# On an Equivalence Between Single-Server PIR with Side Information and Locally Recoverable Codes

Swanand Kadhe, Anoosheh Heidarzadeh, Alex Sprintson, and O. Ozan Koyluoglu

***Abstract*—Private Information Retrieval (PIR) problem has recently attracted a significant interest in the information-theory community. In this problem, a user wants to privately download one or more messages belonging to a database with copies stored on a single or multiple remote servers. In the single server scenario, the user must have prior side information, *i.e.*, a subset of messages unknown to the server, to be able to privately retrieve the required messages in an efficient way.**

**In the last decade, there has also been a significant interest in Locally Recoverable Codes (LRC), a class of storage codes in which each symbol can be recovered from a limited number of other symbols. More recently, there is an interest in *cooperative* locally recoverable codes, *i.e.*, codes in which multiple symbols can be recovered from a small set of other code symbols.**

**In this paper, we establish a relationship between scalar-linear schemes for the single-server PIR problem and scalar-linear LRCs. In particular, we show the following results: (i) PIR schemes designed for retrieving a single message are 'equivalent' to classical LRCs; and (ii) PIR schemes for retrieving multiple messages are equivalent to cooperative LRCs. These equivalence results allow us to recover upper bounds on the download rate for PIR-SI schemes, and to obtain a novel rate upper bound on cooperative LRCs.**

## I. Introduction

The Private Information Retrieval (PIR) problem is one of the important problems in theoretical computer science [1]. The setting of the problem includes a client that needs to retrieve a message belonging to a database with copies stored on a single or multiple remote servers. The message needs to be retrieved by satisfying the privacy condition, which prevents the server from identifying the index of the retrieved message. The theoretical computer science community focused on the settings with small message sizes with the objective to minimize the total number of bits uploaded to and downloaded from the server (see [2]).

Starting with the seminal work of Sun and Jafar [3], the multiple-server PIR problem has received a significant attention from the information and coding theory community with breakthrough results in the past few years (see e.g., [4]–[7], and references therein). The information-theoretic approach has focused on a practical setting with large message sizes with the goal to minimize the ratio of the total number of downloaded bits to the message size.

Recently, Kadhe et al. [8], [9] considered the single-server PIR with Side Information (PIR-SI) problem, wherein the user knows a random subset of messages that is unknown to the server. It was shown that the side information enables the user to substantially reduce the download cost and still achieve information-theoretic privacy for the requested message. The multi-message extension of PIR-SI, which enables a user to privately download multiple messages from the server, was considered by Heidarzadeh et al. [10] as well as Li and Gastpar [11].

It is well-known in the theoretical computer science community that there is a strong relationship between PIR schemes and a class of error-correcting codes called *locally decodable codes* (LDCs) (see, *e.g.*, the surveys [2], [12]). LDCs allow one to decode an arbitrary message symbol from only a small subset of randomly chosen codeword symbols, even after a fraction of codeword symbols are corrupted by an adversary.

Continuing with this theme, in this paper, we show that single-server PIR-SI schemes are closely related to another class of codes with locality called *locally recoverable codes* (LRCs) [13]. LRCs are a class of erasure codes that enable one to recover an erased codeword symbol from only a small subset of other codeword symbols.

In particular, in an LRC with block-length $n$ and locality $r$, every codeword symbol can be reconstructed from at most $r$ other codeword symbols [13]. Rawat et al. [14], [15] extended the notion of local recovery to *cooperative local recovery*. Specifically, in an LRC with block-length $n$ and $(r, \ell)$-cooperative locality, every subset of $\ell$ codeword symbols can be reconstructed from at most $r$ other codeword symbols.

In this paper, we show that single-message PIR-SI schemes are related to LRCs, whereas multi-message PIR-SI schemes are related to cooperative LRCs. Detailed contributions are outlined in the following.

**Our Contributions:** In this work, we consider the single-server PIR-SI problem in which a user wishes to download $D$ messages from a database of $K$ messages (over a finite field $\mathbb{F}_q$), stored on a single remote server. The user has a random subset of $M$ messages, referred to as *side information*, whose identities are unknown to the server.

We focus our attention to the scalar-linear case wherein the answer from the server is of the form $EX$, where $X =$

S. Kadhe and O. O. Koyluoglu are with the Department of Electrical Engineering and Computer Science at University of California Berkeley, USA; emails: {swanand.kadhe, ozan.koyluoglu}@berkeley.edu.

A. Heidarzadeh and A. Sprintson are with the Department of Electrical and Computer Engineering at Texas A&M University, USA; emails:{anoosheh, spalex}@tamu.edu.

This work is supported in part by National Science Foundation grants CCF-1748585 and CNS-1748692.

