# Private Information Retrieval Under Asymmetric Traffic Constraints

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 private information retrieval (PIR) of a single message (file) out of $M$ messages from $N$ distributed databases under *asymmetric traffic* from databases. In this problem, the *ratios between the traffic* from the databases are constrained, i.e., the ratio of the length of the answer string that the user receives from the $n$th database to the total length of all answer strings from all databases is constrained to be $\tau_n$. For this problem, for fixed $M$, $N$, we develop a general upper bound $\bar{C}(\boldsymbol{\tau})$. Our converse bound is a piece-wise affine function in the traffic ratio vector $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$. For the lower bound, we explicitly show the achievability of $\binom{M+N-1}{M}$ corner points. For the remaining traffic ratio vectors, we perform time-sharing between these corner points. The recursive structure of our achievability scheme is captured via a system of difference equations. The upper and lower bounds exactly match for $M = 2$ and $M = 3$ for any $N$ and any $\boldsymbol{\tau}$.**

## I. Introduction

Private information retrieval (PIR), introduced by Chor et al. in [1], studies the privacy of the downloaded content from public databases. In the classical PIR setting, a user requests a certain message (or file) out of $M$ distinct messages from $N$ non-communicating (non-colluding) and replicated databases without leaking the identity of the desired message to any individual database. The user prepares $N$ queries, one for each database, and each database responds with an answering string. The user needs to be able to reconstruct the entire message by decoding the answer strings from all databases. The efficiency of a retrieval scheme is measured by the retrieval rate, which is the ratio of the number of desired message symbols to the number of total downloaded symbols. Recently, the PIR problem is revisited by information theorists [2]–[7]. In the leading work [8], Sun and Jafar introduce the PIR capacity, which is defined as the supremum of PIR rates over all achievable retrieval schemes, and determine the exact capacity of the classical PIR. Following [8], the fundamental limits of many variants of PIR have been considered [9]–[32].

A common property of the achievability schemes constructed for these PIR problems is that they exhibit a *symmetric structure* across the databases. Now, consider the following scenarios that render symmetry assumption unworkable: *Varying database availability:* Certain databases are available only a fraction of the time other databases are available for downloads. *Different capacities:* The capacities of the links from the databases to the user have different capacities. This may be due to different physical locations of the databases, or due to the quality of the physical layer communication channel. In these cases, the user is forced to deal with each database differently. This breaks the database symmetry assumption and makes load balancing of desired message and side information more challenging. Motivated by these practical scenarios, we consider the PIR problem under *asymmetric traffic constraints*. Formally, we consider a classical PIR setting with $N$ replicated and non-communicating databases storing $M$ messages. We assume that the $n$th database responds with a $t_n$-length answer string. We constrain the lengths of the answer strings such that $t_n = \lambda_n t_1$ for $n \in \{2, \cdots, N\}$. This, in turn, forces the ratios between the traffic from the databases to be $1 : \lambda_2 : \cdots : \lambda_N$. We denote the traffic ratio with respect to the total download by a vector $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$, where $\tau_n = \frac{\lambda_n}{\sum_{j=1}^{N} \lambda_j}$. We note that $\boldsymbol{\tau}$ is in bijection with $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_N)$. We aim at characterizing the capacity of this PIR problem, $C(\boldsymbol{\tau})$.

To that end, we develop a novel upper bound for the capacity $\bar{C}(\boldsymbol{\tau})$. This generalizes the converse proof of [8], which exploits the database symmetry, to incorporate the asymmetric traffic constraints. We characterize the upper bound as a piece-wise affine function in $\boldsymbol{\tau}$. The upper bound implies that asymmetry fundamentally hurts the retrieval rate. Then, we propose explicit achievability schemes for $\binom{M+N-1}{M}$ corner points. Each corner point corresponds to a specific partitioning of the databases according to the number of side information symbols that are used simultaneously within the initial round of the download. We describe the achievability scheme via a system of difference equations in the number of stages at each round of the download (which is parallel to [18]). For any other $\boldsymbol{\tau}$, we employ time-sharing between the corner points that enclose $\boldsymbol{\tau}$. We provide an explicit rate expression for the case of $N = 2$ for arbitrary $M$. The upper and lower bounds match for the cases of $M = 2$ and $M = 3$ for any $N$ and any $\boldsymbol{\tau}$ leading to the exact capacity $C(\boldsymbol{\tau})$ for these cases. 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 [33].

