Multi-Message 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 multi-message private information retrieval (MPIR) from N non-communicating replicated databases. In MPIR, the user is interested in retrieving P messages out of M stored messages without leaking the identity of the retrieved messages. The information-theoretic sum capacity of MPIR C_s^P is the maximum number of desired message symbols that can be retrieved privately per downloaded symbol. For the case $P \geq \frac{M}{2}$, we determine the exact sum capacity of MPIR as $C_s^P = \frac{1}{1+\frac{M-P}{PN}}$. For $P \leq \frac{M}{2}$, we develop lower and upper bounds for all M, P, N. These bounds match if the number of messages M is an integer multiple of the number of desired messages P, in which case, $C_s^P = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{M/P}}$. Our results indicate that joint retrieval of desired messages is more efficient than successive use of single-message retrieval schemes.**

I. Introduction

The privacy of the contents of the downloaded information from curious public databases has attracted considerable research within the computer science community [1], [2]. The problem is motivated by practical examples such as ensuring privacy of investors as they download records in a stock market [2], and privacy of inventors as they look up existing patents in a database. In the classical private information retrieval problem (PIR), a user requests to download a certain message (or file) from N non-communicating databases without leaking any information about the identity of the downloaded message. The contents of the databases are identical. The user performs this operation by preparing queries for all databases that are statistically independent of the message index. The databases respond with answer strings that are used to reconstruct the desired message. A trivial solution for this problem is to download the contents of all databases, which is inefficient from the retrieval rate perspective. The retrieval rate is the ratio of the number of desired message symbols to the number of total downloaded symbols. The capacity of PIR is the maximum retrieval rate over all PIR schemes.

The computer science formulation of this problem assumes that the messages are of length one. The metrics in this case are the download cost, i.e., the sum of lengths of the answer strings, and the upload cost, i.e., the size of the queries. The information-theoretic re-formulation of the problem considers arbitrarily large message sizes, and ignores the upload cost. In the information-theoretic formulation, [3] relates PIR to blind interference alignment. [4] derives the exact capacity of the classical PIR problem. Extensions of the basic PIR include: coded PIR [5], [6], PIR with colluding databases, robust PIR [7], symmetric PIR [8], coded symmetric PIR [9], and coded PIR with colluding databases [10].

This work was supported by NSF Grants CNS 13-14733, CCF 14-22111, CCF 14-22129, and CNS 15-26608.

In some applications, the user may be interested in retrieving multiple messages from the databases. One possible solution to this problem is to use single-message retrieval scheme in [4] successively. We show in this work that multiple messages can be retrieved more efficiently than retrieving them one-by-one in a sequence. This resembles superiority of joint decoding in multiple access channels over multiple simultaneous single-user transmissions [11]. To motivate, consider the example in [4, Section 4.3] where $M = 3$, $N = 2$, and the user is interested in retrieving $P = 1$ message. Here the optimal schemes retrieves 8 desired bits in 14 downloads, hence with a rate $4/7$. When the user wishes to retrieve $P = 2$ messages, if we use the scheme in [4] twice in a row, we retrieve 16 bits in 28 downloads, hence again a *sum rate* of $4/7$. Even considering the fact that the scheme in [4] retrieves 2 bits of the second message *for free* in downloading the first message, i.e., it actually retrieves 10 bits in 14 downloads, hence a sum rate of $5/7$, we show in this work that a better sum rate of $4/5$ can be achieved by joint retrieval of the messages.

Although there is a vast literature on classical PIR, only a few works exist in MPIR, such as: [12] which proposes a multi-block (multi-message) scheme; [13] extends the scheme in [1] to multiple blocks; and [14] designs an efficient non-trivial multi-query computational PIR protocol and gives an information-theoretic lower bound on the communication of any multi-query information retrieval protocol.