This material is based upon work supported while Alex Sprintson was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

$[X_1 \cdots X_K]^T \in \mathbb{F}_q^K$ denotes the set of messages, and $E$ is a $T \times K$ matrix with entries over $\mathbb{F}_q$. When the user wishes to protect only the identities of the requested messages, we show the following results:

- Equivalence between single-message ($D = 1$) PIR with Side Information (SM-PIR-SI) schemes and LRCs:
  1) Any solution $E$ to an SM-PIR-SI problem is a parity check matrix of an LRC with block-length $K$ and locality $M$ (Theorem 1).
  2) Given a parity check matrix $H$ of an LRC with block-length $K$ and locality $M$, it is possible to construct an SM-PIR-SI scheme whose solution $E$ is a column-permutation of $H$ (Theorem 2).
- Equivalence between multi-message ($D \geq 2$) PIR with Side Information (MM-PIR-SI) schemes and cooperative LRCs:
  1) Any solution $E$ to a MM-PIR-SI problem is a parity check matrix of an LRC with block-length $K$ and $(M, D)$-cooperative locality (Theorem 3).
  2) Given a parity check matrix $H$ of an LRC with block-length $K$ and $(M, D)$-cooperative locality, it is possible to construct an MM-PIR-SI scheme whose solution $E$ is a column-permutation of $H$ (Theorem 4).
- As corollaries to Theorems 1 and 3, we derive upper bounds on the download rates for SM-PIR-SI problem (Corollary 1) and MM-PIR-SI problem (Corollary 3), respectively. In addition, as a corollary to Theorem 4, we derive a novel tight upper bound on the rate of a cooperative LRC for the regime $\ell > r$ (see Corollary 4 and Remark 2).

Next, we consider the case when the user wants to protect both the identities of the requested messages and that of the side-information, referred to as $(W, S)$-PIR-SI.[1] We show the following equivalence result:

- Equivalence between $(W, S)$-PIR-SI schemes and maximum distance separable (MDS) codes[2]:
  1) Any solution $E$ to a $(W, S)$-PIR-SI problem is a parity check matrix of an MDS code with block-length $K$ and dimension $M$ (Theorem 5).
  2) Given a parity check matrix $H$ of an MDS code with block-length $K$ and dimension $M$, it is possible to construct a $(W, S)$-PIR-SI scheme where $E = H$ (Theorem 6).

## II. Preliminaries

**Notation:** For a positive integer $K$, denote $\{1, \ldots, K\}$ by $[K]$. Let $\mathbb{F}_q$ denote the finite field of order $q$, where $q$ is a power of a prime. For a set $\{X_1, \ldots, X_K\}$ and a subset $S \subset [K]$, let $X_S = \{X_j : j \in S\}$. For a positive integer $P$, let $\mathbf{1}_P$ and $\mathbf{0}_P$, respectively, denote the all-one and all-zero row vectors of length $P$. Let $e_j$ be a unit vector of length $K$ such that its $j$-th entry is 1 and the other entries are 0. For a set $W = \{W_1, W_2, \ldots, W_D\} \subseteq [K]$, let

$$I_W = \begin{bmatrix} e_{W_1} \\ e_{W_2} \\ \vdots \\ e_{W_D} \end{bmatrix}.$$

For a $T \times K$ matrix $E \in \mathbb{F}_q^{T \times K}$, let $\langle E \rangle$ denote the row-space of $E$. For a subset $S \subset [K]$, let $E_S$ denote the $T \times |S|$ submatrix consisting of columns of $E$ indexed by $S$. For a vector $v$, let Supp $(v)$ denote the support of $\mathbf{v}$. For a subspace $\mathcal{C} \subset \mathbb{F}_q^K$, let $\mathcal{C}^\perp$ be its dual subspace.

### A. Single-Server PIR with Side Information

We briefly overview the single-server PIR with side information problem [8], [16] (see also [9]). Consider a server containing a database that consists of a set of $K$ messages $\boldsymbol{X} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_K]^T$, with each message being independently and uniformly distributed over $\mathbb{F}_q$. A user is interested in *privately* downloading $D$ ($1 \leq D \leq K$) messages $\boldsymbol{X}_W$ from the server for some $W \subseteq [K]$, $|W| = D$. We refer to $W$ as the *demand index set* and $\boldsymbol{X}_W$ as the *demand*. The user has the knowledge of a subset $\boldsymbol{X}_S$ of the messages for some $S \subset [K] \setminus W$, $|S| = M$, $M \leq K - D$. We refer to $S$ as the *side information index set* and $\boldsymbol{X}_S$ as the *side information*.

Let $\boldsymbol{W}$ and $\boldsymbol{S}$ denote the random variables corresponding to the demand and side information index sets, respectively. We assume that the side information index set $\boldsymbol{S}$ is distributed uniformly over over all subsets of $[K]$ of size $M$, i.e.,

$$p_{\boldsymbol{S}}(\boldsymbol{S} = S) = \frac{1}{\binom{K}{M}}, \quad S \subset [K], |S| = M. \tag{1}$$

Further, we assume that the demand index set $\boldsymbol{W}$ has the following conditional distribution given $S$:

$$p_{\boldsymbol{W}|\boldsymbol{S}}(\boldsymbol{W} = W | \boldsymbol{S} = S) = \begin{cases} \frac{1}{\binom{K-M}{D}}, & W \subseteq [K] \setminus S, |W| = D \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We assume that the server does not know the side information realization at the user and only knows the *a priori* distributions $p_{\boldsymbol{S}}(S)$ and $p_{\boldsymbol{W}|\boldsymbol{S}}(W|S)$.

