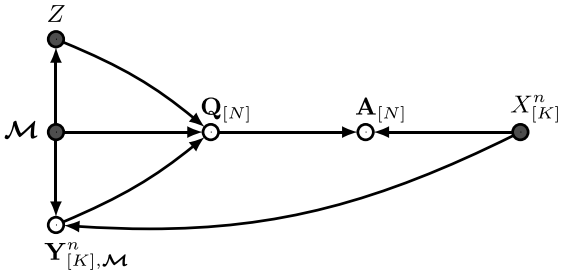

Fig. 5. Dependency graph for all the involved random variables.

$$+ I\left(\mathbf{A}_\mathcal{T}; \mathbf{Q}_{[N]\setminus\mathcal{T}} \middle| \mathbf{X}^n_{[K]}, \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, Z = z, \mathcal{M} = \mathbf{M}\right), \tag{19}$$

where (a) and (b) hold by the chain rule and non-negativity of the mutual information. The first term on the RHS of (19) is equal to zero because of (16) and the second the term on the RHS of (19) is also equal to zero from (1). □

Lemma 2: For each $\mathbf{M} \in \mathfrak{M}$, $z, z' \in [K]$, $\mathcal{T}' \subseteq [N]$, and $\mathcal{T} \subseteq [N]$ such that $|\mathcal{T}| = T$, we have

$$\begin{aligned} &H\left(\mathbf{A}_\mathcal{T} \middle| \mathbf{X}^n_{\mathcal{T}'}, \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, Z = z, \mathcal{M} = \mathbf{M}\right) \\ &\quad = H\left(\mathbf{A}_\mathcal{T} \middle| \mathbf{X}^n_{\mathcal{T}'}, \mathbf{Q}_\mathcal{T}, \mathbf{Y}^n_{[K],\mathcal{M}}, Z = z', \mathcal{M} = \mathbf{M}\right) - o(n). \end{aligned} \tag{20}$$

Proof: We have,

$$\begin{aligned} &I\left(\mathbf{Q}_{[N]}, \mathbf{A}_{[N]}, \mathbf{X}^n_{[K]}, \mathbf{Y}^n_{[K],\mathcal{M}}; Z \middle| \mathcal{M} = \mathbf{M}\right) \\ &\overset{(a)}{=} I\left(\mathbf{Q}_{[N]}, \mathbf{X}^n_{[K]}, \mathbf{Y}^n_{[K],\mathcal{M}}; Z \middle| \mathcal{M} = \mathbf{M}\right) \\ &= I\left(\mathbf{X}^n_{[K]}, \mathbf{Y}^n_{[K],\mathcal{M}}; Z \middle| \mathcal{M} = \mathbf{M}\right) + \\ &I\left(\mathbf{Q}_{[N]}; Z \middle| \mathbf{X}^n_{[K]}, \mathbf{Y}^n_{[K],\mathcal{M}}, \mathcal{M} = \mathbf{M}\right) \\ &\overset{(b)}{\leq} I\left(\mathbf{X}^n_{[K]}, \mathbf{Y}^n_{[K],\mathcal{M}}; Z \middle| \mathcal{M} = \mathbf{M}\right) + o(n) \end{aligned}$$

$$\overset{(c)}{=} I\left(\mathbf{X}^n_{[K]}; Z \middle| \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}} = \mathbf{M}\right) + o(n)$$
$$\overset{(d)}{=} o(n), \tag{21}$$

where

(a) follows from the chain rule and (1);

(b) follows since the queries are of negligible length compared to the file length n and, therefore, $I\left(\mathbf{Q}_{[N]}; Z \middle| \mathbf{X}^n_{[K]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}} = \mathbf{M}\right) \leq H\left(\mathbf{Q}_{[N]}\right) = o(n)$;

(c) follows from the chain rule and because

$$\begin{aligned} &H\left(\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}} \middle| \boldsymbol{\mathcal{M}} = \mathbf{M}, Z\right) \\ &= \sum_z \mathbb{P}[Z = z \big| \boldsymbol{\mathcal{M}} = \mathbf{M}] H\left(\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}} \middle| Z = z, \boldsymbol{\mathcal{M}} = \mathbf{M}\right) \\ &= \sum_z \mathbb{P}[Z = z \big| \boldsymbol{\mathcal{M}} = \mathbf{M}] H\left(\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\right) \\ &= H\left(\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\right), \end{aligned}$$

where the second equality follows since all the files are generated according to the same distribution and, thus, the entropy of the side information does not depend on $\mathbf{M}$ and z.

(d) follows from (16).

Then, we have

$$\begin{aligned} o(n) &\overset{(a)}{=} I\left(\mathbf{A}_{\mathcal{T}}, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \mathbf{X}^n_{\mathcal{T}'}; Z \middle| \boldsymbol{\mathcal{M}} = \mathbf{M}\right) \\ &\overset{(b)}{=} I\left(\mathbf{A}_{\mathcal{T}}; Z \middle| \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \mathbf{X}^n_{\mathcal{T}'}, \boldsymbol{\mathcal{M}} = \mathbf{M}\right) + o(n), \end{aligned} \tag{22}$$

where (a) holds by (21), and (b) holds by (21) and the chain rule. Finally, (22) implies (20). □

Then, we bound $nR\left(\mathbf{Q}_{[N]}\right)$ as (23d), shown at the bottom of the next page, where

(a) follows since conditioning does not increase the entropy;

(b) follows from Fano's inequality in (17);

(c) follows since for $i \triangleq \boldsymbol{\mathcal{M}}(Z)$, we have

$$\begin{aligned} &I\big(\mathbf{Q}_{[N]}, \mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}; X^n_Z \big| Y^n_{Z,i}, Z, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) \\ &= I\big(\mathbf{Q}_{[N]}; X^n_Z \big| Y^n_{Z,i}, Z, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) + \\ &\quad I\big(Y^n_{\bar{\mathbf{Z}},[D]}; X^n_Z \big| \mathbf{Q}_{[N]}, Y^n_{Z,i}, Z, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) \\ &= I\big(Y^n_{\bar{\mathbf{Z}},[D]}; X^n_Z \big| \mathbf{Q}_{[N]}, Y^n_{Z,i}, Z, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) \\ &= 0, \end{aligned}$$

where $\mathbf{Y}^n_{\bar{\mathbf{Z}},[D]} \triangleq \left(\mathbf{Y}^n_{\bar{Z}_1,[D]}, \mathbf{Y}^n_{\bar{Z}_2,[D]}, \ldots, \mathbf{Y}^n_{\bar{Z}_{K-1},[D]}\right)$ and $\mathbf{Y}^n_{\bar{Z}_i,[D]} \triangleq \left(Y^n_{\bar{Z}_i,1}, Y^n_{\bar{Z}_i,2}, \ldots, Y^n_{\bar{Z}_i,D}\right)$, for $i \in [K-1]$, and the second equality holds because, from Fig. 5, $\mathbf{Q}_{[N]} - (Y^n_{Z,i}, Z, \boldsymbol{\mathcal{M}}) - X^n_Z$ forms a Markov chain, and the last equality holds because, from Fig. 5, $Y^n_{\bar{\mathbf{Z}},[D]} - (\mathbf{Q}_{[N]}, Y^n_{Z,i}, Z, \boldsymbol{\mathcal{M}}) - X^n_Z$ forms a Markov chain;