In this paper, we formulate the MPIR problem with non-colluding repeated databases from an information-theoretic perspective. Our goal is to characterize the sum capacity of the MPIR problem C_s^P, which is defined as the maximum ratio of the number of symbols from the P desired messages to the number of total downloaded symbols. When $P \geq \frac{M}{2}$, we determine the exact sum capacity of MPIR as $C_s^P = \frac{1}{1+\frac{M-P}{PN}}$. We use a novel achievable scheme which downloads MDS-coded mixtures of all messages. For the case of $P \leq \frac{M}{2}$, we derive lower and upper bounds that match when $\frac{M}{P}$ is an integer. In this case, the sum capacity is $C_s^P = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{M/P}}$. In other cases, although the exact capacity is still an open problem, we show numerically that the gap between the lower and upper bounds is monotonically decreasing in N and is upper bounded by 0.0082. The scheme for $P \leq \frac{M}{2}$ is inspired by [4]. The main difference of our scheme from [4]

is the number of stages required in each download round. Interestingly, the number of stages is related to the output of a P-order IIR filter [15]. The converse proof generalizes the proof in [4] for $P \geq 1$. We only provide sketches of the proofs here due to space limitations; proof details, extra examples and some figures can be found in the longer version [16].

II. Problem Formulation

Consider a classical PIR setting storing M messages. Each message is a vector $W_i \in \mathbb{F}_q^L$, $i \in \{1, \cdots, M\}$, whose elements are picked uniformly and independently from sufficiently large field $\mathbb{F}_q$. Denote the contents of W_m by the vector $[w_m(1), w_m(2), \cdots, w_m(L)]^T$. The messages are independent and identically distributed such that for $i \in \{1, \cdots, M\}$:

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

where subscript $n_1 : n_2$ represents $\{n_1, \cdots, n_2\}$ henceforth. The messages are stored within N non-colluding databases. Each database stores an identical copy of all M messages. In MPIR, the user aims to retrieve a subset of messages indexed by the index set $\mathcal{P} = \{i_1, \cdots, i_P\} \subseteq \{1, \cdots, M\}$, where $|\mathcal{P}| = P$, without leaking the privacy of the subset $\mathcal{P}$. We assume that the cardinality P is known by all databases. To retrieve $W_{\mathcal{P}} = (W_{i_1}, W_{i_2}, \cdots, W_{i_P})$, the user generates a query $Q_n^{[\mathcal{P}]}$ and sends it to the nth database. The user does not have any knowledge about the messages in advance, hence the messages and the queries are statistically independent,

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

The privacy is satisfied by ensuring statistical independence between the queries and the message-index set $\mathcal{P}$, i.e.,

$$I\left(Q_n^{[\mathcal{P}]}; \mathcal{P}\right) = 0, \quad n \in \{1, \cdots, N\} \tag{3}$$

The nth database responds with an answer string $A_n^{[\mathcal{P}]}$, which is a deterministic function of the queries and the messages,

$$H(A_n^{[\mathcal{P}]} | Q_n^{[\mathcal{P}]}, W_{1:M}) = 0 \tag{4}$$

The user should be able to reconstruct the messages $W_{\mathcal{P}}$ reliably from the collected answers from all databases given the queries. Hence, we write the reliability constraint as,

$$H\left(W_{\mathcal{P}} | A_{1:N}^{[\mathcal{P}]}, Q_{1:N}^{[\mathcal{P}]}\right) = 0 \tag{5}$$

We denote the retrieval rate of the ith message by R_i, where $i \in \mathcal{P}$, which is defined as the ratio of the length of message i to the total download cost of the message set $\mathcal{P}$,

$$R_i = \frac{H(W_i)}{\sum_{n=1}^{N} H\left(A_n^{[\mathcal{P}]}\right)} \tag{6}$$

The sum retrieval rate of $W_{\mathcal{P}}$ is given by,

$$\sum_{i=1}^{P} R_i = \frac{H(W_{\mathcal{P}})}{\sum_{n=1}^{N} H\left(A_n^{[\mathcal{P}]}\right)} = \frac{PL}{\sum_{n=1}^{N} H\left(A_n^{[\mathcal{P}]}\right)} \tag{7}$$

The sum capacity of the MPIR problem is $C_s^P = \sup \sum_{i=1}^{P} R_i$ over all retrieval schemes. We follow the information-theoretic assumptions of large enough message size, field size, and ignore the upload cost as in [4].

