Capacity-Achieving Private Information Retrieval Codes With Optimal Message Size and Upload Cost

Chao Tian, *Senior Member, IEEE*, Hua Sun, *Member, IEEE*, and Jun Chen, *Senior Member, IEEE*

***Abstract*— We propose a new capacity-achieving code for the private information retrieval (PIR) problem, and show that it has the minimum message size (being one less than the number of servers) and the minimum upload cost (being roughly linear in the number of messages) among a general class of capacity-achieving codes, and in particular, among all capacity-achieving linear codes. Different from existing code constructions, the proposed code is asymmetric, and this asymmetry appears to be the key factor leading to the optimal message size and the optimal upload cost. The converse results on the message size and the upload cost are obtained by an analysis of the information theoretic proof of the PIR capacity, from which a set of critical properties of any capacity-achieving code in the code class of interest is extracted. The symmetry structure of the PIR problem is then analyzed, which allows us to construct symmetric codes from asymmetric ones, yielding a meaningful bridge between the proposed code and existing ones in the literature.**

***Index Terms*— Capacity, private information retrieval.**

I. Introduction

The private information retrieval (PIR) problem addresses the following scenario. A total of K messages, each of L bits (or L symbols in some finite alphabet), are replicated at N servers. A user wishes to retrieve one of the messages without revealing the identity of the desired message to any individual server. To retrieve this message, the user generates one query for each server and each server will return an answer to the user, which depends on the stored messages and the received query. To ensure that each server learns nothing about which message is being retrieved in the information theoretic sense, each query must be marginally independent of the desired message index. The PIR problem admits a trivial solution, where the user simply requests all the messages. However, downloading everything obviously incurs too much communication cost, and PIR systems should be designed to communicate as efficiently as possible between the user and the servers.

In PIR systems, the most important measure of communication efficiency is the retrieval rate, defined as the number of message information bits that can be retrieved per bit of downloaded data from the servers. The maximum value of retrieval rate of a PIR system is referred to as its capacity, and the problem of characterizing the capacity is of fundamental importance in this setting. This problem was recently settled in [1] where the capacity was found to be

$$C = \left(1 + \frac{1}{N} + \frac{1}{N^2} + \ldots + \frac{1}{N^{K-1}}\right)^{-1}. \quad (1)$$

Other notable efforts and generalizations on the PIR problem in the coding and information theory literature can be found in, e.g., [2]–[22].

In the previous works where the PIR capacity is concerned, such as [1]–[4], [16]–[18], it is usually assumed that the message length L is sufficiently large (L is allowed to go to infinity). As a consequence, the corresponding code constructions in the literature are usually built by recursively layering message symbols and parity symbols using symmetry relations, resulting in codes that can only be applied on very long messages. The number of symbols in each message that a code can be applied on is sometimes referred to as the sub-packetization factor of the code. A smaller message length (sub-packetization factor) means that the code is more versatile, has less constraints, and may lead to more efficient implementation in practice. Another design factor that is of practical importance is the possible number of queries that each server needs to accommodate, i.e., the cardinalities of the query sets. Small cardinalities of the query sets imply that, firstly, the amount of information that needs to be sent to the servers (often referred to as the upload cost) is small during the query operation, and secondly, the servers only need to compute a small set of functions, both of which lead to simpler and more efficient system implementation.

In this work, we consider the construction of capacity-achieving PIR codes, and the contribution is three-fold.

1) Firstly, we propose a novel capacity-achieving PIR code construction, which has a small message size of $(N-1)$

Manuscript received August 11, 2018; revised March 4, 2019; accepted May 11, 2019. Date of publication May 22, 2019; date of current version October 18, 2019. The work of C. Tian was supported in part by the National Science Foundation under Grant CCF-18-32309. This paper was presented in part at the 2019 IEEE International Conference on Communications, Shanghai, China.

C. Tian is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: chao.tian@tamu.edu).

H. Sun is with the Department of Electrical Engineering, University of North Texas, Denton, TX 76203 USA (e-mail: hua.sun@unt.edu).

J. Chen is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: junchen@ece.mcmaster.ca).

Communicated by B. K. Dey, Associate Editor for Coding Techniques.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2918207

bits and a low upload cost of $N(K-1)\log_2 N$ bits. The coding alphabet of the proposed code can in fact be chosen to be any finite group or finite field, and particularly, to be the binary field which results in a binary linear code. Different from existing code constructions, the code proposed in this work is asymmetric, and this asymmetry appears to be the key to the significant reductions in the message size and the upload cost compared to other capacity-achieving codes.

2) Secondly, through a novel and delicate analysis of the converse proof of the PIR capacity, we identify a set of critical properties for a class of capacity-achieving codes on abelian groups, which we refer to as decomposable codes. Based on these properties we further derive novel converses for the message size and the upload cost. These converse bounds match the corresponding values in the proposed code, thus establishing the optimality of the proposed code construction in terms of the message size and the upload cost within the corresponding code classes, in particular, among capacity-achieving (either scalar or vector) linear codes.
3) Last but not least, the relation between symmetric PIR codes and asymmetric PIR codes is analyzed in details. The symmetry in this problem setting in fact includes three different components, namely symmetry in server indices, symmetry in file indices, and symmetry in the query answers. The analysis reveals certain fundamental structures in the problem setting that were largely overlooked in the existing literature. Using these symmetry relations, we show that the proposed code (in fact any asymmetric code) can be used to build more symmetric PIR codes, which offers a bridge between the proposed code and the existing code constructions.

It should be noted that efforts on the PIR problem in the theoretical computer science community focus on an alternative formulation where the message length L is assumed to be small and fixed, usually a single bit, but the number of messages and the number of servers are allowed to grow asymptotically [23]. In this setting, the overall communication cost can be viewed as consisting of upload cost and download cost, the latter of which is inversely proportional to the retrieval rate, and they can be traded off between each other. In fact, there exists a complex relation among the three quantities of the message size, the upload cost, and the download cost. For example, for the solution of retrieving everything, the upload cost is 0 as nothing needs to be sent to the servers (i.e., there is no randomness in the queries) and the download cost is KL as all messages are retrieved, and the message size can be $L=1$ bit each. Characterizing even the sum cost of upload and download for the case $L=1$ in the original theoretical computer science formulation appears to be intractable, and instead order-wise bounds have been investigated; there have been considerable efforts and many significant results after the ground-breaking work of [23]; see, e.g., [24]–[29]. Against this general backdrop, our result can be viewed as the first to precisely determine the relation among the message size, the upload cost, and the download cost, for the extreme point when the download cost is minimized.

The rest of the paper is organized as follows. Section II provides the problem definition and the necessary notation. Section III gives the proposed PIR code construction. The converse results on the minimum message size and the minimum upload cost are given in Section IV, and the symmetry relations are discussed in Section V. Finally, Section VI concludes the paper.

II. Model and Preliminaries

In this section, we provide a formal problem definition, as well as the necessary notation for subsequent discussions. A slightly different indexing method is chosen in this work: instead of the more conventional indexing of starting at 1, the indexing here starts at 0. This does not make any essential difference in the problem and the solution, however it will lead to notional simplicity when we present the new code construction.

A. System Model

The private information retrieval model can be formally described as follows. There are a total of N servers, each storing a copy of K messages, denoted as $W_0, W_1, \ldots, W_{K-1}$, respectively. A user wishes to retrieve a message W_k, $k \in \{0, 1, \ldots, K-1\}$, however at the same time wishes to keep the identity of the message being retrieved as a secret to any one of the servers. For this purpose, the user, using a random input F as the key, chooses a set of queries, $Q_{0:N-1} = (Q_0, Q_1, \ldots, Q_{N-1})$, one per server, and sends the queries to the servers. Server-n responds with an answer A_n, which depends on the messages stored at the server and the received query. Using all the answers $A_{0:N-1} = (A_0, A_1, \ldots, A_{N-1})$ from all the servers, together with the values of F and k, the user then reconstructs W_k. The privacy requirement stipulates that at each server, the probability distributions on the allowed queries are identical for all the messages, thus the server cannot learn any information regarding which message is being requested.

We now give a more mathematically precise description of the problem. Denote the set of possible queries for server-n as $\mathcal{Q}_n$, and denote its cardinality as $|\mathcal{Q}_n|$. The cardinality of a set $\mathcal{A}$ will be similarly denoted as $|\mathcal{A}|$ in the rest of the paper. Assume that the random key F is uniformly distributed on a certain finite set $\mathcal{F}$. Moreover, a message W_k consists of L symbols, each symbol belonging to a finite alphabet $\mathcal{X}$; in particular, for messages in computer systems, we usually use $\mathcal{X} = \{0, 1\}$. The messages are mutually independent, each of which is uniformly distributed on $\mathcal{X}^L$. We further allow the query answers to be represented as a variable-length vector, whose elements are in the finite alphabet $\mathcal{Y}$, though our code construction will eventually only use $\mathcal{Y} = \mathcal{X}$.

Definition 1. *An N-server private information retrieval (PIR) code for K messages, each of L-symbols in the alphabet $\mathcal{X}$, consists of*

1) N query functions:

$$\phi_n : \{0, 1, \ldots, K-1\} \times \mathcal{F} \to \mathcal{Q}_n, \qquad n \in \{0, , 1, ..., N-1\}, \tag{2}$$

i.e., the user chooses the query $Q_n^{[k]} = \phi_n(k, \mathsf{F})$ for server-n, using the index of the desired message and the random key F;

2) *N answer length functions:*

$$\ell_n : \mathcal{Q}_n \to \{0, 1, 2, \ldots\}, \quad n \in \{0, 1, ..., N-1\}, \tag{3}$$

i.e., the length of the answer at each server, a non-negative integer, is a deterministic function of the query, but not the particular realization of the messages;

3) *N answer functions:*

$$\varphi_n : \mathcal{Q}_n \times \mathcal{X}^{KL} \to \mathcal{Y}^{\ell_n}, \quad n \in \{0, 1, ..., N-1\}, \tag{4}$$

where $\ell_n = \ell_n(q_n)$ with $q_n \in \mathcal{Q}_n$ being the (random) query for server-n, $\mathcal{Y}$ is the coded symbol alphabet, and in the sequel we shall write the query answer as $A_n^{[k]} \triangleq \varphi_n(Q_n^{[k]}, W_{0:K-1})$ when the message index k is relevant;

4) *A reconstruction function using the answers from the servers together with the desired message index and the random key:*

$$\psi : \prod_{n=0}^{N-1} \mathcal{Y}^{\ell_n} \times \{0, 1, ..., K-1\} \times \mathcal{F} \to \mathcal{X}^L, \tag{5}$$

i.e., $\hat{W}_k = \psi(A_{0:N-1}^{[k]}, k, \mathsf{F})$ is the retrieved message.