(d) follows since for $i \in [K]$ and $j \in [D]$, $H(X^n_i|Y^n_{i,j}) = H(X^n_1|Y^n_{1,j}) = nH(X_1|Y_{1,j})$ because $P_{X^n_i} = P_{X^n_1} = P_X^{\otimes n}$ and, therefore, we have $H\left(X^n_z \middle| Y^n_{z,j}, Z = z, \boldsymbol{\mathcal{M}} = \mathbf{M}\right) = H(X^n_z|Y^n_{z,j}) = H(X^n_1|Y^n_{1,j}) = nH(X_1|Y_{1,j})$, for any z and $\mathbf{M}$;

(e) follows from Lemma 1;

(f) follows from Lemma 2;

(g) follows since one can lower bound the second term on the RHS of (23b) using the following inequality,

$$\begin{aligned} &H\left(\mathbf{A}_{[N]} \middle| X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = \bar{z}_{K-1}, \boldsymbol{\mathcal{M}} = \mathbf{M}\right) \\ &\geq \frac{1}{\binom{N}{T}} \sum_{\mathcal{T}:|\mathcal{T}|=T} H\Big(\mathbf{A}_{\mathcal{T}} \Big| X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \\ &\qquad Z = \bar{z}_{K-2}, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \qquad (24a) \\ &\geq \frac{T}{N} H\Big(\mathbf{A}_{[N]} \Big| X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \\ &\qquad Z = \bar{z}_{K-2}, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big), \qquad (24b) \end{aligned}$$

where (24a) follows by writing (23c) for all $\binom{N}{T}$ different subsets $\mathcal{T} \subset [N]$ with cardinality T and adding up all these inequalities; and (24b) follows from Han's inequality [31, Theorem 17.6.1].

Repeating the steps described in (23d) starting from (23a) with $Z = z'$, where z' changes from the first element till the last element of $\left[\bar{z}_{K-2}, \bar{z}_{K-3}, \ldots, \bar{z}_{d_{[D-1]}}\right]$, to bound the second entropy term on the RHS of (23d), we obtain (25), shown at the bottom of the next page, where (a) follows by induction and repeating the steps described in (23d) starting from (23a) with $Z = z'$, where z' changes from the first element till the last element of $\left[\bar{z}_{d_{[D-1]}-1}, \bar{z}_{d_{[D-1]}-2}, \ldots, \bar{z}_{1+d_{[i-1]}}, z, \bar{z}_{d_{[i-1]}}, \ldots, \bar{z}_1\right]$, where $i \triangleq \mathbf{M}(z)$, and (b) follows from (1).

B. Achievability Proof

A high-level description of the achievability scheme is provided after Theorem 1.

1) Preliminaries: Our achievability is based on nested source coding, which we define first and then use in our achievability proof as a black box. Consider a discrete memoryless source $(\mathcal{X}_1 \times \bigtimes_{\ell \in [D]} \mathcal{Y}_{1,\ell}, P_{X_1, \mathbf{Y}_{1,[D]}})$ with $D+1$ components. Assume that $(X^n_1, \mathbf{Y}^n_{1,[D]})$ are independent and identically distributed (i.i.d.) samples of this source. Then, consider an encoder $\mathcal{E} : \mathcal{X}^n_1 \to \boldsymbol{\mathcal{J}}^{(1)}_{[D]}$, that maps the sequence X^n_1 to $\mathbf{J}^{(1)}_{[D]} \triangleq (J^{(1)}_\ell)_{\ell \in [D]}$, where the asymptotic rate of $J^{(1)}_\ell$, for $\ell \in [D]$, is $H(X_1|Y_{1,\ell}) - H(X_1|Y_{1,\ell-1})$, with the convention $H(X_1|Y_{1,0}) = 0$. Consider also D decoders $\mathcal{D}_\ell : \boldsymbol{\mathcal{J}}^{(1)}_{[\ell]} \times \mathcal{Y}^n_{1,\ell} \to \mathcal{X}^n_1$, for $\ell \in [D]$, where the Decoder $\mathcal{D}_\ell$ assigns an estimate $\hat{X}^n_1$ to $(\mathbf{J}^{(1)}_{[\ell]}, Y^n_{1,\ell})$ such that $\mathbb{P}\big[\hat{X}^n_1 \neq X^n_1\big] \xrightarrow[n\to\infty]{} 0$. In Appendix A, we explain how to obtain such a scheme with nested random binning and how to implement it with nested polar codes when the side information at the decoders forms a Markov chain.

Assume that each file is of length $n = N^K$, with symbols in a sufficiently large finite field $\mathbb{F}_q$. Fix $\delta > 0$.

2) Nested Source Coding: For every file x^n_i, $i \in [K]$, generate D nested source codes as in Section IV-B.1. For each test channel $\ell \in [D]$, we denote the source code of the file x^n_i, $i \in [K]$, by $j^{(i)}_\ell \in \mathcal{J}_\ell \triangleq \left[q^n_\ell\right]$, where $q_\ell \triangleq q^{R_\ell}$. Here,