III. Main Results

Theorem 1 *For the MPIR problem with non-colluding and replicated databases, if the number of desired messages P is at least half of the number of overall stored messages M, i.e., if $P \geq \frac{M}{2}$, then the sum capacity is given by,*

$$C_s^P = \frac{1}{1 + \frac{M-P}{PN}} \tag{8}$$

We note that when $P = 1$, then the constraint of Theorem 1 is equivalent to $M = 2$, and the result in (8) reduces to the known result of [4] for $P = 1$, $M = 2$, which is $\frac{1}{1+\frac{1}{N}}$. We observe that the sum capacity in (8) is a strictly increasing function of N, and $C_s^P \to 1$ as $N \to \infty$. We also observe that the sum capacity in this regime is a strictly increasing function of P, and approaches 1 as $P \to M$.

Corollary 1 *For the MPIR problem with $P \geq \frac{M}{2}$, using the single-message retrieval scheme of [4] P times in a row, which achieves a sum rate of,*

$$R_s^{rep} = \frac{(N-1)(N^{M-1} + P - 1)}{N^M - 1} \tag{9}$$

is strictly sub-optimal with respect to the exact capacity in (8).

For the example in the introduction where $M = 3$, $P = 2$, $N = 2$, our scheme achieves a sum capacity of $\frac{4}{5}$ in (8), which is strictly larger than repeating-based sum rate of $\frac{5}{7}$ in (9).

Theorem 2 *For the MPIR problem with non-colluding and replicated databases, when $P \leq \frac{M}{2}$, the sum capacity is lower and upper bounded as,*

$$\underline{R}_s \leq C_s^P \leq \bar{R}_s \tag{10}$$

where the upper bound $\bar{R}_s$ is given by,

$$\bar{R}_s = \frac{1}{1 + \frac{1}{N} + \cdots + \frac{1}{N^{\lfloor \frac{M}{P} \rfloor - 1}} + \left(\frac{M}{P} - \lfloor \frac{M}{P} \rfloor\right) \frac{1}{N^{\lfloor \frac{M}{P} \rfloor}}} \tag{11}$$

$$= \frac{1}{\frac{1-(\frac{1}{N})^{\lfloor \frac{M}{P} \rfloor}}{1-\frac{1}{N}} + \left(\frac{M}{P} - \lfloor \frac{M}{P} \rfloor\right) \frac{1}{N^{\lfloor \frac{M}{P} \rfloor}}} \tag{12}$$

For the lower bound, define r_i as,

$$r_i = \frac{e^{j2\pi(i-1)/P}}{N^{1/P} - e^{j2\pi(i-1)/P}}, \quad i = 1, \cdots, P \tag{13}$$

where $j = \sqrt{-1}$, and denote γ_i, $i = 1, \cdots, P$, to be the solutions of the linear equations $\sum_{i=1}^{P} \gamma_i r_i^{-P} = (N-1)^{M-P}$, and $\sum_{i=1}^{P} \gamma_i r_i^{-k} = 0$, $k = 1, \cdots, P-1$, then $\underline{R}_s$ is given by,

$$\underline{R}_s = \frac{\sum_{i=1}^{P} \gamma_i r_i^{M-P} \left[\left(1 + \frac{1}{r_i}\right)^M - \left(1 + \frac{1}{r_i}\right)^{M-P}\right]}{\sum_{i=1}^{P} \gamma_i r_i^{M-P} \left[\left(1 + \frac{1}{r_i}\right)^M - 1\right]} \tag{14}$$

Corollary 2 *For the MPIR problem with non-colluding and replicated databases, if $\frac{M}{P}$ is an integer, then bounds in (10) match, and hence*

$$C_s^P = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{\frac{M}{P}}}, \quad \frac{M}{P} \in \mathbb{N} \tag{15}$$

We note that if $\frac{M}{P}$ is an integer then the sum capacity of MPIR is the same as the capacity of the single-message PIR [4] with the number of messages equal to $\frac{M}{P}$. We have numerically evaluated the difference between the upper bound and the achievable rate, i.e., $\bar{R}_s - \underline{R}_s$, and found that the largest difference is 0.0082 which results for $M=5$, $P=2$, $N=2$ case. We also note that this gap decreases in N.

IV. Achievablity Proof for the Case $P \geq \frac{M}{2}$

The scheme requires $L = N^2$, and is completed in 2 rounds. The main ingredient of the scheme is the MDS coding of the desired symbols and side information in the second round.