## II. System Model

Consider a classical PIR model with $N$ non-communicating and replicated databases storing $M$ messages (or files). Each database stores the same set of messages $W_{1:M} =$

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

$\{W_1, \cdots, W_M\}$. Messages $W_{1:M}$ are independent and uniformly distributed over all vectors of size $L$ picked from a finite field $\mathbb{F}_q^L$, i.e., for $i \in \{1, \cdots, M\}$

$$H(W_i) = L, \quad H(W_{1:M}) = ML, \quad (q\text{-ary units}) \tag{1}$$

In PIR, a user wants to retrieve a message $W_i \in W_{1:M}$ without revealing any information about the identity of the message $i$ to any individual database. To that end, the user submits a query $Q_n^{[i]}$ to the $n$th database. The messages and the queries are statistically independent due to the fact that the user does not know the message realizations in advance, i.e.,

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

where $Q_{1:N}^{[i]} = \{Q_1^{[i]}, \cdots, Q_N^{[i]}\}$. The $n$th database responds truthfully by an answer string $A_n^{[i]}$, which is a deterministic function of the query $Q_n^{[i]}$ and all the messages $W_{1:M}$, hence

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

In PIR under asymmetric traffic constraints, the lengths of the answer strings are different. More specifically, we assume that the $n$th database responds with a $t_n$-length answer string, such that $t_n = \lambda_n t_1$, where $\lambda_n$ is the ratio between the traffic from the $n$th database to the traffic from the first database. Without loss of generality, we assume that the databases are ordered descendingly in $\lambda_n$. Hence, $\{\lambda_n\}_{n=1}^N$ is a *non-increasing monotone* sequence with $\lambda_1 = 1$, and $\lambda_n \in [0, 1]$, i.e., for $1 \geq \lambda_2 \geq \cdots \geq \lambda_N$,

$$H(A_n^{[i]}) \leq \lambda_n t_1, \quad i \in \{1, \cdots, M\},\ n \in \{1, \cdots, N\} \tag{4}$$

We define the *traffic ratio* of the $n$th database $\tau_n$ as the ratio between the traffic from the $n$th database and the total traffic from all databases, i.e., $\tau_n = \frac{\lambda_n}{\sum_{j=1}^N \lambda_j}$. We note that there is a one-to-one transformation between the vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_N)$ and the vector $\boldsymbol{\tau} = (\tau_1, \tau_2, \cdots, \tau_N)$.

In order to ensure the privacy, at the $n$th database, the query $Q_n^{[i]}$ designed to retrieve $W_i$ should be indistinguishable from the queries designed to retrieve any other message, i.e.,

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

where $\sim$ denotes statistical equivalence.

In addition, the user should be able to reconstruct $W_i$ from the collected answer strings $A_{1:N}^{[i]}$ with small probability of error. Hence, we have the following reliability constraint,

$$H(W_i | Q_{1:N}^{[i]}, A_{1:N}^{[i]}) = o(L) \tag{6}$$

where $\frac{o(L)}{L} \to 0$ as $L \to \infty$. A retrieval rate $R(\boldsymbol{\tau})$ is achievable if there exists a PIR scheme which satisfies (5) and (6) for some message lengths $L(\boldsymbol{\tau})$ and answer strings of lengths $\{t_n(\boldsymbol{\tau})\}_{n=1}^N$ that satisfy the asymmetric traffic constraint (4), such that

$$R(\boldsymbol{\tau}) = \frac{L(\boldsymbol{\tau})}{\sum_{n=1}^N t_n(\boldsymbol{\tau})} \tag{7}$$

The pair $(L(\boldsymbol{\tau}), t_1(\boldsymbol{\tau}))$ can grow arbitrarily large to conform with the information-theoretic framework.