These functions should satisfy the following two requirements:

1) ***Correctness:*** *For any $k \in \{0, 1, ..., K-1\}$, $\hat{W}_k = W_k$.*
2) ***Privacy:*** *For every $k, k' \in \{0, 1, ..., K-1\}$, $n \in \{0, 1, ..., N-1\}$, and $q \in \mathcal{Q}_n$,*

$$\mathbf{Pr}(Q_n^{[k]} = q) = \mathbf{Pr}(Q_n^{[k']} = q). \tag{6}$$

The correctness condition here requires that the reconstructed message as a random variable is the same as the requested message, and it thus inherently requires that for any realization of F, the equality must hold. It is in fact without loss of generality to restrict $\mathcal{F}$ and $\mathcal{Q}_n$'s to be certain finite sets of integers, however, we allow them to be more general sets, which will facilitate describing more concisely the proposed PIR code construction.

It is also worth noting that the alphabet $\mathcal{Y}$ in the problem definition may be an abstract finite set, with no further structure assigned to it. However, for any such a finite set, we can establish a bijective mapping between $\mathcal{Y}$ and the set $\{0, 1, \ldots, |\mathcal{Y}|-1\}$. By further enforcing an operation between any two elements in the latter set (for example, modulo $|\mathcal{Y}|$ addition), the set $\mathcal{Y}$ can also be assigned an operation through homomorphism. In other words, any abstract set $\mathcal{Y}$ can always be viewed as a finite group, however requiring $\mathcal{Y}$ to be a finite group in the problem definition is unnecessary.

B. Two General Code Classes

We next define precisely the code classes in which the optimality of our proposed code construction is established. These definitions are technical, and the readers may wish to skip them at the initial read and simply consider the more restricted code class of vector linear codes on a finite field, without materially jeopardizing understanding the code construction in Section III. These two definitions only become important in Section IV when the optimality of the proposed code construction is established.

Definition 2. *A PIR code is called decomposable, if $\mathcal{Y}$ is a finite abelian group, and for each fixed $n \in \{0, 1, \ldots, N-1\}$ and $q \in \mathcal{Q}_n$, the answer function $\varphi_n(q, W_{0:K-1})$ can be written in the form*

$$\begin{aligned}\varphi_n(q, W_{0:K-1}) \\ = \left(\varphi_{n,0}^{(q)}(W_{0:K-1}), \varphi_{n,1}^{(q)}(W_{0:K-1}), \ldots, \varphi_{n,\ell_n-1}^{(q)}(W_{0:K-1})\right),\end{aligned} \tag{7}$$

where

$$\begin{aligned}\varphi_{n,i}^{(q)}(W_{0:K-1}) \\ = \varphi_{n,i,0}^{(q)}(W_0) \oplus \varphi_{n,i,1}^{(q)}(W_1) \oplus \ldots \oplus \varphi_{n,i,K-1}^{(q)}(W_{K-1}), \\ i \in \{0, 1, \ldots, \ell_n - 1\},\end{aligned} \tag{8}$$

where $\oplus$ represents addition in the finite group $\mathcal{Y}$, and each $\varphi_{n,i,k}^{(q)}$ is a mapping $\mathcal{X}^L \to \mathcal{Y}$.

The terminology "decomposable" comes from (8) which restricts each coded symbol to be a summation (in the abelian group) of the component functions on the individual messages. Let us consider an example where the two messages, each of a single symbol, are in certain ring $(\mathcal{Y}, \oplus, \otimes)$, and the resulting answer for $n = q = 0$ is

$$\begin{aligned}&\varphi_0(0, (W_0, W_1)) \\ &= (\varphi_{0,0}^{(0)}(W_0, W_1), \varphi_{0,1}^{(0)}(W_0, W_1), \varphi_{0,2}^{(0)}(W_0, W_1)) \\ &= \big((W_0 \otimes W_0) \oplus \alpha, (W_0 \otimes W_0) \oplus (W_1 \otimes W_1) \oplus \alpha, \\ &\qquad (W_1 \otimes W_1) \oplus \alpha\big),\end{aligned}$$

where α is an element of the ring. This code belongs to the code class of decomposable codes, but it is clearly not linear. Note that in the component function $\varphi_{0,1}^{(0)}(W_0, W_1)$, we have

$$\begin{aligned}\varphi_{0,1}^{(0)}(W_0, W_1) &= \varphi_{0,1,0}^{(0)}(W_0) \oplus \varphi_{0,1,1}^{(0)}(W_1) \\ &= (W_0 \otimes W_0) \oplus ((W_1 \otimes W_1) \oplus \alpha).\end{aligned}$$

Definition 3. *If a decomposable PIR code has the property that any component function $\varphi_{n,i,k}^{(q)}$ in (8) either satisfies the condition*

$$\begin{aligned}&\left|\left\{w \in \mathcal{X}^L : \varphi_{n,i,k}^{(q)}(w) = g\right\}\right| \\ &\quad = \left|\left\{w \in \mathcal{X}^L : \varphi_{n,i,k}^{(q)}(w) = g'\right\}\right|, \quad \forall g, g' \in \mathcal{Y},\end{aligned} \tag{9}$$

or it maps everything to the same value, i.e.,

$$\varphi_{n,i,k}^{(q)}(w) = \varphi_{n,i,k}^{(q)}(w'), \quad \forall w, w' \in \mathcal{X}^L, \tag{10}$$

then the PIR code is called uniformly decomposable.

A uniformly decomposable PIR code has the property that the decomposed message mappings $\varphi_{n,i,k}^{(q)}$ will preserve a uniform probability distribution on the coded symbol alphabet, unless the induced random variable is in fact deterministic. The notion of decomposable codes considerably generalizes the notion of linear codes. In particular, linear codes on finite fields are uniformly decomposable, and linear codes defined on modules over a ring [30], [31] are decomposable (and some are uniformly decomposable); it also naturally includes codes defined on cosets of a binary lattice and some nonlinear codes. In Section IV, we establish general outer bounds for decomposable codes, which imply that the proposed code is optimal in the corresponding code classes, and particularly, it is optimal among all linear codes.

Decomposable codes can be simply represented as

$$\varphi_n(q, W_{0:K-1}) = W_{0:K-1} \cdot G_n^{(q)}, \tag{11}$$

where $W_{0:K-1}$ is viewed as a length-K vector whose components are in the alphabet $\mathcal{X}^L$, and $G_n^{(q)}$ is a matrix of dimension $K \times \ell_n$ whose elements $G_{n,i,k}^{(q)}$ are functions $\mathcal{X}^L \to \mathcal{Y}$ with the "$\cdot$" operation between W_k and the matrix element $G_{n,i,k}^{(q)}$ defined as

$$W_k \cdot G_{n,i,k}^{(q)} \triangleq \varphi_{n,i,k}^{(q)}(W_k). \tag{12}$$

Consider another example (a uniformly decomposable code) where $K = 3$, $L = 1$, and the answer for query $q = 0$ for server-0 is

$$\varphi_0(q = 0, W_{[0:2]}) = \varphi_{0,0}^{(0)}(W_0, W_1, W_2) = W_0 \ominus W_2, \tag{13}$$

where $\mathcal{X} = \mathcal{Y}$ is the finite group $\{0, 1, 2, 3\}$ with $\oplus$ and $\ominus$ being the modulo-4 addition and subtraction. Then we have

$$\begin{aligned}
&\varphi_0(q = 0, W_{[0:2]})\\
&= \varphi_{0,0,0}^{(0)}(W_0) \oplus \varphi_{0,0,1}^{(0)}(W_1) \oplus \varphi_{0,0,2}^{(0)}(W_2)\\
&= (W_0) \oplus (0) \oplus (\ominus W_2)\\
&= [W_0, W_1, W_2] \cdot G_0^{(0)}\\
&= [W_0, W_1, W_2] \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.
\end{aligned} \tag{14}$$

In the matrix representation, $G_0^{(0)} = [1, 0, -1]^t$, where 1 stands for the identity function, 0 for the all zero function, and -1 for the negation function.

For linear codes defined on a finite field $\mathcal{X}$, the function $\varphi_{n,i,k}^{(q)}(W_k)$ is the inner product between the length-L message vector W_k, and a fixed length-L coding coefficient vector in the same finite field. In this case, $W_{0:K-1}$ can be alternatively written as a length-KL vector in the alphabet $\mathcal{X}$, and the matrix $G_n^{(q)}$ can be further expanded as a $KL \times \ell_n$ matrix whose elements are also in $\mathcal{X}$, and the finite field addition and multiplication will be used in the matrix multiplication. Such a matrix $G_n^{(q)}$ is in fact simply the familiar generator matrix of (vector) linear codes [32].

It should be noted that most converse results on linear codes in the literature have been established by deriving relations among the ranks of the coding matrices, whereas our converse proof in Section IV is information-theoretic in nature. The benefit of our approach is that it allows us to derive converse bounds for the general class of codes in a single framework.

C. Performance Metrics

The performance of an N-server PIR code can be measured using the following three quantities:

1) The retrieval rate
$$R \triangleq \frac{L \log_2 |\mathcal{X}|}{\log_2 |\mathcal{Y}| \sum_{n=0}^{N-1} \mathbb{E}(\ell_n)}, \tag{15}$$
which is the number of bits of desired message information that can be privately retrieved per bit of downloaded data. This quantity should be maximized, because higher rate implies fewer number of bits to be downloaded when retrieving a message. It was shown in [1] that the retrieval rate is upper-bounded by the PIR capacity C, i.e., $R \leq C$, which is a function of (N, K) given in (1).
2) The message size $L \log_2 |\mathcal{X}|$, which is the number of bits to represent each individual message. This quantity should also be minimized, because PIR schemes for a larger message size can be constructed by concatenating multiple schemes for a smaller message size, but not vice versa. Therefore, in practical applications, a smaller message size implies a more versatile code design; similar considerations of reducing the sub-packetization factor also exist for the regenerating code problem, e.g., [33]–[37], and the coded caching problem, e.g., [38]–[40]. Note that we refer to the parameter L as the message length, while the definition of the message size also takes into account the alphabet size $|\mathcal{X}|$.
3) The upload cost
$$\sum_{n=0}^{N-1} \log_2 |\mathcal{Q}_n|, \tag{16}$$
which is the number of bits required to send the queries to the servers. This quantity should be minimized for an efficient PIR code, since a smaller upload cost implies less user-to-server communication, and simpler server functions as mentioned earlier.