A. General Achievable Scheme

1) *Index preparation:* The user interleaves the contents of each message randomly and independently using random interleaver $\pi_m(.)$ which is known privately to the user only, i.e., $x_m(i) = w_m(\pi_m(i))$, $i \in \{1, \cdots, L\}$, where $X_m = [x_m(1), \cdots, x_m(L)]^T$ is the interleaved message.
2) *Round one:* As in [4], the user downloads one symbol from every message from every database, i.e., the user downloads $(x_1(n), x_2(n), \cdots, x_M(n))$ from the nth database. This implements *message symmetry*, *symmetry across databases*, and satisfies the privacy constraint.
3) *Round two:* The user downloads a coded mixture of new symbols from the desired messages and the undesired symbols downloaded from the other databases.
   a) The user picks an MDS generator matrix $\mathbf{G} \in \mathbb{F}_q^{P\times M}$, which has the property that every $P \times P$ submatrix is full-rank; e.g., the Reed-Solomon generator matrix over $\mathbb{F}_q$, where $q > M$ [17].
   b) The user picks uniformly and independently at random the permutation matrices $\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_{N-1}$ of size $M \times M$. These matrices shuffle the order of the columns of $\mathbf{G}$ to be independent of $\mathcal{P}$.
   c) At database 1, the user downloads an MDS-coded version of P new symbols from $\mathcal{P}$ and $M-P$ undesired symbols that are already known from database 2, i.e., download $\mathbf{G}\mathbf{S}_1 \begin{bmatrix}\mathbf{x}_d & \mathbf{x}_u\end{bmatrix}^T$, where $\mathbf{x}_d = \begin{bmatrix}x_{i_1}(N+1) & \cdots & x_{i_P}(N+1)\end{bmatrix}$, $\mathbf{x}_u = \begin{bmatrix}x_{j_1}(2) & x_{j_2}(2) & \cdots & x_{j_{M-P}}(2)\end{bmatrix}$, and $\{j_1, \cdots, j_{M-P}\}$ is the undesired index set. The user can cancel the undesired messages and be left with a $P \times P$ invertible system of equations. This implements *exploiting side information* as in [4].
   d) The user repeats the last step for each set of side information from database 3 to N, each with different permutation matrix.
   e) By *database symmetry*, the user repeats all steps of round two at each other database.

The verification of decodability and privacy constraints can be found in [16]; we give a simple example in the next subsection. To calculate the achievable rate: the user downloads P desired symbols out of M in the first round and $(N-1)P$ symbols in the second round, which are all desired, then

$$\sum_{i=1}^{P} R_i = \frac{N(P+P(N-1))}{N(M+P(N-1))} = \frac{1}{1+\frac{M-P}{PN}} \tag{16}$$

B. Example: $M=5$, $P=3$ Messages, $N=2$ Databases

Let $\mathcal{P} = \{1,2,3\}$, and a to e denote the contents of W_1 to W_5, respectively. We use 5×5 permutation matrix for $\mathbf{S}_1$ and $\mathbf{G}_{3\times 5}$ Reed-Solomon generator matrix over $\mathbb{F}_5$ as,

$$\mathbf{G}_{3\times 5} = \begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 1 & 2 & 3 & 4 & 0\\ 1 & 4 & 4 & 1 & 0\end{bmatrix} \tag{17}$$

The query table is shown in Table I below with permutation $2,5,1,3,4$. The reliability and privacy constraints are satisfied due to the MDS property that implies that any subset of 3 messages corresponds to a 3×3 invertible submatrix, if the remaining symbols are decodable from the other database. The achievable rate is $\frac{12}{16} = \frac{3}{4}$ which equals the sum capacity $\frac{1}{1+\frac{M-P}{PN}}$ in (8). This sum capacity strictly outperforms repetition-based achievable sum rate $\frac{18}{31}$ found from (9).

TABLE I
The query table for $M=5, P=3, N=2$.

