Noisy Private Information Retrieval

Karim Banawan Sennur Ulukus
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
kbanawan@umd.edu *ulukus@umd.edu*

***Abstract*—We consider the problem of noisy private information retrieval (NPIR) from N non-communicating databases, each storing the same set of M messages. In this model, the answer strings are not returned through noiseless bit pipes, but rather through noisy memoryless channels. We aim at characterizing the PIR capacity for this model as a function of the statistical information measures of the noisy channels. We derive a general upper bound for the retrieval rate in the form of a max-min optimization. We use the achievable schemes for the PIR problem under asymmetric traffic constraints and the random coding arguments to derive a general lower bound for the retrieval rate. The upper and lower bounds match for $M = 2$ and $M = 3$, for any N, and any noisy channel. The results imply that separation between channel coding and retrieval is optimal except for adapting the traffic ratio from the databases.**

I. Introduction

In the era of big data, efficient data-mining techniques are present everywhere, from social media to online-shopping and search history. These new challenges motivate studying the privacy issues that arise in modern networks. Private information retrieval (PIR), introduced by Chor et al. [1] and remained an important research avenue in computer science community (e.g., [2]–[4]), is a canonical problem to study the privacy of the downloaded content from public databases. In classical PIR, a user wishes to retrieve a file privately from N distributed and non-colluding databases each storing the same set of M messages (files), in a way that no database can learn the identity of the user's desired file. To that end, the user submits queries for the databases that do not reveal the user's interest in the desired file. The databases respond with *correct* answer strings via *noiseless links*, from which the user reconstructs the desired file. PIR schemes are designed to be more efficient than the trivial scheme of downloading all the files stored in the databases in terms of the retrieval rate, which is the ratio between the number of downloaded bits from the desired message and the total download.

Recently, the PIR problem has attracted a renewed interest within the information theory community [5]–[9]. In the effort of characterizing the fundamental limits of the problem, Sun-Jafar introduced the notion of PIR capacity C_{PIR} in [10], which is defined as the supremum of all PIR rates over all achievable retrieval schemes. [10] proved that for the classical PIR model, $C_{\text{PIR}} = (1+\frac{1}{N}+\cdots+\frac{1}{N^{M-1}})^{-1}$. The achievability scheme is a greedy algorithm that employs a *symmetric query* structure for all databases. Following [10], the capacities of many interesting variants of the classical PIR problem have been considered, such as [11]–[36].

In all previous works, the links from the databases to the user are assumed to be noiseless. This assumption may not be valid in practice. For instance, while browsing (retrieving) the internet, some packets may be dropped randomly. This scenario can be abstracted by passing the answer strings through an erasure channel. Alternatively, the data packets may be randomly corrupted, which can be modeled as a binary symmetric channel that flips randomly some positions of the answer strings. Consequently, a more realistic setting is to assume that the databases return their answer strings through memoryless noisy channels with known transition probabilities. The noisy nature of the channel induces random errors along the received answer strings. This poses many interesting questions, such as: How to devise schemes that mitigate the errors introduced by the channel with a small sacrifice from the retrieval rate? Is there a separation between the channel coding needed for reliable transmission over noisy channels and the private retrieval scheme, or is there a necessity for joint processing? How do the statistical properties of the noisy channels fundamentally affect the retrieval rate?

In this paper, we introduce the noisy PIR (NPIR) problem. In noisy PIR, the nth database is connected to the user via a discrete memoryless channel with known transition probability distribution $p(y_n|x_n)$. Hence, the user needs to decode the desired message *reliably* by observing the noisy versions of the returned answer strings. Intuitively, since a channel with worse channel condition needs a lower code rate to combat the channel errors, we do not expect the lengths of the answer strings to be the same from all the databases. Therefore, in this work, we allow the traffic from each database to be *asymmetric* as in [36]. In this work, we aim at characterizing the capacity of the noisy PIR problem in terms of the statistical information measures of the noisy channels, the number of messages M, and the number of databases N. To that end, we first derive a general upper bound for the retrieval rate in the form of a max-min problem. The converse proof is inspired by the converse proof in [35], in particular in the way the asymmetry is handled. We show the achievability proof by random coding arguments and enforcing the uncoded responses to operate on one of the corner points of the PIR problem under asymmetric traffic constraints. The upper and lower bounds match for $M = 2$ and $M = 3$ messages, for arbitrary N databases, and any noisy channel. Our results show that the channel coding needed to mitigate the channel errors and the retrieval scheme

This work was supported by NSF Grants CNS 13-14733, CCF 14-22111, CNS 15-26608 and CCF 17-13977.

are *almost separable* in the sense that the noisy channel affects only the traffic ratio requested from each database and not the explicit coding technique. Interestingly, the upper and lower bounds depend only on the capacity of the noisy channels and not on the explicit transition probability of the channels. We only provide sketches of the proofs here due to space limitations; proof details, illustrative remarks, extra examples and some figures can be found in the longer version [37].