The code construction we shall propose in this work is optimal in the following senses:

1) It is capacity-achieving $R = C$, i.e., the retrieval rate is optimal;
2) It has the smallest, thus optimal, message size among all capacity-achieving uniformly decomposable codes;
3) It has the smallest, thus optimal, upload cost among all capacity-achieving decomposable codes.

III. A New Capacity-Achieving PIR Code

In this section, we provide the details of the proposed codes. Before presenting the code construction under general parameters, we provide a motivating example for the case of $(N, K) = (2, 2)$.

A. A Motivating Example $(N, K) = (2, 2)$

Let us consider two different codes for the $(N, K) = (2, 2)$ case, both of which are capacity-achieving.

1) A simple but new code is as given in Table I, where 0 in the transmission means no symbol is transmitted.

TABLE I

ANSWERS FOR MESSAGE A AND B WHEN $(N, K) = (2, 2)$ IN A SIMPLE NEW CODE

	Requesting A		Requesting B	
	Server-1	Server-2	Server-1	Server-2
$\mathsf{F} = 0$	0	a	0	b
$\mathsf{F} = 1$	$a + b$	b	$a + b$	a

Here the two messages are $A = (a)$ and $B = (b)$, each of which has only 1 symbol. The random key is binary, uniformly distributed in the key set $\mathcal{F} = \{0, 1\}$. It can be seen that the expected download cost is $0.5 + 1 = 1.5$, and thus the rate is $2/3$, which achieves the capacity.

2) In comparison, in the code constructed in [1], the two messages $A = (a_1, a_2, a_3, a_4)$ and $B = (b_1, b_2, b_3, b_4)$ each have 4 symbols. The random key set $\mathcal{F}$ is the collection of permutation $\pi(\cdot)$, which is used to select the one-to-one correspondence between $\{1, 2, 3, 4\}$ and $\{\square, \diamondsuit, \clubsuit, \heartsuit\}$. With this correspondence determined, the code is as given in Table II. The download cost is 6 symbols, and the rate is thus $2/3$.

It is observed that in the simple new code, the message sizes for different queries are allowed to vary, while the code constructed in [1] uses answers of the same length, regardless of the key realization and the query.

B. The New PIR Code

The code we propose, which will be referred to as the N-ary-indexed PIR code, has the following parameter:

$$L = N - 1. \tag{17}$$

The query sets at the servers are defined as

$$\mathcal{Q}_n \triangleq \left\{ q_{n,0:K-1} \in \{0, \dots, N-1\}^K \middle| \left(\sum_{k=0}^{K-1} q_{n,k} \right)_N = n \right\}, \quad n \in \{0, 1, \dots, N-1\}, \tag{18}$$

where $(\cdot)_N$ means the modulo N operation, and for convenience we have written $(q_{n,0}, q_{n,1}, \dots, q_{n,K-1})$ as $q_{n,0:K-1}$. In other words, the queries are length-K vectors, whose elements are in the set $\{0, 1, \dots, N-1\}$; the query set for server-n is all such vectors whose elements sum up to n under modulo N. It is easy to see that

$$|\mathcal{Q}_n| = N^{K-1}, \qquad n \in \{0, 1, \dots, N-1\}, \tag{19}$$

since the first $K-1$ digits of the query, i.e., $(q_{n,0}, q_{n,1}, \dots, q_{n,K-2})$, can take any value in the set $\{0, 1, \dots, N-1\}^{K-1}$, however, for a fixed server-n, the last digit is then uniquely determined in the set $\mathcal{Q}_n$.

The sample space of the random key is defined as $\mathcal{F} = \{0, 1, \dots, N-1\}^{K-1}$, and thus the random key F can be written as

$$\mathsf{F} = (\mathsf{F}_0, \mathsf{F}_1, \dots, \mathsf{F}_{K-2}), \tag{20}$$

where $\mathsf{F}_k \in \{0, 1, \dots, N-1\}$, $k = 0, 1, \dots, K-2$. Each message W_k, $k \in \{0, 1, \dots, K-1\}$, is a length-$L$ vector and thus by pre-pending a dummy variable $W_{k,0} \triangleq 0$, can be written as

$$W_k = (W_{k,0}, W_{k,1}, \dots, W_{k,N-1}), \tag{21}$$

where $(W_{k,1}, \dots, W_{k,N-1})$ is the true information payload of the message W_k. Without loss of generality, we shall assume $\mathcal{X} = \{0, 1, \cdots, |\mathcal{X}| - 1\}$, which, together with the modulo addition operation $\oplus$, forms a finite group $(\mathcal{X}, \oplus)$. This includes the particularly attractive choice of $\mathcal{X} = \{0, 1\}$, where each symbol is a bit and the group addition is simply binary XOR, and in this case, the binary group can also be viewed as the binary field.

We next provide the precise forms of the four coding functions with the parameter and the relevant sets defined above, which constitute the proposed code:

1) The query function ϕ_n for $n \in \{0, 1, \dots, N-1\}$ is

$$\begin{aligned} Q_n^{[k]} &= \phi_n(k, \mathsf{F}) \\ &= \left(\mathsf{F}_0, \mathsf{F}_1, \mathsf{F}_{k-1}, \left(n - \mathsf{F}_k^*\right)_N, \mathsf{F}_k, \dots, \mathsf{F}_{K-2}\right), \end{aligned} \tag{22}$$

where $\mathsf{F}_k^* \triangleq \left(\sum_{i=0}^{K-2} \mathsf{F}_i \right)_N$. In other words, all digits except the k-th digit in the query vector are copied from F, while the k-th digit is set to match the unique value in the query set at this server. This query can be equivalently written as $Q_{n,0:K-1}^{[k]}$ since it is a length-K vector.

2) The answer length function ℓ_n for $n \in \{0, 1, \dots, N-1\}$ is

$$\ell_n(n, q) = \begin{cases} 0 & (n, q) = (0, (0, 0, \dots, 0)) \\ 1 & \text{otherwise} \end{cases}. \tag{23}$$

In other words, there is only one query at the 0-th server that will induce $\ell_0 = 0$, while all other queries at all other servers will induce an answer of a single symbol.

3) The answer function φ_n for $n \in \{0, 1, \dots, N-1\}$ is

$$\begin{aligned} A_n^{[k]} &= \varphi_n(Q_{n,0:K-1}^{[k]}, W_{0:K-1}) \\ &= W_{0,Q_{n,0}^{[k]}} \oplus W_{1,Q_{n,1}^{[k]}} \oplus \dots \oplus W_{K-1,Q_{n,K-1}^{[k]}} \\ &= W_{k,(n-\mathsf{F}_k^*)_N} \oplus \big(W_{0,\mathsf{F}_0} \oplus \\ &\qquad \dots \oplus W_{k-1,\mathsf{F}_{k-1}} \oplus W_{k+1,\mathsf{F}_k} \oplus \dots \oplus W_{K-1,\mathsf{F}_{K-2}}\big), \end{aligned} \tag{24}$$

where $\oplus$ is the addition operation in the group $\mathcal{X}$. For conciseness, we shall define

$$F \triangleq W_{0,\mathsf{F}_0} \oplus \dots \oplus W_{k-1,\mathsf{F}_{k-1}} \oplus W_{k+1,\mathsf{F}_k} \oplus \dots \oplus W_{K-1,\mathsf{F}_{K-2}}. \tag{25}$$

4) The answers from the servers are

$$A_n^{[k]} = W_{k,(n-\mathsf{F}_k^*)_N} \oplus F, \quad n \in \{0, 1, \dots, N-1\}. \tag{26}$$

The message W_k can now be reconstructed by computing

$$W_{k,(n-\mathsf{F}_k^*)_N} = A_n^{[k]} \ominus A_{\mathsf{F}_k^*}^{[k]} = A_n^{[k]} \ominus F, \quad n \in \{0, 1, \dots, N-1\}, \tag{27}$$

where $\ominus$ denotes subtraction in the abelian group $\mathcal{X}$.

TABLE II
ANSWERS FOR MESSAGE A AND B WHEN $(N, K) = (2, 2)$ IN [1]

Requesting A		Requesting B	
Server-1	Server-2	Server-1	Server-2
$a_\square, b_\square, a_\clubsuit + b_\diamondsuit$	$a_\diamondsuit, b_\diamondsuit, a_\heartsuit + b_\square$	$a_\square, b_\square, a_\diamondsuit + b_\clubsuit$	$a_\diamondsuit, b_\diamondsuit, a_\square + b_\heartsuit$

TABLE III
THE QUERY SETS AND THE ANSWERS AT THE SERVERS

Server-0		Server-1		Server-2	
Q_0	answers	Q_1	answers	Q_2	answers
000	0	001	$a_0 \oplus b_0 \oplus c_1$	002	$a_0 \oplus b_0 \oplus c_2$
012	$a_0 \oplus b_1 \oplus c_2$	010	$a_0 \oplus b_1 \oplus c_0$	011	$a_0 \oplus b_1 \oplus c_1$
021	$a_0 \oplus b_2 \oplus c_1$	022	$a_0 \oplus b_2 \oplus c_2$	020	$a_0 \oplus b_2 \oplus c_0$
102	$a_1 \oplus b_0 \oplus c_2$	100	$a_1 \oplus b_0 \oplus c_0$	101	$a_1 \oplus b_0 \oplus c_1$
111	$a_1 \oplus b_1 \oplus c_1$	112	$a_1 \oplus b_1 \oplus c_2$	110	$a_1 \oplus b_1 \oplus c_0$
120	$a_1 \oplus b_2 \oplus c_0$	121	$a_1 \oplus b_2 \oplus c_1$	122	$a_1 \oplus b_2 \oplus c_2$
201	$a_2 \oplus b_0 \oplus c_1$	202	$a_2 \oplus b_0 \oplus c_2$	200	$a_2 \oplus b_0 \oplus c_0$
210	$a_2 \oplus b_1 \oplus c_0$	211	$a_2 \oplus b_1 \oplus c_1$	212	$a_2 \oplus b_1 \oplus c_2$
222	$a_2 \oplus b_2 \oplus c_2$	220	$a_2 \oplus b_2 \oplus c_0$	221	$a_2 \oplus b_2 \oplus c_1$

The correctness of this code is almost immediate, once we observe that in (27), as n ranges in the set $\{0, 1, \ldots, N-1\}$, the corresponding value $\left(n - \mathsf{F}_k^*\right)_N$ exhausts all possible values in $\{0, 1, \ldots, N-1\}$ as well. This implies all the elements $W_{k,n}$, $n \in \{0, 1, \ldots, N-1\}$ are recovered, and thus the message is correctly reconstructed. The privacy of the code is also almost immediate, as for any $k \in \{0, 1, \ldots, K-1\}$, $n \in \{0, 1, \ldots, N-1\}$, and $q \in \mathcal{Q}_n$,

$$\mathbf{Pr}(Q_n^{[k]} = q) = N^{-K+1}, \tag{28}$$

i.e., the queries are sent to a server with a uniform distribution on the respective query set. Since at each server, each answer is sent with probability N^{-K+1}, and only one answer in server-0 has length 0 while all other answers have length 1, the rate of the code is

$$\begin{aligned} R &= \frac{N-1}{(1 - N^{-K+1}) + (N-1)} = \frac{N-1}{N - N^{-K+1}} \\ &= \left(1 + \frac{1}{N} + \frac{1}{N^2} + \ldots + \frac{1}{N^{K-1}}\right)^{-1} = C, \end{aligned} \tag{29}$$

i.e., achieving the capacity. The upload cost is simply

$$N \log_2 N^{K-1} = N(K-1)\log_2 N, \tag{30}$$

which is roughly linear in K for any fixed N.