Database 1	Database 2
a_1, b_1, c_1, d_1, e_1	a_2, b_2, c_2, d_2, e_2
$a_3+b_3+c_3+d_2+e_2$	$a_4+b_4+c_4+d_1+e_1$
$2a_3+c_3+3d_2+4e_2$	$2a_4+c_4+3d_1+4e_1$
$4a_3+c_3+4d_2+e_2$	$4a_4+c_4+4d_1+e_1$

V. Achievability Proof for the Case $P \leq \frac{M}{2}$

The scheme generalizes the ideas in [4]. Different than [4], our scheme uses unequal number of stages per each round of download. Interestingly, the number of stages at each round can be thought of as the output of an IIR filter.

A. General Achievable Scheme

1) *Index preparation:* This is the same as in Section IV-A.
2) *Number of stages:* We calculate the number of stages needed in each round. This can be done systematically by finding the output of the IIR filter characterized by:

$$y[n] = \frac{1}{N-1}\sum_{i=1}^{P}\binom{P}{i}y[n-i] \tag{18}$$

   with initial conditions $y[-P] = (N-1)^{M-P}$, $y[-P+1] = \cdots = y[-1] = 0$. The number of stages in the ith round is $\alpha_i = y[(M-P)-i]$. The connection between the number of stages and filter (18) can be found in [16].
3) *Initialization:* From database 1, the user downloads one symbol from each message in $\mathcal{P}$ and sets $i=1$.
4) *Message symmetry:* In the ith round, the user downloads sum of i terms from different symbols from the first

database. To satisfy the privacy constraint, the user downloads the remaining $\binom{M-P}{i}$ combinations of the ith round from the undesired symbol set $\bar{\mathcal{P}}$.

5) *Repetition of stages:* In the first database, the user repeats the operation at the ith round according to α_i. This results in downloading $\alpha_i\binom{M-P}{i}$ undesired equations, and $\alpha_i[\binom{M}{i}-\binom{M-P}{i}]$ desired equations.
6) *Symmetry across databases:* The user invokes symmetry across all databases to download $\alpha_i\binom{M-P}{i}$ new undesired equations, and $\alpha_i[\binom{M}{i}-\binom{M-P}{i}]$ new desired equations from each database. These undesired equations will be used as side information in subsequent rounds.
7) *Exploiting side information:* We form the desired equations as a sum of the desired symbols and the undesired symbols that can be decoded from other databases in the former $(i-1)$ rounds. If the user sums two or more symbols from $\mathcal{P}$, the user downloads one new symbol from one message only and the remaining symbols from $\mathcal{P}$ should be derived from other databases. Consequently, in the $(i+1)$th round, the user mixes one symbol of $\mathcal{P}$ with the sum of i undesired symbols from the ith round. This should be repeated for all $\binom{P}{1}$ desired symbols. Then, the user mixes each sum of 2 desired symbols with the sum of $(i-1)$ undesired symbols generated in the $(i-1)$th round. This should be repeated for all the $\binom{P}{2}$ combinations of the desired symbols, and so on.
8) *Repeating steps:* Repeat steps 4, 5, 6, 7 with setting $i=i+1$ until $i=M-P-1$.
9) *Last round:* We note that rounds $M-P+1$ to $M-1$ do not generate useful side information. Hence, $\alpha_{M-P+1}=\cdots=\alpha_{M-1}=0$. In the Mth round, which corresponds to summing all M messages, the user mixes P symbols from $\mathcal{P}$ (only one of them is new) and $M-P$ undesired symbol mixture that was generated in $(M-P)$th round.
10) *Shuffling the order of queries:* The order of the queries are shuffled uniformly, so that all queries are equally likely regardless of $\mathcal{P}$, hence privacy is guaranteed.

The verification of decodability and privacy constraints and calculation of the achievable scheme can be found in [16]; we give a simple example in the next sub-section.