The capacity of the PIR problem under asymmetric traffic constraints $C(\boldsymbol{\tau})$ is defined as the supremum of all achievable retrieval rates, i.e., $C(\boldsymbol{\tau}) = \sup\ R(\boldsymbol{\tau})$.

## III. Main Results

**Theorem 1 (Upper bound)** *For the PIR problem under non-increasing asymmetric traffic constraints $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$, the PIR capacity $C(\boldsymbol{\tau})$ is upper bounded by $\bar{C}(\boldsymbol{\tau})$*

$$\bar{C}(\boldsymbol{\tau}) = \min_{n_i \in \{1, \cdots, N\}} \frac{1 + \frac{\gamma(n_1)}{n_1} + \frac{\gamma(n_2)}{n_1 n_2} + \cdots + \frac{\gamma(n_{M-1})}{\prod_{i=1}^{M-1} n_i}}{1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}} \tag{8}$$

*where $\gamma(\ell) = \frac{\sum_{n=\ell+1}^N \lambda_n}{\sum_{n=1}^N \lambda_n} = \sum_{n=\ell+1}^N \tau_n$ corresponds to the sum of the traffic ratios from databases $[\ell + 1 : N]$.*

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

The following corollary asserts that there is a strict capacity loss due to the asymmetric traffic constraints if the traffic ratio of the weakest link falls below a certain threshold. The proof can be found in [33].

**Corollary 1 (Asymmetry hurts)** *For the PIR problem under monotone non-increasing asymmetric traffic constraints $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$, if $\tau_N < \tau^*$, such that*

$$\tau^* = \frac{N^{M-1} - 1}{N^M - 1}, \quad N > 1 \tag{9}$$

*then $C(\boldsymbol{\tau}) < C$, where $C = \frac{1}{1 + \frac{1}{N} + \cdots + \frac{1}{N^{M-1}}}$ is the PIR capacity without the asymmetric traffic constraints in [8].*

**Theorem 2 (Lower bound)** *For the PIR problem under asymmetric traffic constraints, 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 $y_\ell[k]$ as the number of stages of the achievable scheme that downloads $k$-sums from the $n$th database, 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:*

$$\begin{aligned}
y_0[k] =& (n_0 - 1) y_0[k-1] + \sum_{j \in \mathcal{S} \setminus \{0\}} (n_j - n_{j-1}) y_j[k-1] \\
y_1[k] =& (n_1 - n_0 - 1) y_1[k-1] + \sum_{j \in \mathcal{S} \setminus \{1\}} (n_j - n_{j-1}) y_j[k-1] \\
y_\ell[k] =& n_0 \xi_\ell \delta[k - \ell - 1] + (n_\ell - n_{\ell-1} - 1) y_\ell[k-1] \\
&+ \sum_{j \in \mathcal{S} \setminus \{\ell\}} (n_j - n_{j-1}) y_j[k-1], \quad \ell \geq 2
\end{aligned} \tag{10}$$

*where $\delta[\cdot]$ denotes the Kronecker delta function. The initial conditions of (10) are $y_0[1] = \prod_{s \in \mathcal{S}} \binom{M-2}{s-1}$, and $y_j[k] = 0$ for $k \leq j$. Consequently, the traffic ratio $\tau_n(\mathbf{n})$ corresponding to the sequence $\mathbf{n} = \{n_i\}_{i=0}^{M-1}$, where $n_{j-1} + 1 \leq n \leq n_j$ is given by:*

$$\tau_n(\mathbf{n}) = \frac{\sum_{k=1}^M \binom{M}{k} y_j[k]}{\sum_{\ell \in \mathcal{S}} \sum_{k=1}^M \binom{M}{k} y_\ell[k] (n_\ell - n_{\ell-1})} \tag{11}$$