We summarize the properties of the proposed PIR code construction in the following theorem.

Theorem 1. *The N-ary-indexed PIR code is correct, privacy-preserving, and capacity-achieving. Among all capacity-achieving uniformly decomposable PIR codes, it has the smallest message size, which is $N-1$. Among all capacity-achieving decomposable PIR codes, it has the lowest upload cost, which is $N(K-1)\log_2 N$.*

The optimality in terms of the message size and the upload cost is proved in Section IV. The capacity-achieving code in [1] has a message size of $L = N^K$ and an upload cost of $NK\log_2(\frac{N^K!}{N^{K-1}!})$, while the one in [21] has a message size[1] of $L = N^{K-1}$ and an upload cost of $NK\log_2(\frac{N^{K-1}!}{N^{K-2}!})$. Therefore, the proposed code construction is able to provide an exponential order of improvements over the existing ones in the literature.

C. An Example for $(N, K) = (3, 3)$

Here we use $(N, K) = (3, 3)$ to illustrate the general code construction. The code will have $L = N - 1 = 2$, and we shall denote $W_0 = (a_1, a_2)$, $W_1 = (b_1, b_2)$, $W_2 = (c_1, c_2)$, where all the elements are in the binary field $\{0, 1\}$. As described in the general code construction, we extend these messages by pre-pending one dummy element to each of them, denoted as $a_0 = b_0 = c_0 = 0$, to form

$$W_0 = (a_0, a_1, a_2), W_1 = (b_0, b_1, b_2), W_2 = (c_0, c_1, c_2). \tag{31}$$

In Table III, we provide the query set $\mathcal{Q}_n$ at each server, as well as the corresponding answers.

Let us consider the case where the random key is chosen to be $\mathsf{F} = (0, 2)$, and the message being requested is W_1, then the three queries sent to the servers are

$$q_0 = (0, 1, 2), \quad q_1 = (0, 2, 2), \quad q_2 = (0, 0, 2), \tag{32}$$

i.e., the middle digit in the query is chosen to be the unique value in each query set, and the other two digits are set

[1] The definition of retrieval rate (15) is given in terms of the inverse of the expected number of downloaded symbols (over all random queries), which is in line with the approach taken in [1]. In [21], an alternative definition was adopted, where the retrieval rate was defined in terms of the inverse of the maximum number of downloaded symbols (among all possible queries). Under the alternative definition of [21], the minimum message size was shown to be N^{K-1} for any capacity-achieving codes. In a sense, our result shows that this subtle difference in the problem definition in fact induces a significant difference in terms of the optimal message sizes.

according to $\mathsf{F} = (0, 2)$. The answers are thus

$$\begin{aligned} A_0 &= a_0 \oplus b_1 \oplus c_2 = b_1 \oplus c_2, \\ A_1 &= a_0 \oplus b_2 \oplus c_2 = b_2 \oplus c_2, \\ A_2 &= a_0 \oplus b_0 \oplus c_2 = c_2. \end{aligned} \tag{33}$$

It is clear that b_1 and b_2 can be recovered from these answers by subtracting $A_2 = c_2$ from A_0 and A_1. The code is also privacy-preserving, since regardless of the message being requested, a query element is being sent with probability $1/9$. The retrieval rate is also easy to compute as

$$R = \frac{2}{\frac{1}{9} * 2 + \frac{8}{9} * 3} = \frac{9}{13}, \tag{34}$$

which matches the capacity of this system.

Remark: The queries in each row of Table III are intentionally arranged to have the first two digits being the same, for ease of inspection.

IV. Lower Bounding the Message Size and the Upload Cost

The minimum upload cost and the minimum message size are closely related to the retrieval rate of a PIR code. For example, a naive PIR code where everything is downloaded can have upload cost of 0, and message size of 1, however a more efficient PIR code will need to induce a larger message size and a higher upload cost. In this work, we consider the minimum upload cost and the minimum message size when the retrieval rate is maximized and when the codes are decomposable, i.e., capacity-achieving decomposable codes. We will show, through a delicate set of relations among the coding function matrices $G_n^{(q)}$'s, that the capacity-achieving requirement forces the PIR codes to have certain algebraic structure, which can be utilized to derive the desired lower bounds.

A. Properties of Capacity-Achieving Decomposable Codes

We first provide a detailed analysis of capacity-achieving codes, from which three important properties are derived, given in two lemmas. The analysis is a refinement of the converse proof given in [1], however, with the emphasis on the necessary conditions for optimal codes. A similar approach was used in [41] to analyze optimal joint source-channel codes, and in [43] to facilitate reverse-engineering code designs in coded caching systems.

Lemma 1. *For any PIR code, we have*

$$\begin{aligned} &I\left(W_{0:k-1,k+1:K-1}; A_{0:N-1}^{[k]} \middle| W_k, \mathsf{F}\right) \\ &\quad \le L(1/R - 1)\log_2|\mathcal{X}|, \quad k \in \{0, 1, \ldots, K-1\}. \end{aligned} \tag{35}$$

Moreover, for any PIR code that the equality holds for all $k \in \{0, 1, \ldots, K-1\}$ in (35), let $q_{0:N-1} = (q_0, q_1, \ldots, q_{N-1})$ be a set of queries for which $\mathbf{Pr}(Q_{0:N-1}^{[k]} = q_{0:N-1}) > 0$ for some $k \in \{0, 1, \ldots, K-1\}$, then the code must have

- ***P1***. *Independence of the retrieved data: the N random variables $A_0^{(q_0)}, A_1^{(q_1)}, \ldots, A_{N-1}^{(q_{N-1})}$ are mutually independent, where $A_n^{(q_n)}$ is the answer from server-n when the query $Q_n^{[k]} = q_n$.*

The proof of this lemma is given in the appendix. The property ***P1*** is obtained by setting the inequality (35) to equality, which forces the intermediate steps to also become equality, and then extracting the independence implied by such information theoretic equality.

Remark: For decomposable codes, we can further write

$$\begin{aligned} &\left(A_0^{(q_0)}, A_1^{(q_1)}, \ldots, A_{N-1}^{(q_{N-1})}\right) \\ &= \Big(W_{0:K-1} \cdot G_0^{(q_0)}, W_{0:K-1} \cdot G_1^{(q_1)}, \\ &\qquad \ldots, W_{0:K-1} \cdot G_{N-1}^{(q_{N-1})}\Big). \end{aligned} \tag{36}$$

Also note that for linear codes, the independence relation given above implies that the columns of the matrices $G_0^{(q_0)}, G_1^{(q_1)}, \ldots, G_{N-1}^{(q_{N-1})}$ are linearly independent.

Recall that for decomposable codes, the answer for a query $Q_n = q$ at server-n can be written as $W_{0:K-1} \cdot G_n^{(q)}$, or more concisely, sometimes represented by the coding function matrix $G_n^{(q)}$ alone. The next lemma involves submatrices of $G_n^{(q)}$, with the rows corresponding to a subset of the messages removed, say $\{W_i, i \in \mathcal{A}\}$; we shall write such a submatrix as $G_{n|\mathcal{A}}^{(q)}$. For example, if $(N, K) = (3, 3)$, and $\mathcal{A} = 1$, then $G_{1|1}^{(1)}$ is the submatrix of $G_1^{(1)}$ with the middle row corresponding to the message W_1 removed.

Lemma 2. *Let $\pi : \{0, 1, \ldots, K-1\} \to \{0, 1, \ldots, K-1\}$ be a permutation function. For any PIR code, for any $k \in \{1, 2, \ldots, K-1\}$,*

$$\begin{aligned} &NI\left(W_{\pi(k:K-1)}; A_{0:N-1}^{[\pi(k-1)]} \middle| W_{\pi(0:k-1)}, \mathsf{F}\right) \\ &\quad \ge I\left(W_{\pi(k+1:K-1)}; A_{0:N-1}^{[\pi(k)]} \middle| W_{\pi(0:k)}, \mathsf{F}\right) + L\log_2|\mathcal{X}|. \end{aligned} \tag{37}$$

Moreover, for any decomposable code for which the equality holds for any k and $\pi(\cdot)$ in (37), let $q_{0:N-1} = (q_0, q_1, \ldots, q_{N-1})$ be a set of queries for which $\mathbf{Pr}(Q_{0:N-1}^{[k]} = q_{0:N-1}) > 0$ for the query of the message W_k, and $G_0^{(q_0)}, G_1^{(q_1)}, \ldots, G_{N-1}^{(q_{N-1})}$ be the corresponding answer coding matrices, then

- ***P2***. *Identical information for the residuals: the N random variables*

$$\begin{aligned} &W_{0:k-1,k+1:K-1} \cdot G_{0|k}^{(q_0)}, W_{0:k-1,k+1:K-1} \cdot G_{1|k}^{(q_1)}, \\ &\qquad \ldots, W_{0:k-1,k+1:K-1} \cdot G_{N-1|k}^{(q_{N-1})} \end{aligned} \tag{38}$$

are deterministic of each other;

- ***P3***. *Independence of the requested message signals: the random variables*

$$\begin{aligned} &W_k \cdot G_{0|0:k-1,k+1:K-1}^{(q_0)}, W_k \cdot G_{1|0:k-1,k+1:K-1}^{(q_1)}, \\ &\qquad \ldots, W_k \cdot G_{N-1|0:k-1,k+1:K-1}^{(q_{N-1})} \end{aligned} \tag{39}$$

are independent.

The proof of this lemma can be found in the appendix.

Remark: The property of decomposable codes was used in the proof of Lemma 2, where the answers are decomposed into separate components according to the messages W_k's, with which relations among these answers are derived. Such decomposition does not apply on other code classes in general, and thus the proof cannot be carried through directly using the same argument.