$R_0 \triangleq 0$ and for $t \in [K]$,

$$\sum_{i=1}^{\ell} R_i = H\big(X_t|Y_{t,\ell}\big) + \delta. \tag{26}$$

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

3) Query Structure Construction: The client constructs the query in D different levels. In the first level, we apply to the database $\mathbf{SC}_1$ the same query structure as in [9], which consists of K sublevels. In the level $\ell \in [2:D]$, we apply to the database $\mathbf{SC}_\ell$ the same query structure as in [17], which also consists of K sublevels. Specifically, as in [17], the k_ℓ^{th} sublevel consists of sums of k_ℓ symbols, which are called k_ℓ-sums. There are $\binom{K}{k_\ell}$ different types of k_ℓ-sums

$$\begin{aligned}
&nR\left(\mathbf{Q}_{[N]}\right)\\
&\geq H\left(\mathbf{A}_{[N]}\right)\\
&\overset{(a)}{\geq} H\Big(\mathbf{A}_{[N]}\big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big)\\
&= H\Big(\mathbf{A}_{[N]}, X^n_{\bar{z}_{K-1}}\big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big)\\
&\quad - H\Big(X^n_{\bar{z}_{K-1}}\big|\mathbf{Q}_{[N]}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) \qquad (23a)\\
&\overset{(b)}{\geq} H\Big(\mathbf{A}_{[N]}, X^n_{\bar{z}_{K-1}}\big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n)\\
&= H\Big(X^n_{\bar{z}_{K-1}}\big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big)\\
&\quad + H\Big(\mathbf{A}_{[N]}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n)\\
&\overset{(c)}{=} H\Big(X^n_{\bar{z}_{K-1}}\big|Y^n_{\bar{z}_{K-1},D}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big)\\
&\quad + H\Big(\mathbf{A}_{[N]}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n)\\
&\overset{(d)}{=} nH\Big(X_1\big|Y_{1,D}\Big) + H\Big(\mathbf{A}_{[N]}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n) \qquad (23b)\\
&\geq nH\Big(X_1\big|Y_{1,D}\Big) + H\Big(\mathbf{A}_{\mathcal{T}}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n)\\
&\overset{(e)}{=} nH\Big(X_1\big|Y_{1,D}\Big) + H\Big(\mathbf{A}_{\mathcal{T}}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n)\\
&\overset{(f)}{=} nH\Big(X_1\big|Y_{1,D}\Big) + H\Big(\mathbf{A}_{\mathcal{T}}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-2}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n)\\
&\geq nH\Big(X_1\big|Y_{1,D}\Big) + H\Big(\mathbf{A}_{\mathcal{T}}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-2}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n) \qquad (23c)\\
&\overset{(g)}{\geq} nH\Big(X_1\big|Y_{1,D}\Big) + \frac{T}{N} H\Big(\mathbf{A}_{[N]}\big|X^n_{\bar{z}_{K-1}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-2}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(n), \qquad (23d)
\end{aligned}$$

$$\begin{aligned}
&R\left(\mathbf{Q}_{[N]}\right)\\
&\geq H\big(X_1|Y_{1,D}\big) + \frac{T}{N}\Bigg[H\big(X_1|Y_{1,D}\big) + \frac{T}{N}\Bigg[H\big(X_1|Y_{1,D}\big) + \cdots + \frac{T}{N}\Big[H\big(X_1|Y_{1,D}\big)\\
&\qquad + \frac{1}{n} H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\mathbf{Z}_{\left[d_{[D-1]}:K-1\right]}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{d_{[D-1]}-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big)\Big]\Bigg]\Bigg] - o(1)\\
&= H\big(X_1|Y_{1,D}\big)\left[1 + \frac{T}{N} + \left(\frac{T}{N}\right)^2 + \cdots + \left(\frac{T}{N}\right)^{d_D-1}\right]\\
&\qquad + \frac{1}{n}\left(\frac{T}{N}\right)^{d_D} H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\mathbf{Z}_{\left[d_{[D-1]}:K-1\right]}}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{d_{[D-1]}-1}, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(1)\\
&\overset{(a)}{\geq} \sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{\sum_{i=\ell+1}^{D} d_i}\left[1 + \frac{T}{N} + \left(\frac{T}{N}\right)^2 + \cdots + \left(\frac{T}{N}\right)^{d_\ell-1}\right]\\
&\qquad + \frac{1}{n}\left(\frac{T}{N}\right)^{K-1} H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{[K]}, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_1, \boldsymbol{\mathcal{M}}=\mathbf{M}\Big) - o(1)\\
&\overset{(b)}{=} \sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big)\left(\frac{T}{N}\right)^{\sum_{i=\ell+1}^{D} d_i}\left[1 + \frac{T}{N} + \left(\frac{T}{N}\right)^2 + \cdots + \left(\frac{T}{N}\right)^{d_\ell-1}\right] - o(1) \qquad (25)
\end{aligned}$$

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 $\sum_{k_\ell=1}^{K}\binom{K}{k_\ell}(N-T)^{k_\ell-1}T^{K-k_\ell}$.

4) Query Specialization: For $\ell\in[D]$, we do the query structure construction and query specialization without considering the availability of any side information as in [17], and denote this scheme by Π_ℓ. Then, we do query redundancy removal based on the availability of noiseless side information similar to [17]. 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 are decoded in the previous levels, the client knows $\mathbf{X}^n_{[d_{[\ell]}]}$ and, therefore, $\left(j_{\ell+1}^{(1)},\dots,j_{\ell+1}^{(d_{[\ell]})}\right)$, which is used as noiseless side information to recover $\left(j_{\ell+1}^{(d_{[\ell]}+1)},\dots,j_{\ell+1}^{(K)}\right)$ 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 $\mathbf{U}_{\ell,j}\in\mathbb{F}_q^{p_{\ell,1}}$ 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 [32], 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 $\mathbf{SC}_1$, $p_{1,2}=0$.

5) Decoding: For $\ell\in[D]$, after reconstructing $(j_i^{(t)})_{i\in[\ell]}$, for $t\in\boldsymbol{\mathcal{M}}^{-1}(\ell)$, given $\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}$, the client forms $\hat{X}_t^n$, an estimate of the sequence X_t^n by using the nested source decoders with (26), and thus $\mathbb{P}\big[\hat{X}_t^n\neq X_t^n\big]\xrightarrow[n\to\infty]{}0$.

6) Rate Calculation: Similar to [17], 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, we have,

$$\begin{aligned}p_{\ell,1}&=\sum_{k_\ell=1}^{K}\binom{K}{k_\ell}(N-T)^{k_\ell-1}T^{K-k_\ell}\\&=\frac{\sum_{k_\ell=0}^{K}\binom{K}{k_\ell}(N-T)^{k_\ell}T^{K-k_\ell}-T^K}{N-T}\\&=\frac{N^K-T^K}{N-T},\end{aligned}\tag{27a}$$

similarly,

$$\begin{aligned}p_{\ell,2}&=\sum_{k_\ell=1}^{d_{[\ell-1]}}\binom{d_{[\ell-1]}}{k_\ell}(N-T)^{k_\ell-1}T^{K-k_\ell}\\&=T^{K-d_{[\ell-1]}}\sum_{k_\ell=1}^{d_{[\ell-1]}}\binom{d_{[\ell-1]}}{k_\ell}(N-T)^{k_\ell-1}T^{d_{[\ell-1]}-k_\ell}\\&=\frac{T^{K-d_{[\ell-1]}}\left(N^{d_{[\ell-1]}}-T^{d_{[\ell-1]}}\right)}{N-T}.\end{aligned}\tag{27b}$$