To download the set of messages $\boldsymbol{X}_W$ given the side information $\boldsymbol{X}_S$, the user sends a query $Q^{[W,S]}$ from a finite alphabet $\mathcal{Q}$ to the server. The server responds to the query it receives with an answer $A^{[W,S]}$ over $\mathbb{F}_q^T$. Let $\boldsymbol{Q}^{[W,S]}$ and $\boldsymbol{A}^{[W,S]}$ be the corresponding random variables.

Our main focus is on non-interactive (single round), scalar-linear schemes. In particular, the answer $\boldsymbol{A}^{[W,S]}$ from the server can be specified as

$$\boldsymbol{A}^{[W,S]} = E\boldsymbol{X}, \tag{3}$$

where the matrix $E \in \mathbb{F}_q^{T \times K}$ depends on the query $\boldsymbol{Q}^{[W,S]}$. We refer to $E$ as a solution to the PIR-SI problem. Note that

[1] Here, $W$ denotes the demand index set and $S$ denotes the side information index set. We use the term $(W, S)$-PIR-SI to reflect the fact that the user wants to protect $(W, S)$ jointly.

[2] An MDS code can be considered as an LRC with locality $r = k$.

$T$, the number of rows of $E$, denotes the number of symbols downloaded from the server.

**Definition 1.** *[PIR-SI] Any scheme consisting of a query and an answer is referred to as the PIR with side information (PIR-SI) scheme if the query and answer satisfy the following two conditions.*

1. $W$-***privacy****: The server cannot infer any information about the demand index set from the query it receives i.e.,*
$$I\left(\boldsymbol{W}; \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]}\right) = 0. \tag{4}$$
2. $(W,S)$-***privacy****: The server cannot infer any information about the demand index set as well as the side information index set from the query it receives i.e.,*
$$I\left(\boldsymbol{W}, \boldsymbol{S}; \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]}\right) = 0. \tag{5}$$
3. ***Recoverability:*** *From the answer $A^{[W,S]}$ and the side information $X_S$, the user should be able to decode the desired set of messages $X_W$, i.e.,*
$$H\left(\boldsymbol{X}_W \mid \boldsymbol{A}^{[W,S]}, \boldsymbol{Q}^{[W,S]}, \boldsymbol{X}_S\right) = 0. \tag{6}$$

We refer to the case of $D = 1$ as single-message PIR-SI, while the case of $D \geq 2$ as multi-message PIR-SI.

The *rate* of a PIR-SI scheme is defined as the ratio of the message length ($\log q$ bits) to the total length of the answers (in bits) as follows:[3]

$$R = \frac{D \log q}{H\left(\boldsymbol{A}^{[\boldsymbol{W},\boldsymbol{S}]}\right)}. \tag{7}$$

The *capacity* of $W$-PIR-SI, denoted by $C_W$, is defined as the supremum of rates over all $W$-PIR-SI schemes for a given $K$ and $M$.

### B. Locally Recoverable Codes

Let $\mathcal{C}$ denote a linear $[n,k,d]_q$ code over $\mathbb{F}_q$ with block-length $n$, dimension $k$, and minimum distance $d$. For any codeword $\mathbf{c} \in \mathcal{C}$, $\mathbf{c}_i$ is said to be the $i$-th symbol of the codeword $\mathbf{c}$.

We say that the $i$-th symbol of a code $\mathcal{C}$ has locality $r$ if its value can be recovered from some other $r$ symbols of $\mathcal{C}$. The formal definition of locality is as follows (see [13]).

**Definition 2.** *[Locality] We say that the $i$-th coordinate of a code $\mathcal{C}$ has locality $r$ if there exists a set $R(i) \subset [n] \setminus \{i\}$, $|R(i)| \leq r$, such that, for every codeword $\mathbf{c} \in \mathcal{C}$, $\mathbf{c}_i = \sum_{l \in R(i)} \lambda_l \mathbf{c}_l$, where $\lambda_l \in \mathbb{F}_q \setminus \{0\}$, $\forall\, l \in R(i)$. We say that $R(i)$ is a repair group of the $i$-th coordinate and define $\Gamma(i) = \{\mathbf{c}_i \cup R(i)\}$.*

*We say that an $[n,k,d]_q$ code has (all-symbol) locality $r$ if each of its $n$ coordinates has locality $r$. An LRC with these parameters is referred to as an $(n,k,r)$ LRC.*

Equivalently, we say that the coordinate $i$ has locality $r$, if the dual code $\mathcal{C}^{\perp}$ contains a codeword $\mathbf{c}'$ of Hamming weight at most $r+1$ such that the $i$-th coordinate is in the support of $\mathbf{c}'$.