*and the achievable rate corresponding to $\boldsymbol{\tau}(\mathbf{n})$ is given by:*

$$R(\boldsymbol{\tau}(\mathbf{n})) = \frac{\sum_{\ell\in\mathcal{S}}\sum_{k=1}^{M}\binom{M-1}{k-1}y_\ell[k](n_\ell - n_{\ell-1})}{\sum_{\ell\in\mathcal{S}}\sum_{k=1}^{M}\binom{M}{k}y_\ell[k](n_\ell - n_{\ell-1})} \tag{12}$$

*Moreover, for $\boldsymbol{\tau} = \sum_{i=1}^{N}\alpha_i\boldsymbol{\tau}(\mathbf{n}_i)$ for $\alpha_i \geq 0$, for all $i$, and $\sum_{i=1}^{N}\alpha_i = 1$, the following is a lower bound on $C(\boldsymbol{\tau})$,*

$$C(\boldsymbol{\tau}) \geq R(\boldsymbol{\tau}) = \sum_{i=1}^{N}\alpha_i R(\boldsymbol{\tau}(\mathbf{n}_i)) \tag{13}$$

We present the achievable scheme of Theorem 2 in Section V. The theorem characterizes an achievable rate for the corner points $\boldsymbol{\tau}(\mathbf{n})$ corresponding to any monotone non-decreasing sequence $\mathbf{n} = \{n_i\}_{i=0}^{M-1} \subset \{1,\cdots,N\}^M$. For any other traffic ratio vector $\boldsymbol{\tau}$, the achievability scheme is obtained by time-sharing between the nearest corner points. We describe the achievable rate by a system of difference equations. The solution of this system of difference equations specifies the traffic ratio vector $\boldsymbol{\tau}(\mathbf{n})$ and the achievable rate corresponding to the monotone non-decreasing sequence $\{n_i\}_{i=0}^{M-1}$. For $N = 2$ and an arbitrary $M$, we provide an explicit rate expression. Let $s_2 = \{1,\cdots,M-1\}$, for the traffic ratio $\tau_2(s_2)$, where

$$\tau_2(s_2) = \frac{\sum_{i=0}^{\lfloor\frac{M-s_2-1}{2}\rfloor}\binom{M}{s_2+2i+1}}{M\binom{M-2}{s_2-1} + \sum_{i=0}^{M-s_2-1}\binom{M}{s_2+1+i}} \tag{14}$$

the PIR capacity $C(\tau_2(s_2))$ is lower bounded by $R(\tau_2(s_2))$:

$$R(\tau_2(s_2)) = \frac{\binom{M-2}{s_2-1} + \sum_{i=0}^{M-s_2-1}\binom{M-1}{s_2+i}}{M\binom{M-2}{s_2-1} + \sum_{i=0}^{M-s_2-1}\binom{M}{s_2+1+i}} \tag{15}$$

Moreover, if $\tau_2(s_2) < \tau_2 < \tau_2(s_2+1)$, and $\alpha \in (0,1)$, such that $\tau_2 = \alpha\tau_2(s_2) + (1-\alpha)\tau_2(s_2+1)$, then

$$C(\tau_2) \geq R(\tau_2) = \alpha R(\tau_2(s_2)) + (1-\alpha)R(\tau_2(s_2+1)) \tag{16}$$

The proof of this result can be found in [33].