Therefore, the normalized download cost for the level ℓ is,

$$\begin{aligned}R^{(\ell)}&=\frac{R_\ell N(p_{\ell,1}-p_{\ell,2})}{n}\\&\overset{(a)}{=}\frac{R_\ell N(p_{\ell,1}-p_{\ell,2})}{N^K}\\&\overset{(b)}{=}\frac{R_\ell\left(1-\left(\frac{T}{N}\right)^{K-d_{[\ell-1]}}\right)}{\left(1-\frac{T}{N}\right)}\\&\overset{(c)}{=}\Big(H\big(X_1|Y_{1,\ell}\big)-H\big(X_1|Y_{1,\ell-1}\big)\Big)\sum_{i=0}^{K-d_{[\ell-1]}-1}\left(\frac{T}{N}\right)^i,\end{aligned}\tag{28}$$

where

(a) follows since $n=N^K$;
(b) follows from (27);
(c) follows from (26).

Therefore, the total normalized download cost is,

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

7) 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 obtains after each level. The decoding 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 [9] and [17].

V. Proof of Theorem 2

A. Converse Proof

The following equations and lemma are essential for the converse proof. Considering the probability of error in (2), by Fano's inequality [31, Section 2.11], for every $i\in[D]$, we have

$$\max_{z\in[K]}\max_{\mathbf{M}\in\mathfrak{M}}H\big(X_Z^n|\mathbf{Q}_{[N]},\mathbf{A}_{[N]},\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}},\boldsymbol{\mathcal{M}}(Z)=i,Z=z,\boldsymbol{\mathcal{M}}=\mathbf{M}\big)=o(n).\tag{29}$$

Lemma 3: For all $z,z'\in[K]$, $i\in[D]$, $\mathcal{T},\mathcal{T}'\subseteq[N]$, and $\mathbf{M},\mathbf{M}'\in\mathfrak{M}$, such that $\mathbf{M}(z)=\mathbf{M}'(z')$,

$$H\left(\mathbf{A}_{\mathcal{T}}\middle|\mathbf{Q}_{\mathcal{T}},\mathbf{X}^n_{\mathcal{T}'},\mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}},Z=z,\boldsymbol{\mathcal{M}}(Z)=i,\boldsymbol{\mathcal{M}}=\mathbf{M}\right)$$

$$= H\Big(\mathbf{A}_{\mathcal{T}}\Big|\mathbf{Q}_{\mathcal{T}}, \mathbf{X}^n_{\mathcal{T}'}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z', \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}'\Big) - o(n). \tag{30}$$

Proof: Since $H(Z) = \log K$ and $H(\boldsymbol{\mathcal{M}}) \leq \log|\mathfrak{M}|$, then for all $i \in [D]$, and $\mathcal{T}, \mathcal{T}' \subseteq [N]$, we have

$$I\left(\mathbf{A}_{\mathcal{T}}; Z, \boldsymbol{\mathcal{M}}\middle|\mathbf{Q}_{\mathcal{T}}, \mathbf{X}^n_{\mathcal{T}'}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}(Z) = i\right) = o(n). \tag{31}$$

□

Consider $(z, \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_{K-1})$ a realization of $\mathbf{Z}$, and $\mathbf{M}$ a realization of $\boldsymbol{\mathcal{M}}$ such that $\mathbf{M}(z) = i$. Then, we bound $nR(\mathbf{Q}_{[N]})$ as (32c), shown at the bottom of the page, where

(a) follows since conditioning does not increase entropy;

(b) follows from Fano's inequality in (29);

(c) follows since, for $i \triangleq \boldsymbol{\mathcal{M}}(Z)$, we have,

$$\begin{aligned} &I\big(\mathbf{Q}_{[N]}, \mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}; X^n_Z\big|Y^n_{Z,i}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) \\ &= I\big(\mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}; X^n_Z\big|Y^n_{Z,i}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) \\ &\quad + I\big(\mathbf{Q}_{[N]}; X^n_Z\big|\mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}, Y^n_{Z,i}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\big) \qquad (33a) \\ &\leq o(n), \qquad (33b) \end{aligned}$$

where (33b) holds because the first term on the RHS of (33a) is equal to zero since the files are independent of one another, and the second term on the RHS of (33a) is less than or equal to $o(n)$ since the queries are of negligible normalized download cost;

(d) follows because $H\Big(X^n_Z\Big|Y^n_{Z,i}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) = H(X^n_z|Y^n_{z,i}) = H(X^n_1|Y^n_{1,i}) = nH(X_1|Y_{1,i})$, for any $z \in [K]$;

(e) follows since for all $\mathbf{M} \in \mathfrak{M}$, $z \in [K]$, $\mathcal{T}' \subseteq [N]$, and $\mathcal{T} \subseteq [N]$, such that $|\mathcal{T}| = T$, and $\mathbf{M}(z) = i$ we have

$$I\Big(\mathbf{A}_{\mathcal{T}}; \mathbf{Q}_{[N]\setminus\mathcal{T}}\Big|\mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, \mathbf{X}^n_{\mathcal{T}'}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \leq H(\mathbf{Q}_{[N]\setminus\mathcal{T}}) = o(n);$$

(f) follows from Lemma 3 with $\mathbf{M}_1$ defined as in (34), shown at the bottom of the page, where $\tau_{a,b} \circ \mathbf{M}$ is the transposition that exchanges $\mathbf{M}(a)$ and $\mathbf{M}(b)$ in the second row of the matrix $\mathbf{M}$;