B. Example: $M=4$, $P=2$ Messages, $N=2$ Databases

The number of stages needed per each round of download is characterized by the output of the IIR filter $y[n]=2y[n-1]+y[n-2]$ with the initial conditions $y[-2]=1, y[-1]=0$. Then, $\alpha_k=y[2-k]$. We calculate the output iteratively as,

$$\alpha_4=y[-2]=1, \quad \alpha_3=y[-1]=0 \tag{19}$$
$$\alpha_2=y[0]=2y[-1]+y[-2]=1 \tag{20}$$
$$\alpha_1=y[1]=2y[0]+y[-1]=2 \tag{21}$$

In round one, the user downloads individual symbols from all messages at each database. The user downloads a_1,b_1,c_1,d_1, and a_2,b_2,c_2,d_2 from database 1, since $\alpha_1=2$. This is repeated for database 2. In the second round, the user downloads the sum of each two symbols. At database 1, the undesired symbols from database 2 in the first round are exploited in the sum. The equations are either in the form of $a+(c,d)$, or $b+(c,d)$, each uses different stage from round 1. We note that the user downloads a_5+b_3 which uses b_3 as side information although W_2 is desired. Round two concludes with downloading c_5+d_5, c_6+d_6, which are used as side information in the last round. The third round is missing and the user proceeds to the fourth round directly. Here, the user downloads sum of four symbols using the side information downloaded in the second round and any decoded symbols for the other desired message. The query table for this example is in Table II below. The achievable rate in this case is $\frac{20}{30}=\frac{2}{3}=\frac{1}{1+\frac{1}{N}}$, which matches the upper bound. This rate outperforms the repetition-based rate which is $\frac{3}{5}$ from (9).

TABLE II
THE QUERY TABLE FOR $M=4, P=2, N=2$.

Database 1	Database 2
a_1,b_1,c_1,d_1	a_3,b_3,c_3,d_3
a_2,b_2,c_2,d_2	a_4,b_4,c_4,d_4
a_5+b_3	a_1+b_7
a_6+c_3	a_8+c_1
a_7+d_3	a_9+d_1
b_5+c_4	b_8+c_2
b_6+d_4	b_9+d_2
c_5+d_5	c_6+d_6
$a_3+b_{10}+c_6+d_6$	$a_{10}+b_1+c_5+d_5$

VI. Converse Proof

We follow the notations and simplifications of [4], [6] as,

$$\mathcal{Q}\triangleq\left\{Q_n^{[\mathcal{P}]}:\mathcal{P}\subseteq\{1,\cdots,M\},|\mathcal{P}|=P,n\in\{1,\cdots,N\}\right\} \tag{22}$$

Without loss of generality, the following hold for MPIR [4]:

1) We can assume that the MPIR scheme is symmetric.
2) To invoke the privacy constraint, we fix the response of one database to be the same irrespective of the desired set of messages $\mathcal{P}$, i.e., $A_n^{[\mathcal{P}_i]}=A_n$, where $|\mathcal{P}_i|=P$ for every $i\in\{1,2,\cdots,\beta\}$ for some $n\in\{1,\cdots,N\}$, and $\beta=\binom{M}{P}$. In the sequel, we fix the answer string of the first database, i.e., $A_1^{[\mathcal{P}]}=A_1,\ \forall\mathcal{P}$

The following lemma is a consequence of the symmetry.

Lemma 1 (Symmetry [4]) *For any* $W_{\mathcal{S}}=\{W_i:i\in\mathcal{S}\}$

$$H(A_n^{[\mathcal{P}]}|W_{\mathcal{S}},\mathcal{Q})=H(A_1^{[\mathcal{P}]}|W_{\mathcal{S}},\mathcal{Q}),\ n\in\{1,\cdots,N\} \tag{23}$$
$$H(A_1|\mathcal{Q})=H(A_n^{[\mathcal{P}]}|\mathcal{Q}),\quad n\in\{1,\cdots,N\},\forall\mathcal{P} \tag{24}$$

We construct the converse proof by induction over $\lfloor\frac{M}{P}\rfloor$ in a similar way to [4]. The base induction step is obtained for $1\leq\frac{M}{P}\leq 2$. We obtain an inductive relation for the case $\frac{M}{P}>2$. The converse proof extends the proof in [4] for $P>1$.

A. Converse Proof for the Case $1\leq\frac{M}{P}\leq 2$

The following lemma gives a lower bound on the interference within an answer string. The proof can be found in [16].