**Corollary 2 (Capacity for $M = 2$ and $M = 3$ messages)** *For the PIR problem with asymmetric traffic constraints $\boldsymbol{\tau}$, the capacity $C(\boldsymbol{\tau})$ for $M = 2$ and $M = 3$, and for any arbitrary $N$ is given by:*

$$C(\boldsymbol{\tau}) = \begin{cases} \min\limits_{n_0\in\{1,\cdots,N\}} \dfrac{1+\frac{\sum_{n=n_0+1}^{N}\tau_n}{n_0}}{1+\frac{1}{n_0}}, & M=2 \\ \min\limits_{n_0\leq n_1\in\{1,\cdots,N\}} \dfrac{1+\frac{\sum_{n=n_0+1}^{N}\tau_n}{n_0}+\frac{\sum_{n=n_1+1}^{N}\tau_n}{n_0 n_1}}{1+\frac{1}{n_0}+\frac{1}{n_0 n_1}}, & M=3 \end{cases} \tag{17}$$

The proof of the optimality of our achievable scheme for $M = 2$ and $M = 3$ can be found in [33].

## IV. Converse Proof

We derive an upper bound for PIR with asymmetric traffic constraints. We extend the converse techniques introduced in [8] to account for the asymmetry of the answer strings.

We need the following lemma. The proof of this lemma can be found in [8, Lemma 5]. The proof follows for our case due to the fact that the proof in [8, Lemma 5] deals with the length of the entire downloaded answer strings $A_{1:N}^{[1]}$ and not the individual answer string, see [8, equations (46)-(47)].

**Lemma 1 (Interference lower bound)** *For the PIR problem under asymmetric traffic constraints $\{t_n\}_{n=1}^{N}$, the interference from undesired messages within the answer strings $\sum_{n=1}^{N}t_n - L$ is lower bounded by,*

$$\sum_{n=1}^{N}t_n - L + o(L) \geq I\left(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}|W_1\right) \tag{18}$$

In the following lemma, we prove an inductive relation for the mutual information term on the right hand side of (18).

**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,*

$$\begin{aligned} &I\left(W_{m:M}; Q_{1:N}^{[m-1]}, A_{1:N}^{[m-1]}|W_{1:m-1}\right) \\ &\quad\geq \frac{1}{n_{m-1}}I\left(W_{m+1:M}; Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m}\right) \\ &\qquad + \frac{1}{n_{m-1}}\left(L - t_1\sum_{n=n_{m-1}+1}^{N}\lambda_n\right) - \frac{o(L)}{n_{m-1}} \end{aligned} \tag{19}$$

We note that [8, Lemma 6] can be interpreted as a special case of Lemma 2 with setting $n_{m-1} = N$.

Now, we are ready to derive an explicit upper bound for the retrieval rate under asymmetric traffic constraints. Applying Lemma 1 and Lemma 2 successively for an arbitrary sequence $\{n_i\}_{i=1}^{M-1} \subset \{1,\cdots,N\}^{M-1}$ and observing that $\sum_{n=1}^{N}t_n = t_1\sum_{n=1}^{N}\lambda_n$ under the asymmetric traffic constraints, we have the following

$$\begin{aligned} &t_1\sum_{n=1}^{N}\lambda_n - L + \tilde{o}(L) \\ &\overset{(18)}{\geq} I\left(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}|W_1\right) &(20)\\ &\overset{(19)}{\geq} \frac{1}{n_1}\left(L - t_1\sum_{n=n_1+1}^{N}\lambda_n\right) + \frac{1}{n_1}I\left(W_{3:M}; Q_{1:N}^{[2]}, A_{1:N}^{[2]}|W_{1:2}\right) \\ &\overset{(19)}{\geq} \cdots &(21)\\ &\overset{(19)}{\geq} \frac{1}{n_1}\left(L - t_1\sum_{n=n_1+1}^{N}\lambda_n\right) + \frac{1}{n_1 n_2}\left(L - t_1\sum_{n=n_2+1}^{N}\lambda_n\right) \\ &\qquad + \cdots + \frac{1}{\prod_{i=1}^{M-1}n_i}\left(L - t_1\sum_{n=n_{M-1}+1}^{N}\lambda_n\right) &(22) \end{aligned}$$

where $\tilde{o}(L) = \left(1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1}n_i}\right)o(L)$, (20) follows from Lemma 1, and the remaining bounding steps follow from successive application of Lemma 2.