(g) follows since the second term on the RHS of (32a) can be lower bounded by using the following inequality

$$H\Big(\mathbf{A}_{[N]}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big)$$

$$\begin{aligned} &nR(\mathbf{Q}_{[N]}) \\ &\geq H(\mathbf{A}_{[N]}) \\ &\overset{(a)}{\geq} H\Big(\mathbf{A}_{[N]}\Big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \\ &= H\Big(\mathbf{A}_{[N]}, X^n_z\Big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \\ &\quad - H\Big(X^n_z\Big|\mathbf{Q}_{[N]}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \\ &\overset{(b)}{\geq} H\Big(\mathbf{A}_{[N]}, X^n_z\Big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) - o(n) \\ &= H\Big(X^n_z\Big|\mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \\ &\quad + H\Big(\mathbf{A}_{[N]}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) - o(n) \\ &\overset{(c)}{=} H\Big(X^n_z\Big|Y^n_{Z,i}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) \\ &\quad + H\Big(\mathbf{A}_{[N]}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) - o(n) \\ &\overset{(d)}{=} nH(X_1|Y_{1,i}) + H\Big(\mathbf{A}_{[N]}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) - o(n) \qquad (32a) \\ &\geq nH(X_1|Y_{1,i}) + H\Big(\mathbf{A}_{\mathcal{T}}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) - o(n) \\ &\overset{(e)}{=} nH(X_1|Y_{1,i}) + H\Big(\mathbf{A}_{\mathcal{T}}\Big|X^n_z, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = z, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}\Big) - o(n) \\ &\overset{(f)}{=} nH(X_1|Y_{1,i}) + H\Big(\mathbf{A}_{\mathcal{T}}\Big|X^n_z, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = \bar{z}_{K-1}, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}_1\Big) - o(n) \\ &\geq nH(X_1|Y_{1,i}) + H\Big(\mathbf{A}_{\mathcal{T}}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = \bar{z}_{K-1}, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}_1\Big) - o(n) \qquad (32b) \\ &\overset{(g)}{\geq} nH(X_1|Y_{1,i}) + \frac{T}{N}H\Big(\mathbf{A}_{[N]}\Big|X^n_z, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z = \bar{z}_{K-1}, \boldsymbol{\mathcal{M}}(Z) = i, \boldsymbol{\mathcal{M}} = \mathbf{M}_1\Big) - o(n), \qquad (32c) \end{aligned}$$

$$\mathbf{M}_1 \triangleq \tau_{z,\bar{z}_{K-1}} \circ \mathbf{M} = \begin{pmatrix} (\bar{z}_1 : \bar{z}_{d_1}) & \ldots & \left(z, \bar{z}_{1+d_{[i-1]}}, \ldots, \bar{z}_{-1+d_{[i]}}\right) & \ldots & \left(\bar{z}_{d_{[D-1]}} : \bar{z}_{-1+d_{[D]}}\right) \\ (1, \ldots, 1) & \ldots & (D, i, \ldots, i) & \ldots & (D, \ldots, D, i) \end{pmatrix}, \tag{34}$$

$$\geq \frac{1}{\binom{N}{T}} \sum_{\mathcal{T}:|\mathcal{T}|=T} H\Big(\mathbf{A}_{\mathcal{T}}\big|X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_1\Big) \quad (35a)$$

$$\geq \frac{T}{N} H\Big(\mathbf{A}_{[N]}\big|X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, Z=\bar{z}_{K-1}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_1\Big), \quad (35b)$$

where (35a) follows by writing (32b) for all the $\binom{N}{T}$ different subsets $\mathcal{T} \subseteq [N]$ with cardinality T and adding up all these inequalities; and (35b) follows from Han's inequality [31, Theorem 17.6.1].

Then, similar to (32c), for $\ell \in [K-2]$ with $\mathbf{X}^n_{\bar{\mathbf{z}}_{[K:K-1]}} = \emptyset$, we bound the second term on the RHS of (32c) as provided in (36), shown at the bottom of the next page, where (b) to (g) follow with similar arguments of (b) to (g) in (32c) with $\mathbf{M}_{\ell+1} \triangleq \tau_{\bar{z}_{K-\ell},\bar{z}_{K-\ell-1}} \circ \mathbf{M}_\ell$ and $\mathbf{X}^n_{\bar{\mathbf{z}}_{[K:K-1]}} = \emptyset$. The justification of (c) is, however, different. Specifically, for $i \triangleq \boldsymbol{\mathcal{M}}(Z)$, we have,

$$\begin{aligned} &I\big(\mathbf{Q}_{[N]}, \mathbf{X}^n_{[K]\setminus Z}, \mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}; X_Z^n \big| Y^n_{Z,i}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\big) \\ &= I\big(\mathbf{X}^n_{[K]\setminus Z}, \mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}; X_Z^n \big| Y^n_{Z,i}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\big) \\ &\quad + I\big(\mathbf{Q}_{[N]}; X_Z^n \big| \mathbf{X}^n_{[K]\setminus Z}, \mathbf{Y}^n_{\bar{\mathbf{Z}},[D]}, Y^n_{Z,i}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\big) \quad (37a) \\ &\leq o(n), \quad (37b) \end{aligned}$$

where (37b) holds because the first term on the RHS of (37a) is equal to zero since the files are independent of one another, and the second term on the RHS of (37a) is less than or equal to $o(n)$ since the queries are of negligible normalized download cost;

Then, we repeat (36) to bound the second entropy term on the RHS of (36), as provided in (38), shown at the bottom of page 2901, where

- (a) follows by repeating (36) with $Z = z'$, where z' changes from the first element till the last element of $\big[\bar{z}_{K-1}, \bar{z}_{K-2}, \ldots, \bar{z}_{1+d_{[i-1]}}\big]$;
- (b) follows by repeating (36) with $Z = z'$, where z' changes from the first element till the last element of $\big[\bar{z}_{d_{[i-1]}}, \bar{z}_{d_{[i-1]}-1}, \ldots, \bar{z}_{1+d_{[i-2]}}\big]$;
- (c) follows by induction and repeating (b);
- (d) follows from (1).

B. Achievability Proof

We use the same coding scheme as that of Theorem 1 in Section IV-B with $U \triangleq \boldsymbol{\mathcal{M}}(Z)$ levels instead of D levels. Therefore, from (28) the total normalized download cost is,

$$\begin{aligned} R(U) &= \sum_{\ell=1}^{U} R^{(\ell)} \\ &= \sum_{\ell=1}^{U} \Big(H\big(X_1|Y_{1,\ell}\big) - H\big(X_1|Y_{1,\ell-1}\big)\Big) \times \\ &\qquad \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1-\sum_{i=1}^{\ell-1} d_i}\right) \\ &= \sum_{\ell=1}^{U-1} H\big(X_1|Y_{1,\ell}\big) \left(\frac{T}{N}\right)^{K-\sum_{i=1}^{\ell} d_i} \times \\ &\qquad \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{d_\ell - 1}\right) + H\big(X_1|Y_{1,U}\big) \times \\ &\qquad \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1-\sum_{i=1}^{U-1} d_i}\right). \quad (39) \end{aligned}$$