II. System Model

Consider a classical PIR model with N replicated and non-communicating databases storing M messages. Each database stores the same set of messages $W_{1:M} = \{W_1, \cdots, W_M\}$. The mth message W_m is an L-length vector picked uniformly from $\mathbb{F}_2^L$. The messages $W_{1:M}$ are i.i.d., i.e., for $m \in \{1, \cdots, M\}$,

$$H(W_m) = L, \quad H(W_{1:M}) = ML \tag{1}$$

In PIR, in order to retrieve W_i, the user submits N queries $Q_{1:N}^{[i]} = \{Q_1^{[i]}, \cdots, Q_N^{[i]}\}$, one for each database. The queries and the messages are statistically independent because the user has no knowledge about $W_{1:M}$,

$$I(W_{1:M}; Q_{1:N}^{[i]}) = 0, \quad i \in \{1, \cdots, M\} \tag{2}$$

The nth database responds with a t_n-length answer string $A_n^{[i]} = (X_{n,1}^{[i]}, \cdots, X_{n,t_n}^{[i]})$. The nth answer string is a deterministic function of $(W_{1:M}, Q_n^{[i]})$, i.e., for all n and i,

$$H(A_n^{[i]} | W_{1:M}, Q_n^{[i]}) = 0 \tag{3}$$

In noisy PIR, the user receives the nth answer string via a discrete memoryless channel with a transition probability $p(y_n|x_n)$. Thus, the user receives a noisy answer string $\tilde{A}_n^{[i]} = (Y_{n,1}^{[i]}, \cdots, Y_{n,t_n}^{[i]})$. Therefore, we have,

$$P\left(\tilde{A}_n^{[i]} | A_n^{[i]}\right) = \prod_{\eta_n=1}^{t_n} p\left(y_{n,\eta_n}^{[i]} | x_{n,\eta_n}^{[i]}\right) \tag{4}$$

Denote the channel capacity of the nth response channel by C_n, where $C_n = \max_{p(x_n)} I(X_n; Y_n)$ and X_n, Y_n are the single-letter input-output for the nth channel. Let the channel capacities be ordered such that $C_1 \geq C_2 \geq \cdots \geq C_N$ and $\mathbf{C} = (C_1, \cdots, C_N)$ be the vector of the channel capacities.

The user and the databases agree on suitable lengths $\{t_n\}_{n=1}^N$ for the answer strings, which may be different in general. Define the traffic ratio vector $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$, where $\tau_n = \frac{t_n}{\sum_{j=1}^N t_j}$, $n \in \{1, \cdots, N\}$.

To ensure the privacy, $Q_n^{[i]}$ should not reveal any information about i. We can write the privacy constraint as,

$$(Q_n^{[i]}, A_n^{[i]}, W_{1:M}) \sim (Q_n^{[j]}, A_n^{[j]}, W_{1:M}),\ i, j \in \{1, \cdots, M\} \tag{5}$$

In addition, the user should reconstruct the desired message W_i by observing $\tilde{A}_{1:N}^{[i]}$ with arbitrarily small probability of error $P_e(L)$, i.e., $P_e(L) \to 0$ as $L \to \infty$, hence

$$H(W_i | Q_{1:N}^{[i]}, \tilde{A}_{1:N}^{[i]}) \leq 1 + P_e(L) \cdot L = o(L) \tag{6}$$

where $\frac{o(L)}{L} \to 0$ as $L \to \infty$.

The retrieval rate $R(\boldsymbol{\tau}, \mathbf{C})$ is achievable if there exists a sequence of retrieval schemes, indexed by L, that satisfy (5), (6) with answer string lengths $\{t_n\}_{n=1}^N$ that yield $\boldsymbol{\tau}$, thus,

$$R(\boldsymbol{\tau}, \mathbf{C}) = \lim_{L \to \infty} \frac{L}{\sum_{n=1}^N t_n} \tag{7}$$

Consequently, the retrieval rate $R(\mathbf{C})$ is the supremum of $R(\boldsymbol{\tau}, \mathbf{C})$ over all traffic ratio vectors in $\mathbb{T} = \{(\tau_1, \cdots, \tau_N) : \tau_n \geq 0\ \forall n, \sum_{n=1}^N \tau_n = 1\}$. The PIR capacity $C_{\text{PIR}}(\mathbf{C}) = \sup\ R(\mathbf{C})$ over all achievable retrieval schemes.