Lemma 2 (Interference Lower Bound) *For the MPIR problem with $P \geq \frac{M}{2}$, the uncertainty of the interfering messages $W_{\bar{\mathcal{P}}}$ within the answer string $A_1^{[\mathcal{P}]}$ is lower bounded as,*

$$H(A_1^{[\mathcal{P}]}|W_{\mathcal{P}}, \mathcal{Q}) \geq \frac{(M-P)L}{N} \tag{25}$$

Now, we prove the converse of the case $P \geq \frac{M}{2}$:

$$\begin{aligned}
ML =& H(W_{1:M}|\mathcal{Q}) && (26)\\
=& H(W_{1:M}|\mathcal{Q}) - H(W_{1:M}|A_{1:N}^{[\mathcal{P}_1]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]}, \mathcal{Q}) && (27)\\
=& I(W_{1:M}; A_{1:N}^{[\mathcal{P}_1]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]}|\mathcal{Q}) && (28)\\
=& H(A_{1:N}^{[\mathcal{P}_1]}, A_{1:N}^{[\mathcal{P}_2]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]}|\mathcal{Q}) && (29)\\
=& H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{Q}) + H(A_{2:N}^{[\mathcal{P}_2]}, \cdots, A_{2:N}^{[\mathcal{P}_\beta]}|A_{1:N}^{[\mathcal{P}_1]}, \mathcal{Q}) && (30)\\
\leq& \sum_{n=1}^{N} H(A_n^{[\mathcal{P}_1]}|\mathcal{Q}) + H(A_{1:N}^{[\mathcal{P}_2]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]}|W_{\mathcal{P}_1}, \mathcal{Q})\\
& - H(A_1|W_{\mathcal{P}_1}, \mathcal{Q}) && (31)
\end{aligned}$$

The answer strings $(A_{1:N}^{[\mathcal{P}_2]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]})$ are sufficient to construct all messages $W_{1:M}$ irrespective to $\mathcal{P}_1$. Therefore,

$$H(A_{1:N}^{[\mathcal{P}_2]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]}|W_{\mathcal{P}_1}, \mathcal{Q}) = (M-P)L \tag{32}$$

Consequently and using Lemma 2, the desired converse is,

$$\sum_{i=1}^{P} R_i \leq \frac{PL}{\sum_{n=1}^{N} H\left(A_n^{[\mathcal{P}_1]}|\mathcal{Q}\right)} \leq \frac{1}{1+\frac{M-P}{PN}} \tag{33}$$

B. Converse Proof for the Case $\frac{M}{P} > 2$

We first need the following lemma, whose proof is in [16].

Lemma 3 (Interference Conditioning) *The remaining uncertainty on the answer strings $A_{2:N}^{[\mathcal{P}_2]}$ after conditioning on the messages indexed by $\mathcal{P}_1$, such that $\mathcal{P}_1 \cap \mathcal{P}_2 = \phi$, $|\mathcal{P}_1| = |\mathcal{P}_2| = P$ is upper bounded by,*

$$H(A_{2:N}^{[\mathcal{P}_2]}|W_{\mathcal{P}_1}, \mathcal{Q}) \leq (N-1)[NH(A_1|\mathcal{Q}) - PL] \tag{34}$$

Now, we derive the inductive relation which is used together with the base induction step to derive the converse for $\frac{M}{P} > 2$. Wlog, $\mathcal{P}_1 = \{1, \cdots, P\}$, $\mathcal{P}_2 = \{P+1, \cdots, 2P\}$. From (29),

$$\begin{aligned}
ML =& H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{Q}) + H(A_{2:N}^{[\mathcal{P}_2]}|A_{1:N}^{[\mathcal{P}_1]}, \mathcal{Q})\\
& + H(A_{2:N}^{[\mathcal{P}_3]}, \cdots, A_{2:N}^{[\mathcal{P}_\beta]}|A_1, A_{2:N}^{[\mathcal{P}_1]}, A_{2:N}^{[\mathcal{P}_2]}, \mathcal{Q}) && (35)\\
\leq& NH(A_1|\mathcal{Q}) + H(A_{2:N}^{[\mathcal{P}_2]}|W_{1:P}, \mathcal{Q})\\
& + H(A_{2:N}^{[\mathcal{P}_3]}, \cdots, A_{2:N}^{[\mathcal{P}_\beta]}|A_1, W_{1:2P}, \mathcal{Q}) && (36)\\
=& NH(A_1|\mathcal{Q}) + H(A_{2:N}^{[\mathcal{P}_2]}|W_{1:P}, \mathcal{Q}) - H(A_1|W_{1:2P}, \mathcal{Q})\\
& + H(A_{1:N}^{[\mathcal{P}_3]}, \cdots, A_{1:N}^{[\mathcal{P}_\beta]}|W_{1:2P}, \mathcal{Q}) && (37)\\
\leq& NH(A_1|\mathcal{Q}) + (N-1)[NH(A_1|\mathcal{Q}) - PL]\\
& + (M-2P)L - H(A_1|W_{1:2P}, \mathcal{Q}) && (38)
\end{aligned}$$