Theorem 2. *Any capacity-achieving decomposable PIR code must have the properties **P1-P3**.*

Proof: Let $\pi : \{0, 1, \ldots, K-1\} \to \{0, 1, \ldots, K-1\}$ be a permutation. Starting from Lemma 1, we can write

$$\begin{aligned}\frac{L}{R} - L &\geq I\left(W_{\pi(1:K-1)}; A_{0:N-1}^{[\pi(0)]} \middle| W_{\pi(0)}, \mathsf{F}\right)\\ &\geq \frac{L}{N} + \frac{1}{N} I\left(W_{\pi(2:K-1)}; A_{0:N-1}^{[\pi(1)]} \middle| W_{\pi(0:1)}, \mathsf{F}\right)\\ &\geq \cdots\\ &\geq L\left(\frac{1}{N} + \ldots + \frac{1}{N^{K-1}}\right), \end{aligned} \tag{40}$$

where all the other inequalities are by recursively applying Lemma 2, and it follows that $R \leq C$. For any decomposable code that satisfies $R = C$, all the inequalities in Lemma 1 and Lemma 2 must be equality for any permutation π, and according to the lemmas, such decomposable codes must have properties ***P1-P3***. ■

B. Minimum Message Size

We have the following theorem, which provides a lower bound on the minimum message size for capacity-achieving uniformly decomposable codes.

Theorem 3. *The message size of any uniformly decomposable capacity-achieving PIR code is greater than or equal to $(N-1)\log_2|\mathcal{Y}|$; in particular, it must be greater than or equal to $(N-1)$.*

Remark: Clearly this implies that the standard linear codes defined on finite fields are lower bounded by the same values. Note also that the bound $(N-1)\log_2|\mathcal{Y}|$ is dependent on $\mathcal{Y}$ but not $\mathcal{X}$, which reflects the fact that the representation of the message is of little fundamental importance because we can always use an equivalent representation.

Proof: Let us consider a capacity-achieving uniformly decomposable PIR code, and the request to retrieve the message W_k. Recall property ***P2*** which states that $W_{0:k-1,k+1:K-1} \cdot G_{n|k}^{(q_n)}$, $n \in \{0, 1, \ldots, N-1\}$, are deterministic functions of each other. There must be a set of queries $q_{0:N-1}$ with non-zero probability such that

$$H\left(W_{0:k-1,k+1:K-1} \cdot G_{n|k}^{(q_n)}\right) \neq 0, \qquad n = \{0, 1, \ldots, N-1\}, \tag{41}$$

because otherwise, all answers will have the form

$$W_{0:K-1} \cdot G_n^{(q_n)} = W_k \cdot G_{n|0:k-1,k+1:K+1}^{(q_n)} \oplus \Delta, \qquad n = \{0, 1, \ldots, N-1\}, \tag{42}$$

where $\Delta \in \mathcal{Y}$ is a constant; this would imply that the answers only involve the message W_k but not other messages, but such answers clearly cannot be both private and correct.

With such a set of queries $q_{0:N-1}$ that (41) holds, consider property ***P3***, which states that

$$W_k \cdot G_{0|0:k-1,k+1:K-1}^{(q_0)}, W_k \cdot G_{1|0:k-1,k+1:K-1}^{(q_1)}, \ldots, W_k \cdot G_{N-1|0:k-1,k+1:K-1}^{(q_{N-1})} \tag{43}$$

are independent, and our aim is to show that no more than one of their entropies can be zero. To see this, assume otherwise, i.e., at least two of the entropies are zero. Without loss of generality, let us assume that

$$H\left(W_k \cdot G_{0|0:k-1,k+1:K-1}^{(q_0)}\right) = H\left(W_k \cdot G_{1|0:k-1,k+1:K-1}^{(q_1)}\right) = 0, \tag{44}$$

implying that both $W_k \cdot G_{0|0:k-1,k+1:K-1}^{(q_0)}$ and $W_k \cdot G_{1|0:k-1,k+1:K-1}^{(q_1)}$ in fact take a fixed value, independent of the value of W_k. However, this further implies that the retrieved messages from server-0 and server-1 are

$$\begin{aligned} W_{0:K-1} \cdot G_0^{(q_0)} &= W_{0:k-1,k+1:K-1} \cdot G_{0|k}^{(q_0)} \oplus \Delta_1,\\ W_{0:K-1} \cdot G_1^{(q_1)} &= W_{0:k-1,k+1:K-1} \cdot G_{1|k}^{(q_1)} \oplus \Delta_2, \end{aligned} \tag{45}$$

where Δ_1 and Δ_2 are two constants in the abelian group $\mathcal{Y}$. Because of property ***P2***, the two random variables in (45) are in fact deterministic of each other. However this contradicts property ***P1*** which states that the retrieved contents are independent (recall that their entropies are not zero). Thus we can conclude that at least $N-1$ of the entropies of the terms in (43) are not zero. Because the function $\varphi^{(q)}(n, i, k)$ induces a uniform probability distribution on the coded symbol alphabet $\mathcal{Y}$ (unless it takes a deterministic value), and moreover, by the independence property of ***P3***, we can now conclude that the message size must be greater than or equal to $(N-1)\log_2|\mathcal{Y}|$. Since any meaningful alphabet $\mathcal{Y}$ must satisfy $|\mathcal{Y}| \geq 2$, the message size must be greater than or equal to $N-1$. The proof is thus complete. ■

Remark: The property of uniformly decomposable codes is only invoked during the proof in the last step, which requires the component functions to induce a uniform distribution on the coded alphabet, unless it always takes a deterministic value.

C. Minimum Upload Cost

Theorem 4. *The upload cost of any capacity-achieving decomposable PIR code is greater than or equal to $N(K-1)\log_2 N$.*

We need the following notion of distinctness in the proof.

Definition 4. *Two random variables A and B are called information-theoretically distinct, or simply distinct, if $I(A; B) < \max(H(A), H(B))$.*

According to this definition, if a random variable can be obtained from another through an invertible transformation, they are not information-theoretically distinct.

$$\begin{array}{llllll}
\text{Server } 0 & \text{Server } 1 & \text{Server } 2 & \text{...} & \text{Server } N-1 & \\
\hline\hline
\underline{V_{0,0,0,\ldots,0}}, & V_{0,1,0,\ldots,0}, & V_{0,2,0,\ldots,0}, & \ldots & V_{0,N-1,0,\ldots,0} & \to W_1 \\
V_{1,N-1,0,\ldots,0}, & \underline{V_{1,0,0,\ldots,0}}, & V_{1,1,0,\ldots,0}, & \ldots & V_{1,N-2,0,\ldots,0} & \to W_1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
V_{N-1,1,0,\ldots,0}, & V_{N-1,2,0,\ldots,0}, & V_{N-1,3,0,\ldots,0}, & \ldots & \underline{V_{N-1,0,0,\ldots,0}} & \to W_1
\end{array} \tag{51}$$

$$\begin{array}{llllll}
\text{Server } 0 & \text{Server } 1 & \text{Server } 2 & \text{...} & \text{Server } N-1 & \\
\hline\hline
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
V_{N-1,2,N-1,\ldots,0}, & \underline{V_{N-1,2,0,\ldots,0}}, & V_{N-1,2,1,\ldots,0}, & \ldots & V_{N-1,2,N-2,\ldots,0} & \to W_2 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
V_{1,1,N-2,\ldots,0}, & V_{1,1,N-1,\ldots,0}, & \underline{V_{1,1,0,\ldots,0}}, & \ldots & V_{1,1,N-3,\ldots,0} & \to W_2 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots.
\end{array} \tag{53}$$

Proof: We prove that for a capacity-achieving decomposable PIR code for N servers and K messages, the minimum upload cost to each server is at least $(K-1)\log_2 N$, i.e., it is a lower bound on $\log_2 |\mathcal{Q}_n|$, $n \in \{0, 1, \ldots, N-1\}$. To begin the proof, we find a set of queries $q_{0:N-1}$ for the message W_0, and assume that the answers have the property that the interference signal (i.e., the part of the answer that is not the requested message) is not null, i.e.,

$$H\left(W_{1:K-1} \cdot G^{(q_n)}_{n|0}\right) \neq 0, \quad n = \{0, 1, \ldots, N-1\}, \tag{46}$$

which always exists using the same argument as in Theorem 2; *c.f.* (41). This implies that at least for one of the interference signals $k \in \{1, 2, \ldots, K-1\}$ we have

$$H\left(W_k \cdot G^{(q_n)}_{n|0:k-1,k+1:K-1}\right) \neq 0, \quad n = \{0, 1, \ldots, N-1\}, \tag{47}$$

due to property ***P2***. Without loss of generality, let us assume it is $k = K-1$. With this set of queries, following the argument in Theorem 2, at most one of the entropies

$$H\left(W_0 \cdot G^{(q_n)}_{n|1:K-1}\right), \quad n = \{0, 1, \ldots, N-1\} \tag{48}$$

can be zero. Again without loss of generality, assume it is $n = 0$. We shall denote a particular answer from a server as $V_{a_0,a_1,\ldots,a_{K-1}}$, the meaning of which will soon become apparent.

Since the queries $q_{0:N-1}$ are for the message W_0, the answers from the N servers, respectively,

$$V_{0,0,0,\ldots,0}, V_{1,0,0,\ldots,0}, \ldots, V_{N-1,0,0,\ldots,0} \tag{49}$$

can be used to recover W_0, and moreover, the W_0 component functions

$$W_0 \cdot G^{(q_n)}_{n|1:K-1}, \quad n = \{0, 1, \ldots, N-1\} \tag{50}$$

are all information-theoretically distinct by property ***P3***, and the fact that at most one of them can have zero entropy. We indicate this distinctness by the subscript in the answers (49) in the 0-th position.

Due to the privacy constraint, each answer in (49) can also be used to reconstruct W_1, together with some other answers, i.e., as shown in (51) at the top of this page. In each row of (51), the results produced by the component functions on $W_{0,2:N-1}$ in the answers are deterministic functions of each other across different servers, due to property ***P2***. Moreover, for these answers, the component W_{K-1} must satisfy

$$H\left(W_{K-1} \cdot G^{(q_n)}_{n|0:K-2}\right) \neq 0, \quad n = \{0, 1, \ldots, N-1\}, \tag{52}$$

due to our assumption on (47) holding for $k = K-1$. As a consequence, in each row of (51), the component functions $W_1 \cdot G^{(q)}_{n|0,2:K-1}$ are again distinct. Note however, across rows of (51), the component functions on W_1 in the answers are not necessarily distinct or identical. However, the component functions $W_{0:1} \cdot G^{(q)}_{n|2:K-1}$ of the answers in (51) are all distinct, since they have distinct W_0 component functions in different rows, (i.e., for answers in different rows, the W_0 component functions are (50)), while for answers in the same row, the W_1 component functions are distinct. Thus there are at least N answers with distinct component functions $W_{0:1} \cdot G^{(q)}_{n|2:K-1}$ at server-n, which are the answers with the sum of the indices equal to n modulo N, given in the same column in (51).