Ordering terms, we have,

$$\left(1+\frac{1}{n_1}+\frac{1}{n_1n_2}+\cdots+\frac{1}{\prod_{i=1}^{M-1}n_i}\right)L-\tilde{o}(L) \leq \left(1+\frac{\gamma(n_1)}{n_1}+\cdots+\frac{\gamma(n_{M-1})}{\prod_{i=1}^{M-1}n_i}\right)t_1\sum_{n=1}^{N}\lambda_n \quad (23)$$

We conclude the proof by taking $L\to\infty$. Thus, for an arbitrary sequence $\{n_i\}_{i=1}^{M-1}$, we upper bound $R(\boldsymbol{\tau})$ as,

$$\frac{L}{t_1\sum_{n=1}^{N}\lambda_n}\leq\frac{1+\frac{\gamma(n_1)}{n_1}+\frac{\gamma(n_2)}{n_1n_2}+\cdots+\frac{\gamma(n_{M-1})}{\prod_{i=1}^{M-1}n_i}}{1+\frac{1}{n_1}+\frac{1}{n_1n_2}+\cdots+\frac{1}{\prod_{i=1}^{M-1}n_i}} \quad (24)$$

Finally, we get the tightest bound by minimizing over the sequence $\{n_i\}_{i=1}^{M-1}$ over the set $\{1,\cdots,N\}$, this leads to (8).

## V. Achievability Proof

### A. Description of the General Scheme

We describe the general achievable scheme that achieves the retrieval rates in Theorem 2. We first show explicitly the achievability schemes for corner points, i.e., the achievability scheme for every non-decreasing sequence $\{n_i\}_{i=0}^{M-1}\subset\{1,\cdots,N\}^M$. We note that our achievability scheme is different in two key steps: First, we note that database symmetry is not applied over all databases directly as in [8], but rather it is applied over groups of databases, such as, group 0 includes databases 1 through $n_0$, group 1 includes databases $n_0+1$ through $n_1$, etc. Second, we note that each group of databases exploits side information differently in the *initial* round of downloading. More specifically, we note that group 0 of databases do not exploit any side information in the initial round of the download, group 1 exploits 1 side information symbol in the initial round of the download, and so on.

*1) Achievability Scheme for the Corner Points:* Let $s_n\in\{0,1,\cdots,M-1\}$ denote the number of side information symbols that are used simultaneously in the initial round of downloads at the $n$th database. For a given non-decreasing sequence $\{n_i\}_{i=0}^{M-1}\subset\{1,\cdots,N\}^M$, let $s_n=i$ for all $n_{i-1}+1\leq n\leq n_i$ with $n_{-1}=0$ by convention. Denote $\mathcal{S}=\{i: s_n=i \text{ for some } n\in\{1,\cdots,N\}\}$. We follow the round and stage definitions in [18]. Denote $y_\ell[k]$ to be the number of stages in round $k$ downloaded from the $n$th database, such that $n_{\ell-1}+1\leq n\leq n_\ell$. The details of the achievable scheme are as follows:

1) *Initialization:* The user permutes each message independently and uniformly as [8]. From the $n$th database where $1\leq n\leq n_0$, the user downloads $\prod_{s\in\mathcal{S}}\binom{M-2}{s-1}$ symbols from the desired message. The user sets the round index $k=1$, thus $y_0[1]=\prod_{s\in\mathcal{S}}\binom{M-2}{s-1}$ stages.
2) *Message symmetry:* To satisfy the privacy constraint, for each stage initiated in the previous step, the user completes the stage by downloading the $\binom{M-1}{k-1}$ $k$-sum combinations that do not include the desired symbols.
3) *Database symmetry:* We divide the databases into groups. Group $\ell\in\mathcal{S}$ contains databases $n_{\ell-1}+1$ to $n_\ell$. Database symmetry is applied within each group only. The user repeats step 2 over each group of databases.
4) *Exploitation of side information:* The user downloads $(k+1)$-sum consisting of 1 desired symbol and a $k$-sum of undesired symbols that were generated in the $k$th round. Different from [8], for the $n$th database, if $s_n>k$, this database does not exploit the side information generated in the $k$th round. Moreover, for $s_n=k$, extra side information can be used in the $n$th database. The user forms $n_0\prod_{s\in\mathcal{S}\setminus\{s_n\}}\binom{M-2}{s-1}$ stages of side information by constructing $k$-sums of the undesired symbols in round 1 from the databases in group 0.
5) *Repeat* steps 2, 3, 4 after setting $k=k+1$ until $k=M$.
6) *Shuffling the order of the queries:* The user shuffles the order of queries as in [8] to guarantee the privacy.

*2) Achievability Scheme for Non-Corner Points:* The achievability schemes for non-corner points can be derived by time-sharing between the nearest corner points.

### B. Example: $M=3$ Messages, $N=2$ Databases

For the upper bound, we first carry out the minimization in (8) over $n_1,n_2\in\{1,2\}$. Hence, we have the following explicit upper bound on the capacity as function of $\tau_2$.

$$C(\tau_2)\leq\begin{cases}\frac{1}{3}+\frac{2\tau_2}{3}, & 0\leq\tau_2\leq\frac{1}{4}\\ \frac{2}{5}+\frac{2\tau_2}{5}, & \frac{1}{4}\leq\tau_2\leq\frac{3}{7}\\ \frac{4}{7}, & \frac{3}{7}\leq\tau_2\leq\frac{1}{2}\end{cases} \quad (25)$$

To show the achievability of the upper bound in (25), let $a_i$, $b_i$, $c_i$ denote randomly and independently permuted symbols of messages $W_1$, $W_2$, $W_3$, respectively. We show the achievability of the corner points $\tau_2\in\{0,\frac{1}{4},\frac{3}{7},\frac{1}{2}\}$.

*1) Achievability of the Corner Points:*

*a) The $\tau_2=0$ Corner Point:* The second database does not return any answer strings. The achievable scheme is to download all files from the first database, i.e., download $a_1,b_1,c_1$ from database 1. This achieves $R=\frac{1}{3}=C(0)$.

*b) The $\tau_2=\frac{1}{2}$ Corner Point:* $\tau_2=\frac{1}{2}$ means that a symmetric scheme can be applied to both databases. Thus, the optimal achievable scheme is the optimal symmetric scheme in [8], which results in $R=\frac{4}{7}=C(\frac{1}{2})$.

*c) The $\tau_2=\frac{3}{7}$ Corner Point:* (See Table I.) The user can cut the first round of downloads in database 2 and exploit the side information generated from database 1 directly in the form of sums of 2, i.e., the user downloads $a_1,b_1,c_1$ from database 1 and then exploits the undesired symbols as side information by downloading $a_2+b_1$, $a_3+c_1$ from database 2. The user then applies message symmetry and downloads $b_2+c_2$. Since the user uses 1 bit of side information in the initial download round from database 2, $s_2=1$ in this case. Finally, the user exploits the undesired sum $b_2+c_2$ from database 2 as a side information in database 1 and downloads $a_4+b_2+c_2$. Using this scheme the user downloads 4 symbols from database 1 and 3 symbols from database 2, hence $\tau_2=\frac{3}{7}$. The user downloads $L=4$ desired symbols out of 7 downloads, thus $R=\frac{4}{7}=C(\frac{3}{7})$. This scheme is the asymmetric scheme in [12].