III. Main Results and Discussions

Theorem 1 (Upper bound) *For noisy PIR, $C_{PIR}(\mathbf{C})$ is upper bounded by $\bar{C}(\mathbf{C})$ which is given by:*

$$\max_{\boldsymbol{\tau} \in \mathbb{T}} \min_{n_i \in [N]} \frac{\theta(0) + \frac{\theta(n_1)}{n_1} + \frac{\theta(n_2)}{n_1 n_2} + \cdots + \frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}}{1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}} \tag{8}$$

where $\theta(\ell) = \sum_{n=\ell+1}^N \tau_n C_n$

The proof of this upper bound is given in Section IV.

Theorem 2 (Lower bound) *For noisy PIR, for a monotone non-decreasing sequence $\mathbf{n} = \{n_i\}_{i=0}^{M-1} \subset \{1, \cdots, N\}^M$, let $n_{-1} = 0$, and $\mathcal{S} = \{i \geq 0 : n_i - n_{i-1} > 0\}$. Denote $y_\ell[k]$ to be the number of stages of the achievable scheme that downloads k-sums from the nth database in one repetition of the scheme, such that $n_{\ell-1} \leq n \leq n_\ell$, and $\ell \in \mathcal{S}$. Let $\xi_\ell = \prod_{s \in \mathcal{S} \setminus \{\ell\}} \binom{M-2}{s-1}$. The number of stages $y_\ell[k]$ is characterized by the following system of difference equations:*

$$\begin{aligned}
y_0[k] &= (n_0 - 1) y_0[k-1] + \sum_{j \in \mathcal{S} \setminus \{0\}} (n_j - n_{j-1}) y_j[k-1] \\
y_1[k] &= (n_1 - n_0 - 1) y_1[k-1] + \sum_{j \in \mathcal{S} \setminus \{1\}} (n_j - n_{j-1}) y_j[k-1] \\
y_\ell[k] &= n_0 \xi_\ell \delta[k - \ell - 1] + (n_\ell - n_{\ell-1} - 1) y_\ell[k-1] \\
&\quad + \sum_{j \in \mathcal{S} \setminus \{\ell\}} (n_j - n_{j-1}) y_j[k-1], \quad \ell \geq 2
\end{aligned} \tag{9}$$

where $\delta[\cdot]$ denotes the Kronecker delta function. The initial conditions of (9) are $y_0[1] = \prod_{s \in \mathcal{S}} \binom{M-2}{s-1}$, and $y_j[k] = 0$ for $k \leq j$. Then, the achievable rate corresponding to $\mathbf{n}$ is:

$$R(\mathbf{n}, \mathbf{C}) = \frac{\sum_{\ell \in \mathcal{S}} \sum_{k=1}^M \binom{M-1}{k-1} y_\ell[k] (n_\ell - n_{\ell-1})}{\sum_{\ell \in \mathcal{S}} \sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{\sum_{k=1}^M \binom{M}{k} y_\ell[k]}{C_n}} \tag{10}$$

Consequently, the capacity $C_{PIR}(\mathbf{C})$ is lower bounded by:

$$C_{PIR}(\mathbf{C}) \geq R(\mathbf{C}) = \max_{n_0 \leq \cdots \leq n_{M-1} \in \{1, \cdots, N\}} R(\mathbf{n}, \mathbf{C}) \tag{11}$$

The proof of Theorem 2 is given in Section V. We note that the bounds on $C_{\text{PIR}}(\mathbf{C})$ do not depend explicitly on the transition probability of the channels $p(y_n|x_n)$, but rather depend on the capacity of the noisy channels C_n. Furthermore, the bounds imply that the channel coding needed for combating channel errors is *almost separable* from the retrieval scheme.

The channel coding problem and the retrieval problem are coupled only through agreeing on a traffic ratio vector $\boldsymbol{\tau}$. Other than $\boldsymbol{\tau}$, the channel coding acts as an outer code for the responses of the databases to the user queries.