where (38) follows from Lemma 3. Hence, (38) is written as,

$$NH(A_1|\mathcal{Q}) \geq \left(1+\frac{1}{N}\right) PL + \frac{1}{N} H(A_1|W_{1:2P}, \mathcal{Q}) \tag{39}$$

Now, (39) constructs an inductive relation, since evaluating $NH(A_1|W_{1:2P}, \mathcal{Q})$ is the same as $NH(A_1|\mathcal{Q})$ with $(M-2P)$ messages. We can write the induction hypothesis for MPIR with M messages as,

$$NH(A_1|\mathcal{Q}) \geq PL \left[\sum_{i=0}^{\lfloor \frac{M}{P} \rfloor - 1} \frac{1}{N^i} + \left(\frac{M}{P} - \left\lfloor \frac{M}{P} \right\rfloor \right) \frac{1}{N^{\lfloor \frac{M}{P} \rfloor}} \right] \tag{40}$$

We proceed with proving this relation for $M+1$ messages. From the induction hypothesis, we have

$$\begin{aligned}
& NH(A_1|W_{1:2P}, \mathcal{Q})\\
& \geq PL \left[\sum_{i=0}^{\lfloor \frac{M+1}{P} \rfloor - 3} \frac{1}{N^i} + \left(\frac{M+1}{P} - \left\lfloor \frac{M+1}{P} \right\rfloor \right) \frac{1}{N^{\lfloor \frac{M+1}{P} \rfloor - 2}} \right] && (41)
\end{aligned}$$

Substituting this in (39) concludes the induction argument. Thus, the upper bound for the MPIR problem is given by,

$$\sum_{i=1}^{P} R_i \leq \frac{PL}{NH(A_1|\mathcal{Q})} = \frac{1}{\sum_{i=0}^{\lfloor \frac{M}{P} \rfloor - 1} \frac{1}{N^i} + \left(\frac{M}{P} - \lfloor \frac{M}{P} \rfloor \right) \frac{1}{N^{\lfloor \frac{M}{P} \rfloor}}} \tag{42}$$

REFERENCES

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

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

[3] H. Sun and S. Jafar. Blind interference alignment for private information retrieval. 2016. Available at arXiv:1601.07885.

[4] H. Sun and S. Jafar. The capacity of private information retrieval. 2016. Available at arXiv:1602.09134.

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

[6] K. Banawan and S. Ulukus. The capacity of private information retrieval from coded databases. *IEEE Trans. on Info. Theory*. Submitted September 2016. Also available at arXiv:1609.08138.

[7] H. Sun and S. Jafar. The capacity of robust private information retrieval with colluding databases. 2016. Available at arXiv:1605.00635.

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

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

[10] R. Freij-Hollanti, O. Gnilke, and D. Karpuk C. Hollanti. Private information retrieval from coded databases with colluding servers. 2016. Available at arXiv:1611.02062.

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[12] R. Henry, Y. Huang, and I. Goldberg. One (block) size fits all: PIR and SPIR with variable-length records via multi-block queries. In *NDSS*, 2013.

[13] L. Wang, T. K. Kuppusamy, Y. Liu, and J. Cappos. A fast multi-server, multi-block private information retrieval protocol. In *IEEE Globecom*, Dec 2015.

[14] J. Groth, A. Kiayias, and H. Lipmaa. Multi-query computationally-private information retrieval with constant communication rate. In *International Workshop on Public Key Cryptography*. Springer, 2010.

[15] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Pearson Higher Education, 2010.

[16] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*. Submitted February 2017. Also available at arXiv:1702.01739.

[17] I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. *SIAM*, 8(2):300–304, 1960.