Next consider each answer in (51), which can also be used to recover W_2 due to the privacy requirement. For example, if we focus on the answers $V_{N-1,2,0,\ldots,0}$ and $V_{1,1,0,\ldots,0}$, which are from server 1 and server 2, respectively, they can be used to recover W_2 with some other answers, as shown in (53) at the top of this page. Again these answers are distinct through a similar argument as before. Using this argument on all the answers in (51) for the retrieval of W_2, it can be seen that across all the servers, there are at least N^3 answers, whose component functions $W_{0:2} \cdot G^{(q)}_{n|3:K-1}$ are all distinct, and each server has at least N^2 answers whose corresponding component functions are distinct. We can continue this line of argument for messages $W_3, W_4, \ldots, W_{K-2}$, resulting in a total of N^{K-1} answers at all the servers (N^{K-2} at each server) in the form of

$$V_{a_0,a_1,\ldots,a_{K-2},0}, \quad a_k \in \{0, 1, \ldots, N-1\}, \tag{54}$$

whose component functions $W_{0:K-2} \cdot G^{(q)}_{n|K-1}$ are distinct.

Next consider the reconstruction of the message W_{K-1}, for which we need to be more careful. In this case, we cannot assume the interference signals in the retrieval are not null, because W_{K-1} is now the requested message, and the condition (52) becomes insufficient; thus the component functions of W_{K-1} in the answers cannot be guaranteed to be all distinct during a retrieval. However, notice that due to the distinctness of the component functions $W_{0:K-2} \cdot G^{(q)}_{n|K-1}$ in all the answers in (54), at most one of these component functions can have zero entropy, i.e., in these answers, there is at most one of them satisfying,

$$H\left(W_{0:K-2} \cdot G^{(q)}_{n|K-1}\right) = 0. \tag{55}$$

For all other answers that (55) does not hold, our previous induction argument based on the distinctness of the signal components still applies. For the one exception answer where (55) holds, which we assume without of generality to be $V_{0,0,\ldots,0}$, the answers to recover W_{K-1} can be labeled as

$$V_{0,0,\ldots,0}, V_{0,0,\ldots,1}, \ldots, V_{0,0,\ldots,N-1}, \tag{56}$$

which may not be all distinct since there may be more than one item with zero entropy. However, since they are placed at different servers, each one of them is distinct from all other answers at the same server.

We have shown that at server-n, $n \in \{0, 1, \ldots, N-1\}$, there are at least N^{K-1} distinct answers $V_{a_0,a_1,\ldots,a_{K-1}}$ in the form of

$$\left(\sum_{k=0}^{K-1} a_k\right)_N = n, \tag{57}$$

implying $|\mathcal{Q}_n| \geq N^{K-1}$. Our proof is now complete. ∎

Remark: Although Theorem 4 is stated in terms of the total upload cost, in the proof, we have actually shown that the upload cost at each individual server is greater than or equal to $(K-1)\log_2 N$.

V. Symmetry and Symmetrized Codes

The proposed code construction is able to achieve exponential improvements over the existing capacity-achieving PIR codes in the literature, in terms of both the message size and the upload cost. The question we wish to address in this section is what the root cause is for these improvements. It is clear that the existing codes in the literature, such as [1]–[4], [16]–[18], are all symmetric, while our proposed code is not symmetric. It is thus natural to suspect that this symmetry vs. asymmetry relation is the root cause, however, in order to better understand this issue, we have to identify and evaluate carefully the symmetry relations in the problem.

Recall our discussion on the minimum upload cost, which is related to $|\mathcal{Q}_n|$. For simplicity, we shall refer to the distinct answers (or precisely, distinct answer functions) at a server as the varieties of the answers at this server.[2] This concept plays an instrumental role in the subsequent discussion.

There are in fact three kinds of symmetry relations in this problem setting:

1) Server-symmetry: obtained by permuting the servers;
2) Message-symmetry: obtained by permuting the messages;
3) Variety-symmetry: obtained by compositing the varieties of answers.

Among the three types of symmetry relations, the variety-symmetry is the most interesting, and appears unique to the PIR problem. Through this symmetry, it can be shown that without loss of optimality on the retrieval rate, we can always assume that the varieties are requested with a uniform distribution at any given server. These three symmetry components can be operated in composition, and space sharing of all possible permuted codes eventually can yield a highly symmetric code. In this section we shall provide a precise characterization of these three types of symmetry relations, and discuss several consequences of these relations. Technically, this is accomplished by constructing a new set of coding functions, which either by space-sharing over some permutations or some other mechanism, will induce certain symmetry relation on the coding rates and the probability distribution.

Central to these symmetry relations are the following random variables

$$\begin{gathered}\{W_0, W_1, \ldots, W_{K-1}\},\\ \{A_0^{(0)}, A_0^{(1)}, \ldots, A_0^{(|\mathcal{Q}_0|-1)}\},\\ \{A_1^{(0)}, A_1^{(1)}, \ldots, A_1^{(|\mathcal{Q}_1|-1)}\},\\ \ldots,\\ \{A_{N-1}^{(0)}, A_{N-1}^{(1)}, \ldots, A_{N-1}^{(|\mathcal{Q}_{N-1}|-1)}\},\end{gathered} \tag{58}$$

where $A_n^{(q)}$ is the answer at server-n for the query $Q_n = q$. Note that $A_n^{(q)}$ is a deterministic function of the messages $W_{0:K-1}$; this should be distinguished from $A_n^{[k]}$ which is the (randomized) answer for the request of the message W_k at server-n, and not a deterministic function of the messages $W_{0:K-1}$.

The symmetrization techniques given this section should not be viewed as design requirements stipulated by practical system design considerations, but rather should be viewed as theoretical tools to pinpoint the key difference between our code construction and the existing ones, and perhaps to help future investigations on the capacities of privacy-preserving primitives, as they appear to be rather general.

A. Server-Symmetry

Let $\pi(\cdot)$ be a permutation function on the set $\{0, 1, \ldots, N-1\}$, which is the set of server indices. For any PIR code which is specified by the four coding functions in Definition 1, a new set of coding functions can be specified as

$$\hat{\phi}_n = \phi_{\pi(n)}, \quad \hat{\ell}_n = \ell_{\pi(n)}, \quad \hat{\varphi}_n = \varphi_{\pi(n)}, \qquad n \in \{0, 1, \ldots, N-1\}, \tag{59}$$

and

$$\hat{\psi}(A_{0:N-1}, k, \mathsf{F}) = \psi(A_{\pi^{-1}(0:N-1)}, k, \mathsf{F}). \tag{60}$$

Let us examine an example where $N = 4$, and let

$$\pi([0, 1, 2, 3]) = [3, 0, 1, 2]. \tag{61}$$

Then we have

$$\hat{\phi}_0 = \phi_3, \quad \hat{\phi}_1 = \phi_0, \quad \hat{\phi}_2 = \phi_1, \quad \hat{\phi}_3 = \phi_2, \tag{62}$$

[2] The term variety here should be distinguished from the algebraic variety concept in algebraic geometry.

that is, the query sent to server-0 in this new code is what was sent to server-3, etc. Similarly,

$$\hat{\ell}_0 = \ell_3, \quad \hat{\ell}_1 = \ell_0, \quad \hat{\ell}_2 = \ell_1, \quad \hat{\ell}_3 = \ell_2,$$
$$\hat{\varphi}_0 = \varphi_3, \quad \hat{\varphi}_1 = \varphi_0, \quad \hat{\varphi}_2 = \varphi_1, \quad \hat{\varphi}_3 = \varphi_2, \tag{63}$$

that is, the function to produce the answer (and the length of the answer) at server-0 in the permuted code is what was used at server-3 for the same query value, etc.; moreover, for the reconstruction function

$$\hat{\psi}(A_0, A_1, A_2, A_3, k, \mathsf{F}) = \psi(A_1, A_2, A_3, A_0, k, \mathsf{F}), \tag{64}$$

that is, the reconstructed message $\hat{W}_k$ using random key F, is in fact obtained by operating the original function on the permuted answers, i.e., using the answer obtained from server-0 in the place of what was for the answer from server-3, etc.

It is easy to see that this new set of coding functions is indeed privacy-preserving and correct, since there is no essential change in the coding operations. A direct consequence of the definition of the new code is reflected on the equivalence of the induced random variables in the two codes as

$$\begin{aligned}
&\{\hat{W}_0, \hat{W}_1, \ldots, \hat{W}_{K-1}\},\\
&\{\hat{A}_0^{(0)}, \hat{A}_0^{(1)}, \ldots, \hat{A}_0^{(|\hat{\mathcal{Q}}_0|-1)}\},\\
&\quad \{\hat{A}_1^{(0)}, \hat{A}_1^{(1)}, \ldots, \hat{A}_1^{(|\hat{\mathcal{Q}}_1|-1)}\},\\
&\qquad \ldots, \{\hat{A}_{N-1}^{(0)}, \hat{A}_{N-1}^{(1)}, \ldots, \hat{A}_{N-1}^{(|\hat{\mathcal{Q}}_{N-1}|-1)}\}\\
&= \{W_0, W_1, \ldots, W_{K-1}\},\\
&\{A_{\pi(0)}^{(0)}, A_{\pi(0)}^{(1)}, \ldots, A_{\pi(0)}^{(|\mathcal{Q}_{\pi(0)}|-1)}\},\\
&\quad \{A_{\pi(1)}^{(0)}, A_{\pi(1)}^{(1)}, \ldots, A_{\pi(1)}^{(|\mathcal{Q}_{\pi(1)}|-1)}\},\\
&\qquad \ldots, \{A_{\pi(N-1)}^{(0)}, A_{\pi(N-1)}^{(1)}, \ldots, A_{\pi(N-1)}^{(|\mathcal{Q}_{\pi(N-1)}|-1)}\}.
\end{aligned} \tag{65}$$

Next consider the following code constructed through the space-sharing technique using a base code. Let each message consist of a total of NL symbols, and apply a permuted version of the base code on each length-L sequence (and over the K messages), which corresponds to one of the cyclic permutations on $\{0, 1, \ldots, N-1\}$. This space-sharing code is clearly privacy-preserving and correct, and it has the property that $|\hat{\mathcal{Q}}_0| = |\hat{\mathcal{Q}}_1| = \ldots = |\hat{\mathcal{Q}}_{N-1}| = \prod_{n=0}^{N-1} |\mathcal{Q}_n|$, i.e., the upload costs to all the servers are the same. Moreover, the expected retrieval rates are also the same across all the servers, i.e., $\mathbb{E}(\hat{\ell}_0) = \mathbb{E}(\hat{\ell}_1) = \ldots = \mathbb{E}(\hat{\ell}_{N-1})$.