Corollary 1 (Capacity for $M = 2, 3$ messages) *For noisy PIR, the capacity $C_{PIR}(\mathbf{C})$ for $M = 2$, and a fixed N is:*

$$C(\mathbf{C}) = \max_{n_i \in [N]} \frac{n_0 n_1}{\sum_{n=1}^{n_0} \frac{n_0+1}{C_n} + \sum_{n=n_0+1}^{n_1} \frac{n_0}{C_n}} \tag{12}$$

and for $M = 3$, $C(\mathbf{C})$ is given by,

$$\max_{n_i \in [N]} \frac{n_0 n_1 n_2}{\sum_{n=1}^{n_0} \frac{n_0 n_1 + n_0 + 1}{C_n} + \sum_{n=n_0+1}^{n_1} \frac{n_0 n_1 + n_0}{C_n} + \sum_{n=n_1+1}^{n_2} \frac{n_0 n_1}{C_n}} \tag{13}$$

The proof of Corollary 1 follows from the optimality of PIR-WTC-II scheme in [36] by replacing $1 - \mu_n$ by C_n.

IV. Converse Proof

The main idea of the converse hinges on the fact that the traffic from the databases should be dependent on the relative channel qualities (i.e., channel capacities) of the response channels. Thus, we extend the converse proof in [35] to account for the noisy observations. We will need the following lemmas, whose proofs can be found in [37].

Lemma 1 (Interference bound) *For NPIR, the mutual information between the interfering messages $W_{2:M}$ and the noisy answers $\tilde{A}_{1:N}^{[1]}$ given the desired message W_1 is bounded by,*

$$I\left(W_{2:M}; Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]} | W_1\right) \leq \sum_{n=1}^{N} t_n C_n - L + o(L) \tag{14}$$

Lemma 2 (Induction lemma) *For all $m \in \{2, \ldots, M\}$ and for an arbitrary $n_{m-1} \in \{1, \cdots, N\}$, the mutual information term in Lemma 1 can be inductively lower bounded as,*

$$\begin{aligned} &I\left(W_{m:M}; Q_{1:N}^{[m-1]}, \tilde{A}_{1:N}^{[m-1]} | W_{1:m-1}\right) \\ &\quad \geq \frac{1}{n_{m-1}} I\left(W_{m+1:M}; Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m}\right) \\ &\qquad + \frac{1}{n_{m-1}} \left(L - \sum_{n=n_{m-1}+1}^{N} t_n C_n\right) - \frac{o(L)}{n_{m-1}} \end{aligned} \tag{15}$$

Now, we derive an explicit upper bound for the retrieval rate. Fixing the length of the nth answer string to t_n and applying Lemma 1 and Lemma 2 successively for an arbitrary sequence $\{n_i\}_{i=1}^{M-1} \subset \{1, \cdots, N\}^{M-1}$, we have the following,

$$\begin{aligned} &\sum_{n=1}^{N} t_n C_n - L + \tilde{o}(L) \\ &\overset{(14)}{\geq} I\left(W_{2:M}; Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]} | W_1\right) \qquad (16) \\ &\overset{(15)}{\geq} \frac{1}{n_1}\left(L - \sum_{n=n_1+1}^{N} t_n C_n\right) + \frac{1}{n_1} I\left(W_{3:M}; Q_{1:N}^{[2]}, \tilde{A}_{1:N}^{[2]} | W_{1:2}\right) \qquad (17) \\ &\overset{(15)}{\geq} \frac{1}{n_1}\left(L - \sum_{n=n_1+1}^{N} t_n C_n\right) + \frac{1}{n_1 n_2}\left(L - \sum_{n=n_2+1}^{N} t_n C_n\right) \\ &\qquad + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}\left(L - \sum_{n=n_{M-1}+1}^{N} t_n C_n\right) \qquad (18) \end{aligned}$$

where $\tilde{o}(L) = \left(1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}\right) o(L)$, (16) follows from Lemma 1, and the remaining steps follow from successive application of Lemma 2. Ordering terms, we have,

$$\begin{aligned} &\left(1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}\right) L \\ &\quad \leq \left(\theta(0) + \frac{\theta(n_1)}{n_1} + \cdots + \frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}\right) \sum_{n=1}^{N} t_n + \tilde{o}(L) \end{aligned} \tag{19}$$

where $\theta(\ell) = \sum_{n=\ell+1}^{N} \tau_n C_n$. We conclude the proof by taking $L \to \infty$. Thus, for an arbitrary sequence $\{n_i\}_{i=1}^{M-1}$, we have

$$R(\boldsymbol{\tau}, \mathbf{C}) = \frac{L}{\sum_{n=1}^{N} t_n} \leq \frac{\theta(0) + \frac{\theta(n_1)}{n_1} + \cdots + \frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}}{1 + \frac{1}{n_1} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}} \tag{20}$$

We get the tightest bound by minimizing over the sequence $\{n_i\}_{i=1}^{M-1}$ over the set $\{1, \cdots, N\}$. Finally, the user and the databases can agree on a $\boldsymbol{\tau} \in \mathbb{T}$ that maximizes $R(\boldsymbol{\tau}, \mathbf{C})$, hence the retrieval rate $R(\mathbf{C})$ is upper bounded by (8).