**Example 1.** *Let us consider a $(7,3)$ Simplex code $\mathcal{C}$, which is a dual of a $(7,4)$ Hamming code. In particular, $\mathcal{C}$ encodes three information symbols $\{a,b,c\}$ into seven symbols as $\{a,b,c,a+b,a+c,b+c,a+b+c\}$. It is easy to see that any symbol can be recovered from two other symbols. For instance, $a$ can be recovered from $b+c$ and $a+b+c$.*[4]

In [13], it is shown that the minimum distance $d_{min}(\mathcal{C})$ of an $(n,k,r)$ LRC $\mathcal{C}$ is upper bounded as

$$d_{min}(\mathcal{C}) \leq n - k - \left\lceil \frac{k}{r} \right\rceil + 2. \tag{8}$$

Further, the authors of prove that any systematic code with locality for information symbols that achieves equality in (8) must follow a specific *structure* [13]. We state below the structure theorem [13, Theorem 9], adapted to the form useful for our setup.

**Proposition 1.** *[13] Let $\mathcal{C}$ be an $(n,k,r)$ code, where $r \mid k$, $r < k$, and $n = k + k/r$. Then, for any $i,j \in [n]$, $i \neq j$, we have either $\Gamma(i) = \Gamma(j)$ or $\Gamma(i) \cap \Gamma(j) = \emptyset$.*

### C. Cooperative Locally Recoverable Codes

Let $\mathcal{C}$ denote a linear $[n,k,d]_q$ code over $\mathbb{F}_q$ with block-length $n$, dimension $k$, and minimum distance $d$. We say that the code has $(r,\ell)$-cooperative locality if for every codeword, it is possible to repair any $\ell$ symbols from at most $r$ other symbols. The formal definition is as follows (see [14]).

**Definition 3.** *We say that an $[n,k,d]$ code $\mathcal{C}$ has $(r,\ell)$-cooperative locality, if for any subset of $\ell$ coordinates $\Delta \subset [n]$, $|\Delta| = \ell$, there exists a set $\Gamma(\Delta) \subset [n]$ satisfying $\Delta \cap \Gamma(\Delta) = \emptyset$, $|\Gamma(\Delta)| \leq r$, such that, for every codeword $\mathbf{c} \in \mathcal{C}$, the symbols $\mathbf{c}_\Delta$ can be recovered using the symbols $\mathbf{c}_{\Gamma(\Delta)}$.*

An LRC with these parameters is referred to as an $(n,k,r,\ell)$ cooperative LRC. Note that when $\ell = n-k$ and $r = k$, then the above definition coincides with that of an MDS code.

In [14], it is shown that the minimum distance $d_{min}(\mathcal{C})$ of an $(n,k,r,\ell)$ cooperative LRC $\mathcal{C}$ is upper bounded as

$$d_{min}(\mathcal{C}) \leq n - k + 1 - \ell\left(\left\lceil \frac{k}{r} \right\rceil - 1\right). \tag{9}$$

## III. Equivalence Between Single-Message PIR-SI and Locally Repairable Codes

In this section, we show that a single-message PIR-SI scheme is *equivalent* to a locally recoverable code (LRC). In particular, we show that any solution to the single-message PIR-SI problem (SM-PIR-SI) must be a parity check matrix of

[3] We focus our attention to the download rate similar to [3]. This is because the download rate dominates the total communication rate when the message size is sufficiently large as compared to the size of a query.

[4] In fact, every symbol of the $(7,3)$ simplex code has three disjoint repair groups [17]. Further, note that, even though the $(7,3)$ simplex code is not optimal with respect to the distance upper bound in (8), it is optimal with respect to a field size dependent rate upper bound established in [17].

an LRC. Furthermore, we show that it is possible to construct a solution to the SM-PIR-SI problem using a parity check matrix of an LRC.

First, we establish the relation from a solution of the SM-PIR-SI problem to a parity check matrix of an LRC.

**Theorem 1.** *Any (scalar-linear) solution $E$ to the single-message PIR-SI problem must be a parity check matrix of an LRC with block length $K$ and locality $M$.*

*Proof:* First, we note that the following necessary condition is imposed by the privacy and recoverability conditions. For any query $Q^{[W,S]}$, the answer $E$ should satisfy the following necessary condition: for any candidate demand index $W' \in [K]$, there must exist a potential side information index set $S' \subseteq [K] \setminus W'$, $|S'| \leq M$ such that it is possible to recover $W'$ from $EX$ and $X_{S'}$. In other words, the following condition must hold:

$$e_{W'} \in \left\langle \begin{bmatrix} E \\ I_{S'} \end{bmatrix} \right\rangle. \tag{10}$$

If the aforementioned necessary condition does not hold, then the server will learn from $E$ that $W'$ is not the user's demand index. Indeed, since $E$ is the solution corresponding to the query $Q^{[W,S]}$, we have

$$\mathbb{P}\left(\boldsymbol{W} = W' \mid \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]} = Q^{[W,S]}\right) = 0, \tag{11}$$

which, in turn, implies that $I\left(\boldsymbol{W}; \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]}\right) > 0$. This violates the $W$-privacy condition (4).

The above condition (10) implies that for every $W' \in [K]$, $\langle E \rangle$ must contain a vector $\mathbf{v}$ of Hamming weight at most $M+1$ such that $W' \in \text{Supp}(\mathbf{v})$. According to Definition 2, $\langle E \rangle^{\perp}$ is an LRC with block-length $K$ and all-symbol locality $M$. ∎

Theorem 1 has the following two immediate implications. First, it allows us to construct a class of LRCs using solutions to the SM-PIR-SI problem. More specifically, given a solution $E$ to the SM-PIR-SI problem with $K$ messages and side information size $M$, one can easily obtain an LRC with block-length $K$ and locality $M$ as $\mathcal{C} = \langle E \rangle^{\perp}$.