TABLE I
THE QUERY TABLE FOR $M = 3$, $N = 2$, $\tau_2 = \frac{3}{7}$.

| Database 1 | Database 2 |
|---|---|
| $a_1, b_1, c_1$ | $a_2 + b_1$<br>$a_3 + c_1$<br>$b_2 + c_2$ |
| $a_4 + b_2 + c_2$ | |

*d) The $\tau_2 = \frac{1}{4}$ Corner Point:* (See Table II.) In this case, the user downloads $a_1, b_1, c_1$ from database 1. In database 2, the user exploits the side information $b_1, c_1$ simultaneously and downloads $a_2 + b_1 + c_1$. In this case $s_2 = 2$, as 2 side information symbols are exploited simultaneously in the initial round of download from database 2. Using this scheme the user downloads 3 symbols from database 1 and 1 symbol from database 2, therefore $\tau_2 = \frac{1}{4}$. The user downloads $L = 2$ desired symbols in 4 downloads, hence $R = \frac{1}{2} = C(\frac{1}{4})$.

TABLE II
THE QUERY TABLE FOR $M = 3$, $N = 2$, $\tau_2 = \frac{1}{2}$.

| Database 1 | Database 2 |
|---|---|
| $a_1, b_1, c_1$ | $a_2 + b_1 + c_1$ |

We illustrate time-sharing by the following example.

*e) Specific Example for Non-Corner Points, $\tau_2 = \frac{1}{3}$:* (See Table III.) The user applies the scheme of $\tau_2 = \frac{3}{7}$, and the scheme of $\tau_2 = \frac{1}{4}$. The scheme downloads 10 symbols from database 1 and 5 symbols from database 2, thus, $\tau_2 = \frac{1}{3}$. The scheme downloads 8 symbols in 15 downloads, hence $R(\frac{1}{3}) = \frac{8}{15} = \frac{2}{5} + \frac{2\tau_2}{5} = C(\frac{1}{3})$.

TABLE III
THE QUERY TABLE FOR $M = 3$, $N = 2$, $\tau_2 = \frac{1}{3}$.

| Database 1 | Database 2 |
|---|---|
| $a_1, b_1, c_1$ | $a_2 + b_1$<br>$a_3 + c_1$<br>$b_2 + c_2$ |
| $a_4 + b_2 + c_2$ | |
| $a_5, b_3, c_3$ | $a_6 + b_3 + c_3$ |
| $a_7, b_4, c_4$ | $a_8 + b_4 + c_4$ |

## REFERENCES


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

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

[3] 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.

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

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

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

[7] H. Sun and S. A. Jafar. The capacity of symmetric private information retrieval. In *IEEE Globecom*, Dec 2016.

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

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

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

[11] 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.

[12] 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, Dec 2017.

[13] Q. Wang and M. Skoglund. Symmetric private information retrieval for MDS coded distributed storage. 2016. Available at arXiv:1610.04530.

[14] H. Sun and S. Jafar. Multiround private information retrieval: Capacity and storage overhead. 2016. Available at arXiv:1611.02257.

[15] 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.

[16] H. Sun and S. Jafar. Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al. 2017. Available at arXiv: 1701.07807.

[17] 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. 2017. Available at arXiv:1701.07636.

[18] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*. To appear. Also available at arXiv:1702.01739.

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

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

[21] K. Banawan and S. Ulukus. The capacity of private information retrieval from Byzantine and colluding databases. *IEEE Trans. on Info. Theory*. Submitted June 2017. Also available at arXiv:1706.01442.

[22] Q. Wang and M. Skoglund. Secure symmetric private information retrieval from colluding databases with adversaries. 2017. Available at arXiv:1707.02152.

[23] R. Tandon. The capacity of cache aided private information retrieval. 2017. Available at arXiv: 1706.07035.

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

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

[26] Y.-P. Wei, K. Banawan, and S. Ulukus. Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. 2017. Available at arXiv:1709.01056.

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

[28] 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.

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

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

[31] 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.

[32] Y.-P. Wei, K. Banawan, and S. Ulukus. Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits. *Jour. on Selected Areas in Communications*, 2017. To appear.

[33] 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.