Calculating the expectation of (39) with respect to U results to

$$\begin{aligned} &\mathbb{E}_U\left[R(U)\right] \\ &= \frac{1}{K} \sum_{u=1}^{D} d_u \left[\sum_{\ell=1}^{u-1} H\big(X_1|Y_{1,\ell}\big) \left(\frac{T}{N}\right)^{K-\sum_{i=1}^{\ell} d_i} \times \right. \\ &\qquad \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{d_\ell - 1}\right) \\ &\qquad \left. + H\big(X_1|Y_{1,u}\big) \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1-\sum_{i=1}^{u-1} d_i}\right)\right] \\ &= \frac{1}{K} \left[\sum_{u=1}^{D} d_u \sum_{\ell=1}^{u-1} H\big(X_1|Y_{1,\ell}\big) \left(\frac{T}{N}\right)^{K-\sum_{i=1}^{\ell} d_i} \times \right. \\ &\qquad \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{d_\ell - 1}\right) + \sum_{u=1}^{D} d_u H\big(X_1|Y_{1,u}\big) \times \\ &\qquad \left. \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1-\sum_{i=1}^{u-1} d_i}\right)\right] \\ &= \frac{1}{K} \sum_{\ell=1}^{D} H\big(X_1|Y_{1,\ell}\big) \left[\left(K - \sum_{i=1}^{\ell} d_i\right) \left(\frac{T}{N}\right)^{K-\sum_{i=1}^{\ell} d_i} \times \right. \\ &\qquad \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{d_\ell - 1}\right) \\ &\qquad \left. + d_\ell \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1-\sum_{i=1}^{\ell-1} d_i}\right)\right]. \end{aligned}$$

Finally, similar to Section IV-B.7, privacy is inherited from the privacy of the scheme in [9] and [17].

VI. Conclusion

We have studied the PIR problem with N servers, where each server has a copy of K files and T of the servers may collude, when the client has a noisy version of each of the K files. The side information is such that each file is passed through one of D possible and distinct test channels, whose statistics are known by the client and the servers. We studied this problem under two different security metrics. Under the first metric, the client wants to keep the index of the desired file and the mapping between the files and the test channels secret from the servers. Under the second metric, the client wants to keep the index of the desired file and the mapping between the files and the test channels secret from the servers, but is willing to reveal the index of the test channel that is associated with

the desired file. We derived the optimal normalized download cost under both privacy metrics. We showed that the optimal normalized download cost under the second privacy metric is smaller than or equal to the optimal normalized download cost under the first privacy metric, which shows that revealing the index of the test channel that is associated with the desired file results in a lower normalized download cost. Our setting and results recover several known settings, including PIR with private noiseless side information and PIR with private side information under storage constraints. We note that the PIR problem with noisy side information when the side information is not required to be kept private is an interesting open problem.

APPENDIX A
NESTED SOURCE CODING SCHEMES

In Appendix A-A, we provide a nested random binning scheme to implement nested source coding, as described in Section IV-B.1. In Appendix A-B, we provide an implementation of this scheme with polar codes for the case that the side information forms the Markov chain $X_t - Y_{t,1} - Y_{t,2} - \cdots - Y_{t,D}$, for $t \in [K]$. As formalized next, this Markov chain is always satisfied for test channels that are degraded with respect to one another as, for instance, in Corollaries 2, 3 for binary erasure or binary symmetric test channels.

Lemma 4: If there exists a test channel $\tilde{C}_{i,j}$ such that $C^{(j)} = \tilde{C}_{i,j} \circ C^{(i)}$, for $i, j \in [D]$ and $i < j$, i.e., $C^{(j)}$ is degraded with respect to $C^{(i)}$, then, without loss of generality, one can assume that $X_t - Y_{t,1} - Y_{t,2} - \cdots - Y_{t,D}$, for $t \in [K]$, forms a Markov chain.

Proof: Since the test channels are degraded with respect to one another, one can redefine X_t, $(Y_{t,i})_{i\in[D]}$, $t \in [K]$, such that $X_t - Y_{t,1} - Y_{t,2} - \cdots - Y_{t,D}$ forms a Markov chain. Note that the probability of error in (2) and the privacy condition in (4b) will not be affected because they do not depend on the joint distribution between X_t and $(Y_{t,i}, Y_{t,j})$, for $t \in [K]$, $i, j \in [D]$, and $i \neq j$. □

A. Nested Random Binning Scheme

Consider a discrete memoryless source $(\mathcal{X}_1 \times \bigtimes_{\ell\in[D]} \mathcal{Y}_{1,\ell}, P_{X_1,\mathbf{Y}_{1,[D]}})$ with $D+1$ components. Assume that $(X_1^n, \mathbf{Y}_{1,[D]}^n)$ are i.i.d. samples of this source. Then, consider

$$
\begin{aligned}
&H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell+1:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big)\\
&= H\Big(\mathbf{A}_{[N]}, X^n_{\bar{z}_{K-\ell}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell+1:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big)\\
&\quad - H\Big(X^n_{\bar{z}_{K-\ell}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell+1:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{A}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big)\\
&\overset{(b)}{\geq} H\Big(\mathbf{A}_{[N]}, X^n_{\bar{z}_{K-\ell}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell+1:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big) - o(n)\\
&= H\Big(X^n_{\bar{z}_{K-\ell}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell+1:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big)\\
&\quad + H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big) - o(n)\\
&\overset{(c)}{=} H\Big(X^n_{\bar{z}_{K-\ell}}\big|Y^n_{\bar{z}_{K-\ell},i}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big)\\
&\quad + H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big) - o(n)\\
&\overset{(d)}{=} nH(X_1|Y_{1,i})\\
&\quad + H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big) - o(n)\\
&\geq nH(X_1|Y_{1,i})\\
&\quad + H\Big(\mathbf{A}_{\mathcal{T}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big) - o(n)\\
&\overset{(e)}{=} nH(X_1|Y_{1,i})\\
&\quad + H\Big(\mathbf{A}_{\mathcal{T}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_\ell\Big) - o(n)\\
&\overset{(f)}{=} nH(X_1|Y_{1,i})\\
&\quad + H\Big(\mathbf{A}_{\mathcal{T}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{\mathcal{T}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell-1}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_{\ell+1}\Big) - o(n)\\
&\geq nH(X_1|Y_{1,i})\\
&\quad + H\Big(\mathbf{A}_{\mathcal{T}}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell-1}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_{\ell+1}\Big) - o(n)\\
&\overset{(g)}{\geq} nH(X_1|Y_{1,i})\\
&\quad + \frac{T}{N} H\Big(\mathbf{A}_{[N]}\big|\mathbf{X}^n_{\bar{\mathbf{z}}_{[K-\ell:K-1]}}, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}, Z=\bar{z}_{K-\ell-1}, \boldsymbol{\mathcal{M}}(Z)=i, \boldsymbol{\mathcal{M}}=\mathbf{M}_{\ell+1}\Big) - o(n), \qquad (36)
\end{aligned}
$$