Now, consider the Partition-and-Code scheme proposed in [9] for the SM-PIR-SI problem. Let $K = \alpha(M+1) + \beta$ for some $\alpha > 0$ and $0 \leq \beta < M+1$. In the P&C scheme, the user first randomly partitions the $K$ messages into $(\alpha+1)$ subsets, each of size at most $M+1$, such that one of the subsets is $W \cup S'$ for some $S' \subseteq S$. The user then asks the server to send the sum of messages in each subset, resulting in the download cost of $\alpha + 1$ symbols.

Note that the Partition-and-Code scheme yields a solution $E$ of size $(\alpha+1) \times K$ with the following form (up to column permutation):

$$E = \begin{bmatrix} \mathbf{1}_{M+1} & \mathbf{0}_{M+1} & \cdots & \mathbf{0}_{\beta} \\ \mathbf{0}_{M+1} & \mathbf{1}_{M+1} & \cdots & \mathbf{0}_{\beta} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{M+1} & \mathbf{0}_{M+1} & \cdots & \mathbf{1}_{\beta} \end{bmatrix}, \tag{12}$$

It is easy to verify that the corresponding LRC $\mathcal{C} = \langle E \rangle^{\perp}$ is a direct-sum of $\alpha+1$ single-parity check codes, each of length at most $M+1$. In other words, $\mathcal{C}$ is a simple LRC that partitions the message symbols into $\alpha+1$ subsets each of size at most $M+1$, and adds a parity check symbol for each subset.

Second, Theorem 1 enables us to use (8) to obtain an upper bound on the capacity of a (scalar-linear) single-message PIR-SI scheme. As we show next, the bound coincides with the upper bound derived in [8], [9].

**Corollary 1.** *The (scalar-linear) capacity of the single-message PIR-SI problem is upper bounded by $\lceil K/(M+1) \rceil^{-1}$.*

*Proof:* Let $E$ be a scalar-linear solution to the SM-PIR-SI problem. Let $\mathcal{C} = \langle E \rangle^{\perp}$. Suppose the minimum distance of $\mathcal{C}$ is $d$. Note that we must have $d \geq 2$. For, if $d = 1$, $E$ must contain a column of all zeros. Let $W'$ denote the index of this all-zero column. However, this implies that $X_{W'}$ cannot be the demand, and this will violate the privacy.[5] Now, since $\langle E \rangle^{\perp}$ is an LRC with block-length $n = K$, dimension $k = K - T$, and locality $r = M$ from Theorem 1, we have from (8) that

$$K \geq K - T + \left\lceil \frac{K-T}{M} \right\rceil - 2 + d.$$

After re-arranging, and noting that $d \geq 2$ and $T$ is an integer, we get

$$T \geq \left\lceil \frac{K}{M+1} \right\rceil.$$

As the messages are independent and uniformly distributed over $\mathbb{F}_q$, we have $H\left(\boldsymbol{A}^{[W,S]}\right) = T \log q$. The result then follows from (7). ∎

**Remark 1.** *The above result can be directly proved using an upper bound on the rate of an LRC with locality $r$ given as $r/(r+1)$ (see [18, Theorem 1]). It is interesting to note that [18, Theorem 1] uses acyclic induced subgraph argument similar to [8], [9].*

We say that a scalar-linear solution to SM-PIR-SI problem is an *optimal* solution, if $T = \lceil K/(M+1) \rceil$. Then, Proposition 1 implies the following structure on any optimal scalar-linear solution.

**Corollary 2.** *When $(M+1) \mid K$, any optimal scalar-linear solution $E$ to the PIR-SI problem can be converted to the following form using elementary row operations and column permutations:*

$$E = \begin{bmatrix} \times & \cdots & \times & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \times & \cdots & \times & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \times & \cdots & \times \end{bmatrix}, \tag{13}$$

*where $\times$ can be any non-zero element in $\mathbb{F}_q$, i.e., $\times \in \mathbb{F}_q \setminus \{0\}$, and the number of non-zero entries in each row is exactly $M+1$.*

[5]Note that here we are using the same argument as in the proof of Theorem 1 (cf. (19)).

Since the solution obtained using the partition-and-code scheme (cf. (12)) has the same form as (13), this shows the *uniqueness* of the solution obtained by the partition-and-code scheme. In other words, any optimal scalar-linear solution can be obtained from the partition-and-code solution using elementary row operations and column permutations.

Next, we establish the relation from a parity check matrix of an LRC to a solution of the SM-PIR-SI problem.