V. Achievability Proof

We begin first by a motivating example of PIR from BSC(0.1) and BSC(0.2) for $M = 3$. For this example, we provide an explicit scheme that uses achievability of Shannon's channel coding theorem for BSC using linear block codes [38]. We extend the scheme for arbitrary M, N, and any noisy response channels, by means of the random coding argument.

A. Example: $M = 3$, $N = 2$, via BSC(p_1), BSC(p_2)

1) Explicit Upper Bound: From Theorem 1, (8) can be explicitly written as the following linear program:

$$\begin{aligned} \max_{\tau_2, R} \; & R \\ \text{s.t. } & R \leq \frac{1}{3}(1 - H(p_1)) + \left[(1 - H(p_2)) - \frac{1}{3}(1 - H(p_1))\right] \tau_2 \\ & R \leq \frac{2}{5}(1 - H(p_1)) + \left[\frac{4}{5}(1 - H(p_2)) - \frac{2}{5}(1 - H(p_1))\right] \tau_2 \\ & R \leq \frac{4}{7}(1 - H(p_1)) + \left[\frac{4}{7}(1 - H(p_2)) - \frac{4}{7}(1 - H(p_1))\right] \tau_2 \\ & 0 \leq \tau_2 \leq 1 \end{aligned} \tag{21}$$

The optimal solution of (21) resides at one of the corner points of the feasible set, hence an upper bound for $C_{\text{PIR}}(p_1, p_2)$ is:

$$\max\left\{\frac{1 - H(p_1)}{3}, \frac{2}{\frac{3}{1-H(p_1)} + \frac{1}{1-H(p_2)}}, \frac{4}{\frac{4}{1-H(p_1)} + \frac{3}{1-H(p_2)}}\right\} \tag{22}$$

2) Achievable Scheme for BSC(0.1), BSC(0.2): The explicit upper bound in (22) implies that $R \leq \frac{4}{\frac{4}{1-H(p_1)}+\frac{3}{1-H(p_2)}}$ which is 0.218 for $p_1 = 0.1$, $p_2 = 0.2$. To achieve this rate, we enforce the ratio between the uncoded traffic, i.e., before channel coding, to be $4:3$. This results in coded traffic ratio of $\frac{4}{1-H(p_1)} : \frac{3}{1-H(p_2)}$, which appears in the denominator of the upper bound. Concurrently, this results in retrieving 4 desired bits per scheme repetition, which appear at the numerator.

The user repeats the following retrieval scheme for ν times. Each repetition of the scheme operates over $L^* = 4$ bits from $W_{1:3}$. The user permutes the indices of the bits of each message independently and uniformly. Let $a_i(j)$, $b_i(j)$, $c_i(j)$ denote the ith bit of block j from the permuted message W_1, W_2, W_3, respectively. Assume that the desired file is W_1. In block j, the user requests to download $a_1(j)$, $b_1(j)$, and $c_1(j)$ from database 1. From database 2, the user exploits the side information generated from database 1 by requesting the sums $a_2(j)+b_1(j)$, $a_3(j)+c_1(j)$, and $b_2(j)+c_2(j)$. Finally, the user exploits the side information generated from database 2 by downloading $a_4(j)+b_2(j)+c_2(j)$ from database 1. The query table for the jth block is summarized in Table I. Denote the number of uncoded bits requested from the nth database by D_n, then $D_1 = 4$, $D_2 = 3$. This guarantees that the ratio between the uncoded traffic is $4:3$. This query structure is private, as all combinations of the sums are included in the queries and the message bits are randomly permuted.

TABLE I
THE QUERY TABLE FOR THE jTH BLOCK.