an encoder $\mathcal{E}: \mathcal{X}_1^n \to \boldsymbol{\mathcal{J}}_{[D]}^{(1)}$, that assigns D random bin indices $\mathbf{J}_{[D]}^{(1)} \triangleq (J_\ell^{(1)})_{\ell\in[D]}$ to the sequence X_1^n, where the asymptotic rate of $J_\ell^{(1)}$, for $\ell \in [D]$, is $H(X_1|Y_{1,\ell}) - H(X_1|Y_{1,\ell-1})$, with the convention $H(X_1|Y_{1,0}) = 0$. Consider D decoders $\mathcal{D}_\ell : \boldsymbol{\mathcal{J}}_{[\ell]}^{(1)} \times \mathcal{Y}_{1,\ell}^n \to \mathcal{X}_1^n$, for $\ell \in [D]$, such that the decoder $\mathcal{D}_\ell$ assigns an estimate $\hat{X}_1^n$ to $(\mathbf{J}_{[\ell]}^{(1)}, Y_{1,\ell}^n)$ if there is a unique $\hat{X}_1^n$ such that $\left(\hat{X}_1^n, Y_\ell^n\right)$ are jointly typical and $\left(J_1^{(1)}, J_2^{(1)}, \ldots, J_\ell^{(1)}\right)$ corresponds to the first ℓ components of $\mathcal{E}(\hat{X}_1^n)$. According to [29], [30, Section 10.4], since the asymptotic sum rate for the bin indices that are used at the decoder $\mathcal{D}_\ell$, i.e., $\mathbf{J}_{[\ell]}^{(1)}$, is $H(X_1|Y_{1,\ell})$, then $\mathbb{P}\left[\hat{X}_1^n \neq X_1^n\right] \xrightarrow[n\to\infty]{} 0$.

B. Nested Polar Coding Scheme

We now provide an implementation for the nested source coding in Section IV-B by using nested polar codes when the side information available at the decoders forms a Markov chain. We will rely on the following result for source coding with side information from [33].

Lemma 5 (Source Coding With Side Information [33]): Consider a probability distribution p_{XY} over $\mathcal{X}\times\mathcal{Y}$ with $|\mathcal{X}| = 2$ and $\mathcal{Y}$ a finite alphabet. Let N be a power of 2 and consider (X^N, Y^N) distributed according to $\prod_{i=1}^N p_{XY}$. Define $U^N \triangleq X^N G_N$, where $G_N \triangleq \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes \log N}$ is the source polarization matrix defined in [33]. Define also for $\delta_N \triangleq 2^{-N^\beta}$ with $\beta \in]0, \frac{1}{2}[$, the set $\mathcal{H}_{X|Y} \triangleq \left\{i \in [N] : H(U_i|U^{i-1}Y^N) > \delta_N\right\}$. Given $U^N[\mathcal{H}_{X|Y}]$ and Y^N, one can form $\hat{U}^N$ by the successive cancellation decoder of [33] such that $\mathbb{P}[\hat{U}^N \neq U^N] \leq N\delta_N$. Moreover, $\lim_{N\to\infty} |\mathcal{H}_{X|Y}|/N = H(X|Y)$.

Let $N = 2^n$. Fix a joint probability distribution $P_{X_1\mathbf{Y}_{1,[D]}} \triangleq P_{X_1} p_{Y_{1,1}|X_1} \prod_{\ell=2}^D P_{Y_{1,\ell}|Y_{1,\ell-1}}$ over $\mathcal{X}_1 \times \boldsymbol{\mathcal{Y}}_{1,[D]}$, where $|\mathcal{X}_1| = 2$, $(\mathcal{Y}_{1,\ell})_{\ell\in[D]}$ are finite alphabets, $\boldsymbol{\mathcal{Y}}_{1,[D]} \triangleq \bigtimes_{\ell\in[D]} \mathcal{Y}_{1,\ell}$, and $\mathbf{Y}_{1,[D]} \triangleq (Y_{1,\ell})_{\ell\in[D]}$. Define $U^N \triangleq X^N G_N$. For $\delta_N \triangleq 2^{-N^\beta}$, $\beta \in]0, \frac{1}{2}[$, define for $\ell \in [D]$,

$$\mathcal{H}_{X|Y_\ell} \triangleq \left\{i \in [N] : H(U_i|U^{i-1}Y_{1,\ell}^N) \geq \delta_N\right\}.$$

Lemma 6: For $\ell \in [D-1]$, we have $\mathcal{H}_{X_1|Y_{1,\ell}} \subset \mathcal{H}_{X_1|Y_{1,\ell+1}}$.

Proof: Let $i \in \mathcal{H}_{X_1|Y_{1,\ell}}$. We have

$$\begin{aligned}\delta_N &\overset{(a)}{\leq} H(U_i|U^{i-1}Y_{1,\ell}^N)\\ &\overset{(b)}{=} H(U_i|U^{i-1}Y_{1,\ell}^N Y_{\ell+1}^N)\\ &\overset{(c)}{\leq} H(U_i|U^{i-1}Y_{1,\ell+1}^N),\end{aligned}$$

where (a) holds because $i \in \mathcal{H}_{X_1|Y_{1,\ell}}$, (b) holds because $I(U^i; Y_{1,\ell+1}^N|U^{i-1}Y_{1,\ell}^N) \leq I(U^N; Y_{1,\ell+1}^N|Y_{1,\ell}^N) = 0$, (c) holds because conditioning does not increase entropy. □

From Lemmas 5 and 6, we deduce the following proposition.

Proposition 1: Let $\ell \in [D-1]$. Define $J_\ell \triangleq U^N[\mathcal{H}_{X_1|Y_{1,\ell}}]$ and $J'_{\ell+1} \triangleq U^N[\mathcal{H}_{X_1|Y_{1,\ell+1}} \backslash \mathcal{H}_{X_1|Y_{1,\ell}}]$. Then, $\lim_{N\to\infty} |J_\ell|/N = H(X_1|Y_{1,\ell})$, $\lim_{N\to\infty} |J'_{\ell+1}|/N = H(X_1|Y_{1,\ell+1}) - H(X_1|Y_{1,\ell})$, and one can reconstruct X^N from $(J_\ell, J'_{\ell+1}, Y_{1,\ell+1}^N)$ with vanishing probability of error as N goes to infinity.