**Theorem 2.** *Let $H$ be a parity check matrix of an LRC with block length $K$ and locality $M$. Then, it is possible to construct a single-message PIR-SI scheme, such that the solution $E$ is a column-permutation of $H$.*

*Proof:* We present a constructive proof. In the rest of the proof, we consider all sets as ordered sets (with a natural ascending order). For a given $W$ and $S$, the user first finds a permutation $\pi$ on $[K]$ as follows. Choose an index $W'$ uniformly at random from $[K]$, independent of $W$ and $S$. Let $R(W')$ be a repair group of $W'$. If a coordinate has multiple repair groups, arbitrarily choose one repair group.[6] By the definition of locality, we have $|R(W')| \leq M$. For simplicity, we assume that every repair group of any symbol is of size $M$.[7] Let $R'(W')$ be a random permutation of $R(W')$. Let $P = [K] \setminus \{W \cup S\}$, and $P'$ be a random permutation of $[K] \setminus \{W' \cup R(W')\}$. Let $\pi$ be the permutation that maps $W$ to $W'$, $S$ to $R'(W')$, and $P$ to $P'$. The user sends $\pi$ as its query $Q^{[W,S]}$. The server then applies $\pi$ to the columns of $H$ to obtain $E$, *i.e.*, $E_i = H_{\pi(i)}$ for each $i \in [K]$, where $H_j$ is the $j$th column of $H$. Then, the server computes the answer as $EX$.

Next, we show that the above scheme satisfies the recoverablity and $W$-privacy conditions. Indeed, by the definition of locality for $W'$, $\langle H \rangle$ contains a vector whose support is $W' \cup R(W')$. Therefore, by the construction of $E$, $\langle E \rangle$ contains a vector whose support is $W \cup S$. Hence, the recoverability condition in (6) is satisfied. For the $W$-privacy, it suffices to show that, for any $W \in [K]$ and any permutation $\pi$,

$$\mathbb{P}\left(\boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]} = \pi \mid \boldsymbol{W} = W\right) = \frac{1}{K!}. \tag{14}$$

This is because using (14), it is easy to show that $\mathbb{P}\left(\boldsymbol{W} = W \mid \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]} = \pi\right) = \mathbb{P}(\boldsymbol{W} = W)$, from which the privacy condition (4) follows.

Now, we give a proof of (14). Observe that the query generation process first maps the demand index to a random index in $[K]$. Let $\boldsymbol{W}'$ denote that random index. Now, given a permutation $\pi$ on $[K]$ as a query, define the following events:

$$\boldsymbol{E}_1 = \left\{\boldsymbol{W}' = \pi(\boldsymbol{W})\right\}, \tag{15}$$

$$\boldsymbol{E}_2 = \left\{\boldsymbol{S} = \pi^{-1}(R(\boldsymbol{W}'))\right\}, \tag{16}$$

$$\boldsymbol{E}_3 = \{[K] \setminus \{\boldsymbol{W} \cup \boldsymbol{S}\} = \pi^{-1}([K] \setminus \{\boldsymbol{W}' \cup R(\boldsymbol{W}')\}\}. \tag{17}$$

[6]This arbitrary choice of a repair group for each coordinate is made *a priori*, and are known to the server as a part of the scheme.

[7]The arguments can be easily generalized to the case when some repair groups are smaller than $M$.

Then, for any $W \in [K]$ and a permutation $\pi$ on $[K]$, the probability of choosing $\pi$ as a query can be written as

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]} = \pi \mid \boldsymbol{W} = W\right) \\
&\quad \overset{(a)}{=} \mathbb{P}(\boldsymbol{E}_1 | \boldsymbol{W} = W) \times \mathbb{P}(\boldsymbol{E}_2 \mid \boldsymbol{E}_1, \boldsymbol{W} = W) \\
&\qquad \times \mathbb{P}(\boldsymbol{E}_3 \mid \boldsymbol{E}_2, \boldsymbol{E}_1, \boldsymbol{W} = W), \\
&\quad \overset{(b)}{=} \frac{1}{K} \times \frac{1}{M! \binom{K-1}{M}} \times \frac{1}{(K-1-M)!}, \\
&\quad = \frac{1}{K!},
\end{aligned}$$

where (a) follows from the query generation procedure, and (b) uses (1) and (2) to compute $\mathbb{P}(\boldsymbol{E}_2 \mid \boldsymbol{E}_1, \boldsymbol{W} = W)$. This completes the proof of (14), and concludes the proof. ∎

## IV. Equivalence Between Multi-Message PIR-SI and Cooperative Locally Recoverable Codes

In this section, we show that a multi-message PIR-SI scheme is a dual of a cooperative LRC, introduced in [14].

First, we show that any solution to the multi-message PIR-SI problem should be a parity check matrix of a code with cooperative locality.