We could also space share over longer messages of $(N!)L$ symbols each, where for each length-L sequence we apply the permuted coding function corresponding to one of the $N!$ permutations on $\{0, 1, \ldots, N-1\}$. By leveraging (65), it is also possible to obtain an invariance in terms of the joint entropy values of the subsets of the random variables. Such refined invariant relations are not necessary for this work, however, similar relations have been shown to be important when deriving information theoretic converse bounds [42], [43] in other information systems.

It should be noted that although the expected numbers of retrieved symbols are the same across the servers (and thus the retrieval rates are the same per server), this does not imply for each individual set of queries $q_{0:N-1}$ with non-zero probability, the numbers of symbols being retrieved are the same as those for another set of queries $q'_{0:N-1}$. To achieve such a fine level of invariance, we will need to invoke the variety-symmetry, to be introduced in Section V-C.

B. Message-Symmetry

Let $\pi(\cdot)$ be a permutation function on the set $\{0, 1, \ldots, K-1\}$, which is the set of message indices. For any PIR code which is specified by the four coding functions in Definition 1, a new set of coding functions can be specified as

$$\bar{\phi}_n = \phi_n, \quad \bar{\ell}_n = \ell_n, \quad \bar{\varphi}_n(q, W_{0:K-1}) = \varphi_n(q, W_{\pi(0:K-1)}),$$
$$n \in \{0, 1, \ldots, N-1\}, \tag{66}$$

and

$$\bar{\psi}(A_{0:N-1}, k, \mathsf{F}) = \psi(A_{0:N-1}, \pi^{-1}(k), \mathsf{F}). \tag{67}$$

Let us examine an example where $K = 3$ and let

$$\pi([0, 1, 2]) = [2, 0, 1]. \tag{68}$$

Then we have for the functions $\bar{\varphi}_n$

$$\bar{\varphi}_n(q, W_0, W_1, W_2) = \varphi_n(q, W_2, W_0, W_1),$$
$$n \in \{0, 1, \ldots, N-1\}, \tag{69}$$

that is, the message W_0 in the new code serves the role of W_1 in the original code, etc.

For the reconstruction functions

$$\begin{aligned}
\bar{\psi}(A_{0:N-1}, 0, \mathsf{F}) &= \psi(A_{0:N-1}, 1, \mathsf{F}),\\
\bar{\psi}(A_{0:N-1}, 1, \mathsf{F}) &= \psi(A_{0:N-1}, 2, \mathsf{F}),\\
\bar{\psi}(A_{0:N-1}, 2, \mathsf{F}) &= \psi(A_{0:N-1}, 0, \mathsf{F}),
\end{aligned} \tag{70}$$

that is, the message W_0 is reconstructed in the same way as that for W_1 in the base code, etc.

This new set of coding functions is again privacy-preserving and correct. A direct consequence of the definition of the permuted code is reflected on the equivalence in the probability distribution of the random variables

$$\begin{aligned}
&\{\bar{W}_0, \bar{W}_1, \ldots, \bar{W}_{K-1}\},\\
&\{\bar{A}_0^{(0)}, \bar{A}_0^{(1)}, \ldots, \bar{A}_0^{(|\mathcal{Q}_0|-1)}\},\\
&\quad \{\bar{A}_1^{(0)}, \bar{A}_1^{(1)}, \ldots, \bar{A}_1^{(|\mathcal{Q}_1|-1)}\}, \ldots,\\
&\qquad \ldots, \{\bar{A}_{N-1}^{(0)}, \bar{A}_{N-1}^{(1)}, \ldots, \bar{A}_{N-1}^{(|\mathcal{Q}_{N-1}|-1)}\}\\
&\overset{d}{=} \{W_{\pi^{-1}(0)}, W_{\pi^{-1}(1)}, \ldots, W_{\pi^{-1}(K-1)}\},\\
&\{A_0^{(0)}, A_0^{(1)}, \ldots, A_0^{(|\mathcal{Q}_0|-1)}\},\\
&\quad \{A_1^{(0)}, A_1^{(1)}, \ldots, A_1^{(|\mathcal{Q}_1|-1)}\},\\
&\qquad \ldots, \{A_{N-1}^{(0)}, A_{N-1}^{(1)}, \ldots, A_{N-1}^{(|\mathcal{Q}_{N-1}|-1)}\},
\end{aligned} \tag{71}$$

where $\overset{d}{=}$ indicates equivalence in distribution, but not necessarily identical.

Next consider the following code constructed through the space-sharing technique using a base code. Let each message consist of a total of $(K!)L$ symbols, and apply a permuted version of the base code on each length-L sequence (and across K messages), which corresponds to one of the possible permutations on $\{0, 1, \ldots, K-1\}$. This space-sharing code is clearly privacy-preserving and correct, however it does not lead to any explicit symmetry relation on the coding rates or the distribution on the queries. It does lead to more subtle invariant relations on the entropies of the subsets of the random variables, e.g., the joint entropy of a subset of the answers and a subset of the messages is invariant to which subset of messages is being involved. This symmetry cannot produce the invariance on the individual varieties we mentioned earlier.

C. Variety-Symmetry

The last symmetry we consider is produced by constructing a different set of queries (and answer varieties) and a new random key $\hat{\mathsf{F}}$ to retrieve the messages. The variety-symmetry is constructed using a different mechanism than the previous two types of symmetry relations.

Recall in the base code, the random key F is uniformly distributed on the alphabet $\mathcal{F}$. In the new code, the random key is uniformly distributed on the following set

$$\grave{\mathcal{F}} \triangleq \Big\{ f_{0:|\mathcal{F}|-1} \in \mathcal{F}^{|\mathcal{F}|} : \\ f_{0:|\mathcal{F}|-1} \text{ is a permutation of the elements of } \mathcal{F} \Big\}. \quad (72)$$

It follows that $|\grave{\mathcal{F}}| = |\mathcal{F}|!$. The new code operates as follows. The message has $|\mathcal{F}|L$ symbols, which is partitioned into $|\mathcal{F}|$ length-L blocks. Suppose a particular random key realization $\grave{\mathsf{F}} = f_{0:|\mathcal{F}|-1}$ is generated for the new code. For index $i \in \{0, 1, \ldots, |\mathcal{F}|-1\}$, the corresponding i-th blocks of the messages are encoded using the base code retrieval strategy determined by the key value $\mathsf{F} = f_i \in \mathcal{F}$.

This new code is clearly correct, and next we show that it is also privacy-preserving. Recall for the request of the message W_k, the query for server-n is a deterministic function of the random key $\mathsf{F} = f$ in the base code. Because in the new code, any valid key $f_{0:|\mathcal{F}|-1}$ is a permutation of all the elements in $\mathcal{F}$, the number of times that a particular query $q \in \mathcal{Q}_n$ appears in such a query sequence $f_{0:|\mathcal{F}|-1}$ at server-n is given by

$$\kappa_{n,k}(q) \triangleq |\{f \in \mathcal{F} : \phi_n(k, f) = q\}|. \quad (73)$$

Because the base code is privacy-preserving, we have

$$\kappa_n(q) \triangleq \kappa_{n,0}(q) = \kappa_{n,1}(q) = \ldots = \kappa_{n,K-1}(q), \\ n \in \{0, 1, \ldots, N-1\}, \quad q \in \mathcal{Q}_n. \quad (74)$$

The composition of any query $q_{0:|\mathcal{F}|-1}$ sent to server-n for the request of the message W_k in this new code, which is a vector of length $|\mathcal{F}|$, is thus given exactly by (74), and the only difference among the queries is the patterns that these elements in $\mathcal{Q}_n$ are arranged. Thus, the query set at server-n is the constant composition set, i.e.,

$$\mathcal{T}_n = \Big\{ q_{0:|\mathcal{F}|-1} \in \mathcal{Q}_n^{|\mathcal{F}|} : \text{ the number of} \\ \text{appearances of any } q \in \mathcal{Q}_n \text{ in } q_{0:|\mathcal{F}|-1} = \kappa_n(q) \Big\}. \quad (75)$$

TABLE IV

ANSWERS FOR MESSAGE A AND B FOR $(N, K) = (2, 2)$ AFTER VARIETY-SYMMETRIZATION

	Requesting A		Requesting B	
	Server-1	Server-2	Server-1	Server-2
$\mathsf{F} = (01)$	$a_2 + b_2$	a_1, b_2	$a_2 + b_2$	a_2, b_1
$\mathsf{F} = (10)$	$a_1 + b_1$	a_2, b_1	$a_1 + b_1$	a_1, b_2

Due to the symmetry in $\grave{\mathcal{F}}$ and $\mathcal{T}_n$, as well as the uniform distribution on $\grave{\mathcal{F}}$, it is clear that the distribution of the query on $\mathcal{T}_n$ is also uniform, regardless of the identity of the requested message. Thus this new code is indeed privacy-preserving. As a direct consequence of the construction, at each server, all the answer varieties also have the same numbers of symbols.

D. Applying the Symmetrization Techniques

Let us revisit our example for $(N, K) = (2, 2)$ given in Section III-A. To make a variety-symmetric code, we let each message be 2 bits, denoted as $A = (a_1, a_2), B = (b_1, b_2)$, respectively. The total number of new varieties at each server is $|\grave{\mathcal{Q}}_n| = 2!$. This new code is illustrated in Table IV. It can be seen that now at each server, the lengths of the answers are indeed the same. We can further apply the server-symmetrization technique, which will produce a code quite similar to that proposed in [1] and illustrated in Table II.

We can apply the variety-symmetrization technique on our proposed code with more general parameters. The message size will increase by a factor of N^{K-1}, resulting in a total message size of $N^{K-1}(N-1)$ in the new symmetrized code. In [21], it was shown that if we insist that the total number of retrieved symbols from all servers is the same for all possible query combinations, then the minimum message size is N^{K-1}. Our proposed code in Section III-B has a much smaller message size of $N-1$, but does not have this property, which turns out to be rather restrictive. On the other hand, the variety-symmetrized code based on our proposed code has a slightly larger message size of $N^{K-1}(N-1)$ than the optimal value in the restricted setting of [21]. This relatively small increase appears to have stemmed from the decoupled design strategy of applying the symmetrization technique on a base code, instead of designing a symmetric code directly.