Proof: We have $|J_\ell|/N = |\mathcal{H}_{X_1|Y_{1,\ell}}|/N \xrightarrow[N\to\infty]{} H(X_1|Y_{1,\ell})$, where the limit holds by [33]. Then, by Lemma 6, we have $|J'_{\ell+1}|/N = |\mathcal{H}_{X_1|Y_{1,\ell+1}} \backslash \mathcal{H}_{X_1|Y_{1,\ell}}|/N =$

$$\begin{aligned}
&R\left(\mathbf{Q}_{[N]}\right)\\
&\overset{(a)}{\geq} H\left(X_1|Y_{1,i}\right) + \frac{T}{N}\left[H\left(X_1|Y_{1,i}\right) + \frac{T}{N}\left[H\left(X_1|Y_{1,i}\right) + \cdots + \frac{T}{N}\left[H\left(X_1|Y_{1,i}\right)\right.\right.\right.\\
&\quad\left.\left.\left. + \frac{1}{n}H\left(\mathbf{A}_{[N]}\middle|\mathbf{X}_{\mathbf{z}_{[1+d_{[i-1]}:K-1]}}^n, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, Z = \bar{z}_{d_{[i-1]}}, \boldsymbol{\mathcal{M}} = \mathbf{M}_{K-d_{[i-1]}}\right)\right]\right]\right] - o(1)\\
&= H\left(X_1|Y_{1,i}\right)\left[1 + \frac{T}{N} + \left(\frac{T}{N}\right)^2 + \cdots + \left(\frac{T}{N}\right)^{-1+\sum_{\ell=i}^D d_i}\right]\\
&\quad + \frac{1}{n}\left(\frac{T}{N}\right)^{\sum_{\ell=i}^D d_i} H\left(\mathbf{A}_{[N]}\middle|\mathbf{X}_{\mathbf{z}_{[1+d_{[i-1]}:K-1]}}^n, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, Z = \bar{z}_{d_{[i-1]}}, \boldsymbol{\mathcal{M}} = \mathbf{M}_{K-d_{[i-1]}}\right) - o(1)\\
&\overset{(b)}{\geq} H\left(X_1|Y_{1,i}\right)\Psi^{-1}\left(\frac{T}{N}, d_{[i:D]}\right) + H\left(X_1|Y_{1,i-1}\right)\left(\frac{T}{N}\right)^{d_{[i:D]}}\Psi^{-1}\left(\frac{T}{N}, d_{i-1}\right)\\
&\quad + \frac{1}{n}\left(\frac{T}{N}\right)^{d_{[i-1:D]}} H\left(\mathbf{A}_{[N]}\middle|\mathbf{X}_{\mathbf{z}_{[1+d_{[i-2]}:K-1]}}^n, X_z^n, \mathbf{Q}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, Z = \bar{z}_{d_{[i-2]}}, \boldsymbol{\mathcal{M}} = \mathbf{M}_{K-d_{[i-2]}}\right) - o(1)\\
&\overset{(c)}{\geq} H\left(X_1|Y_{1,i}\right)\Psi^{-1}\left(\frac{T}{N}, d_{[i:D]}\right) + \sum_{\ell=1}^{i-1} H\left(X_1|Y_{1,\ell}\right)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N}, d_\ell\right)\\
&\quad + \frac{1}{n}\left(\frac{T}{N}\right)^{K-1} H\left(\mathbf{A}_{[N]}\middle|\mathbf{X}_{[K]}^n, \mathbf{Q}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, Z = \bar{z}_1, \boldsymbol{\mathcal{M}} = \mathbf{M}_{K-1}\right) - o(1)\\
&\overset{(d)}{=} H\left(X_1|Y_{1,i}\right)\Psi^{-1}\left(\frac{T}{N}, d_{[i:D]}\right) + \sum_{\ell=1}^{i-1} H\left(X_1|Y_{1,\ell}\right)\left(\frac{T}{N}\right)^{d_{[\ell+1:D]}}\Psi^{-1}\left(\frac{T}{N}, d_\ell\right) - o(1),
\end{aligned} \tag{38}$$

$|\mathcal{H}_{X_1|Y_{1,\ell+1}}|/N - |\mathcal{H}_{X_1|Y_{1,\ell}}|/N \xrightarrow[N\to\infty]{} H(X_1|Y_{1,\ell+1}) - H(X_1|Y_{1,\ell})$, where the limit holds by [33]. Finally, the near lossless reconstruction of X_1^N from $(J_\ell, J'_{\ell+1}) = U^N[\mathcal{H}_{X_1|Y_{1,\ell+1}}]$ follows from Lemma 5. □

REFERENCES

[1] H. ZivariFard and R. A. Chou, "Private information retrieval when private noisy side information is available," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Taiwan, Jun. 2023, pp. 1538–1543.

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

[3] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, 1998.

[4] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.

[5] A. Beimel, "Secret-sharing schemes: A survey," in *Proc. Int. Conf. Coding Cryptol.* Berlin, Germany: Springer, 2011, pp. 11–46.

[6] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," in *Proc. 13th Annu. ACM Symp. Theory Comput.*, 1998, pp. 151–160.

[7] M. O. Rabin, "How to exchange secrets with oblivious transfer," Aiken Comput. Lab, Harvard Univ., Cambridge, MA, USA, Tech. Rep. TR-81, May 1981.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," in *Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2020, pp. 1–6.

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

[21] 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. Inf. Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1042–1046.

[22] 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.

[23] A. Heidarzadeh and A. Sprintson, "The role of reusable and single-use side information in private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, VIC, Australia, Jun./Jul. 2022, pp. 414–419.

[24] 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.

[25] 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.

[26] M. J. 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.

[27] 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.

[28] Y.-P. Wei, K. Banawan, and S. Ulukus, "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.

[29] D. Slepian and J. K. Wolf, "A coding theorem for multiple access channels with correlated sources," *Bell Syst. Tech. J.*, vol. 52, no. 7, pp. 1037–1076, Sep. 1973.

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

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

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

[33] E. Arikan, "Source polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 2010, pp. 899–903.

Hassan ZivariFard received the M.Sc. degree in electrical engineering from the K. N. Toosi University of Technology, Tehran, Iran, and the Ph.D. degree in electrical engineering from The University of Texas at Dallas, Richardson, TX, USA. He is currently a Post-Doctoral Research Scientist with the Department of Electrical Engineering, Columbia University, New York, NY, USA. Prior to that, he was a Visiting Research Scholar with Wichita State University, Wichita, KS, USA.

Rémi A. Chou received the Engineering degree from Supélec, Gif-sur-Yvette, France, in 2011, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2015. From 2015 to 2017, he was a Post-Doctoral Scholar with The Pennsylvania State University, University Park, PA, USA. From 2017 to 2023, he was an Assistant Professor with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, USA. He is currently an Assistant Professor with the Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX, USA.