**Theorem 3.** *Any (scalar-linear) solution $E$ to the multi-message PIR-SI problem with a demand set of size $D$ and a side information set of size $M$ must be a parity check matrix of an LRC with block length $K$ and $(M, D)$-cooperative locality.*

*Proof:* First, we note that the following necessary condition is imposed by the privacy and recoverability conditions. For any query $Q^{[W,S]}$, the answer $E$ should satisfy the following necessary condition: for every candidate demand index set $W' \in [K]$, $|W'| = D$, there must exist a potential side information index set $S' \subseteq [K] \setminus W'$, $|S'| \leq M$ such that it is possible to recover $X_{W'}$ from $EX$ and $X_{S'}$. In other words, the following condition must hold:

$$e_{i_j} \in \left\langle \begin{bmatrix} E \\ I_{S'} \end{bmatrix} \right\rangle, \quad \forall\, i_j \in W'. \tag{18}$$

If the aforementioned necessary condition does not hold, then the server will learn from $E$ that $W'$ is not the user's demand index. Since $E$ is the solution corresponding to the query $Q^{[W,S]}$, we have

$$\mathbb{P}\left(\boldsymbol{W} = W' \mid \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]} = Q^{[W,S]}\right) = 0,$$

which, in turn, implies that $I\left(\boldsymbol{W}; \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]}\right) > 0$. This violates the $W$-privacy condition (4). This violates the privacy condition (4).

The above condition (18) implies that for every subset $W' = \{i_1, i_2, \ldots, i_D\} \subseteq [K]$ of size $D$, $\langle E \rangle$ must contain $D$ vectors $v_1, v_2, \ldots, v_D$ such that $|\cup_{j=1}^{D} \text{Supp}(v_j)| \leq D + M$, and for each $1 \leq j \leq D$, $\text{Supp}(v_j) \cap W' = \{i_j\}$. It is easy to verify from Definition 3 that $\langle E \rangle^{\perp}$ is an $(M, D)$ cooperative LRC with block-length $K$. ∎

**Corollary 3.** *For $M \geq D$, the scalar-linear capacity of the multi-message PIR-SI problem is upper bounded by $D/\lceil DK/(M+D)\rceil$.*

*Proof:* Let $\mathcal{C} = \langle E \rangle^{\perp}$. Note that from Theorem 3, $\mathcal{C}$ must be a code with blocklength $K$ and $(M, D)$-cooperative locality. Using (9), it is shown in [15, Corollary 1] that the rate of a code with $(M, D)$-cooperative locality for $M \geq D$ is upper bounded as $M/(M+D)$. Therefore, we have $T/K \geq 1 - M/(M+D)$. This yields $T \geq \lceil DK/(D+M)\rceil$, which gives the capacity upper bound. ∎

Next, we show that it is possible to construct a solution to the multi-message PIR-SI problem using a parity check matrix of a cooperative locality code.

**Theorem 4.** *Let $H$ be a parity check matrix of an LRC with block-length $K$ and $(D, M)$-cooperative locality. Then, it is possible to construct a multi-message PIR-SI scheme, such that the solution $E$ is a column-permutation of $H$.*

*Proof:* The query generation process and the rest of the proof is similar to the proof of Theorem 1. ∎

**Corollary 4.** *For $\ell > r$, the rate of a linear $(n, k, r, \ell)$ cooperative LRC is upper bounded by $r/n$.*

*Proof:* Let $H$ be a parity check matrix of an $(n, k, r, \ell)$ cooperative LRC. From Theorem 3, $H$ is a solution (up to a column-permutation) of a multi-message PIR-SI problem such that $K = n$, $M = r$, and $D = \ell$. Now, in [16, Lemma 1], it is shown that, when $D > M$, the number of transmissions in any multi-message PIR-SI scheme is at least $K - M$. Therefore, we have $n - k \geq n - r$, from which the result follows. ∎

**Remark 2.** *Corollary 4 yields a better bound on the rate of a cooperative LRC for $\ell > r$ than [15, Corollary 1] given as $r/(r+\ell) + \ell^2/(nr)$. In fact, the rate bound is tight for $n > 2r$. This is because an $(n, r)$ MDS code trivially has $(r, \ell)$-cooperative locality for any $\ell \geq r$.*

Theorem 3 also enables us to obtain *computationally efficient* multi-message PIR-SI solutions. In particular, for $D \leq M$, the schemes in [16] (see also [19]) rely on generalized Reed-Solomon codes, and thus, require a finite field size at least $M + \lceil M/D \rceil$. On the other hand, it is possible to use constructions of cooperative LRCs to obtain PIR-SI schemes over smaller field size.[8] As an example, an $(n = 2^k - 1, k)$ simplex code has $(\ell+1, \ell)$-cooperative locality for any $1 \leq \ell \leq (n-1)/2$ (see [15]). Thus, it is possible to obtain multi-message PIR-SI solutions over the binary field when $K = 2^t - 1$ for a positive integer $t$, $1 \leq D \leq (K-1)/2$, and $M = D + 1$.

## V. Equivalence Between $(W, S)$-Private PIR-SI and MDS Codes

In this section, we show an equivalence between a solution to the $(W, S)$-PIR-SI problem and a maximum distance separable (MDS) code.

First, we establish the relation from a solution of the $(W, S)$-PIR-SI problem to a parity check matrix of an MDS code.