Database 1	Database 2
$a_1(j)$	$a_2(j)+b_1(j)$
$b_1(j)$	$a_3(j)+c_1(j)$
$c_1(j)$	$b_2(j)+c_2(j)$
$a_4(j)+b_2(j)+c_2(j)$	

The nth database concatenates the uncoded responses to the user's queries into a vector $U_n^{[1]}$ of length νD_n, i.e.,

$$\begin{aligned} U_1^{[1]} =&[a_1(1) \;\; b_1(1) \;\; c_1(1) \;\; a_4(1)+b_2(1)+c_2(1) \;\; \cdots \\ & a_1(\nu) \;\; b_1(\nu) \;\; c_1(\nu) \;\; a_4(\nu)+b_2(\nu)+c_2(\nu)]^T \quad (23) \\ U_2^{[1]} =&[a_2(1)+b_1(1) \;\; a_3(1)+c_1(1) \;\; b_2(1)+c_2(1) \;\; \cdots \\ & a_2(\nu)+b_1(\nu) \;\; a_3(\nu)+c_1(\nu) \;\; b_2(\nu)+c_2(\nu)]^T \quad (24) \end{aligned}$$

Using Shannon's theorem for BSC(p) [38, Theorem 4.17], for all $\rho \in (0,1]$, all but less than ρ linear $[n,k]$ codes, $\mathcal{C}$, satisfy $P_e(\mathcal{C}) < \frac{2}{\rho} \cdot 2^{-n\Delta(p,R)}$, for some $\Delta(p,R) > 0$, if $R = \frac{k}{n} < 1 - H(p)$. The nth database picks a $(t_n, \nu D_n)$ linear block code, which belongs to the set of these good codes. The nth database encodes the vector $U_n^{[1]}$ to a coded answer string $A_n^{[1]}$ of length t_n such that $t_n = \left\lceil \frac{\nu D_n}{1-H(p_n)} \right\rceil$. This ensures that $\frac{\nu D_n}{t_n} < 1 - H(p_n)$. The nth database responds with $A_n^{[1]}$ via the noisy channel BSC(p_n). The user receives the noisy answer string $\tilde{A}_n^{[1]}$ from the nth database.

For the decoding, the user employs the nearest-codeword decoder to find an estimate of $A_n^{[1]}$ based on $\tilde{A}_n^{[1]}$. Using union bound, the probability of error in decoding is bounded by:

$$P_e(L) \leq \sum_{n=1}^{2} P_e(\mathcal{C}_n) \leq \frac{2}{\rho} \sum_{n=1}^{2} 2^{-t_n \Delta\left(p_n, \frac{\nu D_n}{t_n}\right)} \quad (25)$$

As $\nu \to \infty$, $L \to \infty$ and $t_n \to \infty$, hence $P_e(L) \to 0$. This ensures the decodability of $U_n^{[1]}$ with high probability. Since the vectors $U_1^{[1]}$, $U_2^{[2]}$ are designed to exploit the side information, the user can cancel the undesired messages and be left only with the correct W_1 with probability of error $P_e(L)$.

The retrieval scheme decodes $L = \nu L^* = 4\nu$ bits from the desired messages. The retrieval scheme downloads $t_n = \left\lceil \frac{\nu D_n}{1-H(p_n)} \right\rceil$ from the nth database, hence as $\nu \to \infty$, we have

$$R = \frac{L}{t_1+t_2} = \frac{\nu L^*}{\frac{\nu D_1}{1-H(p_1)} + \frac{\nu D_2}{1-H(p_2)}} = \frac{4}{\frac{4}{1-H(p_1)} + \frac{3}{1-H(p_2)}} \quad (26)$$

B. General Achievable Scheme

The main idea of the general scheme is to use the uncoded response from the nth database to user's query as an *index* for choosing the transmitted codeword from a codebook generated according to the optimal probability distribution. The uncoded query structure is derived from one of the achievable schemes for the corner points of the asymmetric PIR problem [35].

As in [35], we denote the number of side information symbols that are used simultaneously in the initial round at the nth database by $s_n \in \{0, 1, \cdots, M-1\}$. For a given non-decreasing sequence $\{n_i\}_{i=0}^{M-1} \subset \{1, \cdots, N\}^M$, the databases are divided into groups, such that group ℓ includes databases $n_{\ell-1}+1 : n_\ell$. Let $s_n = i$ for all databases belonging to the ith group. Denote $\mathcal{S} = \{i : s_n = i \text{ for some } n \in \{1, \cdots, N\}\}$. Similar to [20], the kth round is the download queries that admit a sum of k different messages (k-sum [10]). A stage of the kth round is a query block that exhausts all $\binom{M}{k}$ combinations of the k-sum. Denote $y_\ell[k]$ to be the number of stages in round k downloaded from the nth database, such that $n_{\ell-1}+1 \leq n \leq n_\ell$. Our scheme is repeated for ν repetitions, each having the same query structure and operating over message symbols of length L^*. Denote the total requested symbols from the nth database in one repetition of the scheme by $D_n(\mathbf{n})$. The details of the achievable scheme are:

1) *Codebook construction:* According to the optimal probability distribution $p^*(x_n)$ (that maximizes $I(X_n;Y_n)$), the nth database constructs a $\left(2^{\nu D_n(\mathbf{n})}, t_n(\mathbf{n})\right)$ codebook $\mathcal{C}_n$, where $t_n(\mathbf{n}) = \left\lceil \frac{\nu D_n(\mathbf{n})}{C_n} \right\rceil$, at random, i.e., $p(x_{n,1}, \cdots, x_{n,t_n(\mathbf{n})}) = \prod_{\eta_n=1}^{t_n(\mathbf{n})} p^*(x_{n,\eta_n})$. Specifically, the codebook $\mathcal{C}_n$ can be written as:
$$\begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_{t_n(\mathbf{n})}(1) \\ x_1(2) & x_2(2) & \cdots & x_{t_n(\mathbf{n})}(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(2^{\nu D_n(\mathbf{n})}) & \cdots & \cdots & x_{t_n(\mathbf{n})}(2^{\nu D_n(\mathbf{n})}) \end{bmatrix} \quad (27)$$
2) *Initialization at the user side:* The user permutes each message independently and uniformly.

3) *Initial download:* From the nth database, $1 \leq n \leq n_0$, the user requests $\prod_{s\in\mathcal{S}} \binom{M-2}{s-1}$ symbols from the desired message and sets the round index $k = 1$.
4) *Message symmetry:* For each stage initiated in the previous step, the user completes the stage by requesting the $\binom{M-1}{k-1}$undesired k-sum combinations.
5) *Database symmetry:* The user repeats step 4 over each group of databases such that the scheme is symmetric within the same group of databases.
6) *Exploitation of side information:* The undesired symbols downloaded within the kth round are used as side information in the $(k+1)$th round. The user requests $(k+1)$-sum consisting of 1 desired symbol and a k-sum of undesired symbols only that were generated in the kth round. The nth database belonging to the ℓth group exploits the side information generated in the kth round from all databases except itself if $s_n \leq k$. Moreover, for $s_n = k$, extra side information can be used by constructing k-sums of the undesired symbols in round 1 from the databases in group 0.
7) *Repeat* steps 4, 5, 6 after setting $k = k+1$ until $k = M$.
8) Repeat steps $3, \cdots, 7$ for a total of ν repetitions.
9) Shuffle the order of the queries uniformly.
10) *Encoding the responses to the user queries:* The nth database concatenates all the responses to the user queries in a vector $U_n^{[i]}$ of length $\nu D_n(\mathbf{n})$. The nth database uses $U_n^{[i]}$ as an index for choosing codeword from $\mathcal{C}_n$, i.e., the index of the codeword and $U_n^{[i]}$ should be in bijection (e.g., by transforming $U_n^{[i]}$ into a decimal value). Consequently, the nth database responds with,

$$A_n^{[i]} = [x_1(U_n^{[i]}) \quad x_1(U_n^{[i]}) \cdots x_{t_n(\mathbf{n})}(U_n^{[i]})]^T \quad (28)$$

REFERENCES

[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. Private information retrieval. *Journal of the ACM*, 45(6):965–981, November 1998.

[2] W. Gasarch. A survey on private information retrieval. In *Bulletin of the EATCS*, volume 82, pages 72–107, 2004.

[3] R. Ostrovsky and W. Skeith III. A survey of single-database private information retrieval: Techniques and applications. In *International Workshop on Public Key Cryptography*, pages 393–411. Springer, 2007.

[4] S. Yekhanin. Private information retrieval. *Communications of the ACM*, 53(4):68–73, April 2010.

[5] N. B. Shah, K. V. Rashmi, and K. Ramchandran. One extra bit of download ensures perfectly private information retrieval. In *IEEE ISIT*, June 2014.

[6] G. Fanti and K. Ramchandran. Efficient private information retrieval over unsynchronized databases. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1229–1239, October 2015.

[7] T. Chan, S. Ho, and H. Yamamoto. Private information retrieval for coded storage. In *IEEE ISIT*, June 2015.

[8] A. Fazeli, A. Vardy, and E. Yaakobi. Codes for distributed PIR with low storage overhead. In *IEEE ISIT*, June 2015.

[9] R. Tajeddine and S. El Rouayheb. Private information retrieval from MDS coded data in distributed storage systems. In *IEEE ISIT*, July 2016.