**Theorem 5.** *Any (scalar-linear) solution $E$ to the $(W, S)$-PIR-SI problem must be a parity check matrix of a $(K, M)$ MDS code.*

*Proof:* First, we note that the $(W, S)$-privacy condition implies the following necessary condition: for each message $X_i$ and every set $S_i \subseteq [K] \setminus \{i\}$ of size $M$, it is possible to recover $X_i$ from $EX$ and $X_{S_i}$. If this is not the case, then the server learns that the user cannot possess $X_{S_i}$ and demand any $X_W$ such that $i \in W$. Indeed, since $E$ is the solution corresponding to the query $Q^{[W,S]}$, we have

$$\mathbb{P}\left(\boldsymbol{S} = S_i, i \in \boldsymbol{W} \mid \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]} = Q^{[W,S]}\right) = 0, \tag{19}$$

which, in turn, implies that $I\left(\boldsymbol{W}, \boldsymbol{S}; \boldsymbol{Q}^{[\boldsymbol{W},\boldsymbol{S}]}\right) > 0$. This violates the $(W, S)$-privacy condition (5).

The aforementioned necessary condition implies that, for any set $S \subset [K]$ of size $M$, for every $i \in [K] \setminus S$, we should have

$$e_i \in \left\langle \begin{bmatrix} E \\ I_S \end{bmatrix} \right\rangle. \tag{20}$$

Equation (20), in turn, implies that the columns of $E$ in $[K] \setminus S$ must be linearly independent. Since this should hold for each subset $S \subset [K]$ of size $M$, we have that every subset of columns of $E$ of size $K - M$ are linearly independent. Thus, $E$ must be a parity check matrix of a $(K, M)$ MDS code. ∎

Next, we establish a relation from a parity check matrix of an MDS code to a solution of the $(W, S)$-PIR-SI problem. It is worth noting that the achievability schemes in [9], [16] for $(W, S)$-privacy are based on MDS codes.

**Theorem 6.** *Let $H$ be a parity check matrix of a $(K, M)$-MDS code. Then, $E = H$ is a solution to the $(W, S)$-PIR-SI problem.*

*Proof:* First, note that the scheme with $E = H$ is private, since the solution is independent of the particular realization of $\boldsymbol{W}$ and $\boldsymbol{S}$. As the server already knows the size of the side information index set, it does not get any other information about $\boldsymbol{W}$ and $\boldsymbol{S}$ from $E$.

To see the recoverability, note that any $K - M$ columns of $H$ are linearly independent. Thus, given the side information $X_S$ for any $S \subset [K]$ of size $M$, the user can recover all the messages $X_i$, $i \in [K] \setminus S$, including the demand message(s) $X_W$. ∎

## VI. Conclusion

The theoretical computer science community has established a strong relationship between PIR schemes and locally decodable codes. This paper extends this theme by establishing strong relationship between PIR schemes for a recently proposed single-server PIR with side information problem and locally recoverable codes. As corollaries to these results, we

[8] Note that small field size schemes obtained from cooperative LRCs may have smaller download rate than those in [16], [19].

obtain upper bounds on the download rate for PIR-SI schemes, and a novel rate upper bound on cooperative LRCs.

## REFERENCES


[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 965–981, 1998.

[2] S. Yekhanin, "Private information retrieval," *Communications of the ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[3] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. on Info. Theory*, vol. 63, no. 7, pp. 4075–4088, July 2017.

[4] ——, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. on Info. Theory*, vol. 64, no. 4, pp. 2361–2370, April 2018.

[5] R. Tajeddine and S. El Rouayheb, "Robust private information retrieval on coded data," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017.

[6] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *CoRR*, vol. abs/1702.01739, 2017. [Online]. Available: http://arxiv.org/abs/1702.01739

[7] ——, "The capacity of private information retrieval from coded databases," *IEEE Trans. on Info. Theory*, vol. 64, no. 3, pp. 1945–1956, March 2018.

[8] S. Kadhe, B. Garcia, A. Heidarzadeh, S. E. Rouayheb, and A. Sprintson, "Private information retrieval with side information: The single server case," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2017, pp. 1099–1106.

[9] ——, "Private information retrieval with side information," *CoRR*, vol. abs/1709.00112, 2017. [Online]. Available: http://arxiv.org/abs/1709.00112

[10] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *2018 56th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2018.

[11] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *2018 56th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2018.

[12] S. Yekhanin, "Locally decodable codes," in *Computer Science–Theory and Applications*. Springer, 2011, pp. 289–290.

[13] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *Information Theory, IEEE Transactions on*, vol. 58, no. 11, pp. 6925–6934, Nov 2012.

[14] A. S. Rawat, A. Mazumdar, and S. Vishwanath, "On cooperative local repair in distributed storage," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, March 2014, pp. 1–5.

[15] ——, "Cooperative local repair in distributed storage," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 107, Dec 2015.

[16] A. Heidarzadeh, B. Garcia, S. Kadhe, S. Y. E. Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," *CoRR*, vol. abs/1807.09908, 2018.

[17] V. R. Cadambe and A. Mazumdar, "Bounds on the size of locally recoverable codes," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 5787–5794, Nov 2015.

[18] I. Tamo and A. Barg, "A family of optimal locally recoverable codes," *Information Theory, IEEE Transactions on*, vol. 60, no. 8, pp. 4661–4676, Aug 2014.

[19] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," *CoRR*, vol. abs/1808.05797, 2018. [Online]. Available: http://arxiv.org/abs/1808.05797