[10] H. Sun and S. A. Jafar. The capacity of private information retrieval. *IEEE Trans. on Info. Theory*, 63(7):4075–4088, July 2017.

[11] H. Sun and S. Jafar. The capacity of robust private information retrieval with colluding databases. *IEEE Trans. on Info. Theory*, 64(4):2361–2370, April 2018.

[12] H. Sun and S. Jafar. The capacity of symmetric private information retrieval. 2016. Available at arXiv:1606.08828.

[13] K. Banawan and S. Ulukus. The capacity of private information retrieval from coded databases. *IEEE Trans. on Info. Theory*, 64(3):1945–1956, March 2018.

[14] H. Sun and S. A. Jafar. Optimal download cost of private information retrieval for arbitrary message length. *IEEE Trans. on Info. Forensics and Security*, 12(12):2920–2932, December 2017.

[15] Q. Wang and M. Skoglund. Symmetric private information retrieval for MDS coded distributed storage. In *IEEE ICC*, May 2017.

[16] H. Sun and S. A. Jafar. Multiround private information retrieval: Capacity and storage overhead. *IEEE Trans. on Info. Theory*, 64(8):5743–5754, August 2018.

[17] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. Karpuk. Private information retrieval from coded databases with colluding servers. *SIAM Journal on Applied Algebra and Geometry*, 1(1):647–664, 2017.

[18] H. Sun and S. Jafar. Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al. *IEEE Trans. on Info. Theory*, 64(2):1000–1022, February 2018.

[19] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb. Private information retrieval schemes for coded data with arbitrary collusion patterns. In *IEEE ISIT*, June 2017.

[20] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*, 64(10):6842–6862, October 2018.

[21] Y. Zhang and G. Ge. A general private information retrieval scheme for MDS coded databases with colluding servers. 2017. Available at arXiv: 1704.06785.

[22] Y. Zhang and G. Ge. Multi-file private information retrieval from MDS coded databases with colluding servers. 2017. Available at arXiv: 1705.03186.

[23] K. Banawan and S. Ulukus. The capacity of private information retrieval from Byzantine and colluding databases. *IEEE Trans. on Info. Theory*. To appear. Also available at arXiv:1706.01442.

[24] Q. Wang and M. Skoglund. Secure symmetric private information retrieval from colluding databases with adversaries. In *IEEE Allerton*, October 2017.

[25] R. Tandon. The capacity of cache aided private information retrieval. In *IEEE Allerton*, October 2017.

[26] Q. Wang and M. Skoglund. Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers. 2017. Available at arXiv:1708.05673.

[27] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson. Private information retrieval with side information. 2017. Available at arXiv:1709.00112.

[28] Y.-P. Wei, K. Banawan, and S. Ulukus. Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. *IEEE Trans. on Info. Theory*. To appear. Available at arXiv:1709.01056.

[29] Z. Chen, Z. Wang, and S. Jafar. The capacity of private information retrieval with private side information. 2017. Available at arXiv:1709.03022.

[30] Y.-P. Wei, K. Banawan, and S. Ulukus. The capacity of private information retrieval with partially known private side information. 2017. Available at arXiv:1710.00809.

[31] H. Sun and S. A. Jafar. The capacity of private computation. 2017. Available at arXiv:1710.11098.

[32] M. Mirmohseni and M. A. Maddah-Ali. Private function retrieval. 2017. Available at arXiv:1711.04677.

[33] M. Abdul-Wahid, F. Almoualem, D. Kumar, and R. Tandon. Private information retrieval from storage constrained databases–coded caching meets PIR. 2017. Available at arXiv:1711.05244.

[34] Y.-P. Wei, K. Banawan, and S. Ulukus. Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits. *IEEE JSAC*, 36(6):1126–1139, June 2018.

[35] K. Banawan and S. Ulukus. Asymmetry hurts: Private information retrieval under asymmetric-traffic constraints. *IEEE Trans. on Info. Theory*. Submitted January 2018. Also available at arXiv:1801.03079.

[36] K. Banawan and S. Ulukus. Private information retrieval through wiretap channel II: Privacy meets security. *IEEE Trans. on Info. Theory*. Submitted January 2018. Also available at arXiv:1801.06171.

[37] K. Banawan and S. Ulukus. Noisy private information retrieval: Separability of channel coding and information retrieval. *IEEE Trans. on Info. Theory*. Submitted July 2018. Available at arXiv: 1807.05997.

[38] R. Roth. *Introduction to Coding Theory*. Cambridge University Press, 2006.