More generally, we can apply all three symmetrization techniques on any asymmetric code (in any order) to obtain a code that is highly symmetric without jeopardizing the retrieval rate, but at the expense of the message size and the upload cost. From this perspective, the reason behind the small message size and upload cost of the proposed code is indeed its asymmetric nature.

VI. CONCLUSION

We proposed a new capacity-achieving PIR code construction, which has the optimal message size and the optimal upload cost. The key to the reduction of both factors, compared to existing constructions, appears to be the asymmetry in the proposed code. In order to prove converse bounds for the optimal message size and the optimal upload cost, we extracted certain critical structures in the converse proof of the PIR capacity. The symmetry structure in the PIR problem is of interest in its own right, and we provided a careful analysis of this structure, which can be used to symmetrize any PIR code into its symmetric version.

Although in this work we have focused on the most canonical setting of the private information retrieval problem, the proposed code construction using asymmetric structure can be extended to more general settings, such as maximum distance separable code (MDS-coded) databases, for which readers can refer to [44].

Appendix A
Proof of Technical Lemmas

Proof: [Proof of Lemma 1] Without loss of generality, let us consider $k=0$. We start by writing the following chain of inequalities:

$$\begin{aligned}
&I\left(W_{1:K-1}; A^{[0]}_{0:N-1}\middle| W_0, \mathsf{F}\right)\\
&\overset{(a)}{=} I\left(W_{1:K-1}; A^{[0]}_{0:N-1}, W_0\middle|\mathsf{F}\right)\\
&= I\left(W_{1:K-1}; A^{[0]}_{0:N-1}\middle|\mathsf{F}\right) + I\left(W_{1:K-1}; W_0\middle| A^{[0]}_{0:N-1}, \mathsf{F}\right)\\
&\overset{(b)}{=} I\left(W_{1:K-1}; A^{[0]}_{0:N-1}\middle|\mathsf{F}\right)\\
&= H\left(A^{[0]}_{0:N-1}\middle|\mathsf{F}\right) - H\left(A^{[0]}_{0:N-1}\middle| W_{1:K-1}, \mathsf{F}\right)\\
&= H\left(A^{[0]}_{0:N-1}\middle|\mathsf{F}\right) - H\left(W_0, A^{[0]}_{0:N-1}\middle| W_{1:K-1}, \mathsf{F}\right)\\
&\qquad\qquad + H\left(W_0\middle| A^{[0]}_{0:N-1}, W_{1:K-1}, \mathsf{F}\right)\\
&\overset{(c)}{=} H\left(A^{[0]}_{0:N-1}\middle|\mathsf{F}\right) - H\left(W_0 | W_{1:K-1}, \mathsf{F}\right)\\
&\overset{(d)}{\le} \left[\frac{L}{R} - L\right]\log_2|\mathcal{X}|,
\end{aligned} \tag{76}$$

where (a) is because the components of $(W_0, W_1, \ldots, W_{K-1}, \mathsf{F})$ are mutually independent, (b) and (c) are due to the retrieval correctness requirement and the fact that $A_{0:N-1}$ is a deterministic function of $(W_{0:K-1}, \mathsf{F})$, and (d) is by the definition of the retrieval rate.

To see the independence condition ***P1***, let us consider (d), and we can write

$$\begin{aligned}
&H(A^{[0]}_{0:N-1}|\mathsf{F})\\
&= \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1})\\
&\qquad\qquad \cdot H\left(A^{[0]}_{0:N-1}\middle| Q^{[0]}_{0:N-1} = q_{0:N-1}\right)\\
&= \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1})\\
&\qquad\qquad \cdot H\left(A^{(q_0)}_0, A^{(q_1)}_1, \ldots, A^{(q_{N-1})}_{N-1}\right)\\
&\overset{(e)}{\le} \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1}) \sum_{n=0}^{N-1} H\left(A^{(q_n)}_n\right)\\
&\le \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1}) \sum_{n=0}^{N-1} \ell_n \log_2|\mathcal{Y}|\\
&= \log_2|\mathcal{Y}| \sum_{n=0}^{N-1} \mathbb{E}(\ell_n)\\
&= \frac{L}{R}\log_2|\mathcal{X}|.
\end{aligned} \tag{77}$$

For the equality to hold, it is clear that (e) must be equality for any $q_{0:N-1}$ of non-zero probability, and thus the independence condition ***P1*** must hold for $k=0$. However, by choosing a permutation π on $\{0, 1, \ldots, K-1\}$ such that $\pi(k)=0$ and using the same line of proof, it can be concluded that the independence condition holds for all coding matrices of any given requested message. The proof is thus complete. ■

Proof: [Proof of Lemma 2] Without loss of generality, let us consider the identity permutation function $\pi(k)=k$. We can start by writing the following chain of information inequalities:

$$\begin{aligned}
&NI\left(W_{k:K-1}; A^{[k-1]}_{0:N-1}\middle| W_{0:k-1}, \mathsf{F}\right)\\
&\overset{(a)}{\ge} \sum_{n=0}^{N-1} I\left(W_{k:K-1}; A^{[k-1]}_n\middle| W_{0:k-1}, \mathsf{F}\right)\\
&\overset{(b)}{=} \sum_{n=0}^{N-1} I\left(W_{k:K-1}; A^{[k]}_n\middle| W_{0:k-1}, \mathsf{F}\right)\\
&\overset{(c)}{=} \sum_{n=0}^{N-1} H\left(A^{[k]}_n\middle| W_{0:k-1}, \mathsf{F}\right)\\
&\overset{(d)}{\ge} \sum_{n=0}^{N-1} H\left(A^{[k]}_n\middle| W_{0:k-1}, \mathsf{F}, A^{[k]}_{0:n-1}\right)\\
&= \sum_{n=0}^{N-1} I\left(W_{k:K-1}; A^{[k]}_n\middle| W_{0:k-1}, \mathsf{F}, A^{[k]}_{0:n-1}\right)\\
&= I\left(W_{k:K-1}; A^{[k]}_{0:N-1}\middle| W_{0:k-1}, \mathsf{F}\right)\\
&\overset{(e)}{=} I\left(W_{k:K-1}; W_k, A^{[k]}_{0:N-1}\middle| W_{0:k-1}, \mathsf{F}\right)\\
&= L\log_2|\mathcal{X}| + I\left(W_{k+1:K-1}; A^{[k]}_{0:N-1}\middle| W_{0:k}, \mathsf{F}\right),
\end{aligned} \tag{78}$$

where (c) is because the answers are deterministic functions of the messages and the random key F, (e) is due to the retrieval correctness requirement, and the equality (b) can be justified as follows. We can write that

$$\begin{aligned}
&I\left(W_{k:K-1}; A^{[k-1]}_n\middle| W_{0:k-1}, \mathsf{F}\right)\\
&= H\left(A^{[k-1]}_n\middle| W_{0:k-1}, \mathsf{F}\right) - H\left(A^{[k-1]}_n\middle| W_{0:K-1}, \mathsf{F}\right)\\
&= H\left(A^{[k-1]}_n\middle| W_{0:k-1}, \mathsf{F}\right)\\
&\overset{(f)}{=} H\left(A^{[k-1]}_n\middle| W_{0:k-1}, Q^{[k-1]}_n\right)\\
&\overset{(g)}{=} H\left(A^{[k]}_n\middle| W_{0:k-1}, Q^{[k]}_n\right)\\
&\overset{(h)}{=} H\left(A^{[k]}_n\middle| W_{0:k-1}, \mathsf{F}\right) - H\left(A^{[k]}_n\middle| W_{0:K-1}, \mathsf{F}\right)\\
&= I\left(W_{k:K-1}; A^{[k]}_n\middle| W_{0:k-1}, \mathsf{F}\right),
\end{aligned} \tag{79}$$

where (f) is due to the Markov string $(A^{[k]}_n, W_{0:K-1}) \leftrightarrow Q^{[k]}_n \leftrightarrow \mathsf{F}$, (g) is because of the privacy constraint, and (h) is because of the afore-mentioned Markov string and the fact that $Q^{[k]}_n$ is a deterministic function of F.

The inequalities (a) and (d) are due to the standard non-negativity property of mutual information. However, the necessary conditions stated in the lemma can be derived from

$$\begin{aligned}
0 &= H\left(A^{[0]}_{0:N-1}\middle|A^{[0]}_{n}, W_0, \mathsf{F}\right)\\
&= \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1}) H\left(A^{[0]}_{0:N-1}\middle|A^{[0]}_{n}, W_0, Q^{[0]}_{0:N-1} = q_{0:N-1}\right)\\
&\overset{(i)}{=} \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1}) H\left(W_{1:K-1}\cdot\left[G^{(q_0)}_{0|0}, G^{(q_1)}_{1|0}, \ldots, G^{(q_{N-1})}_{N-1|0}\right]\middle|W_{1:K-1}\cdot G^{(q_n)}_{n|0}, W_0, Q^{[0]}_{0:N-1} = q_{0:N-1}\right)\\
&\overset{(j)}{=} \sum_{q_{0:N-1}} \mathbf{Pr}(Q^{[0]}_{0:N-1} = q_{0:N-1}) H\left(W_{1:K-1}\cdot\left[G^{(q_0)}_{0|0}, G^{(q_1)}_{1|0}, \ldots, G^{(q_{N-1})}_{N-1|0}\right]\middle|W_{1:K-1}\cdot G^{(q_n)}_{n|0}\right). \qquad (80)
\end{aligned}$$

$$\begin{aligned}
0 &= I\left(A^{[K-1]}_{N-1}; A^{[K-1]}_{0:N-2}\middle|W_{0:K-2}, Q^{[K-1]}_{0:N-1} = q_{0:N-1}\right)\\
&= I\left(W_{K-1}\cdot G^{(q_{N-1})}_{N-1|0:K-2}; W_{K-1}\cdot\left[G^{(q_0)}_{0|0:K-2}, G^{(q_1)}_{1|0:K-2}, \ldots, G^{(q_{N-2})}_{N-2|0:K-2}\right]\right). \qquad (82)
\end{aligned}$$

these two inequalities. First consider when (a) is equality, from which we have that any decomposable code must satisfy the condition in (80), shown at the top of this page, where in (i) we have utilized the fact that the component functions $W_0\cdot G^{(q_n)}_{n|1:K-1}$ can be meaningfully subtracted from the answers in the abelian group, and (j) is because W_0 is now independent of everything else after the corresponding component functions are eliminated in the answers, and the dependence on $q_{0:N-1}$ is fully absorbed in the answer function matrix $G^{q_{0:N-1}}_n$. This implies that for any set of queries $Q^{[0]}_{0:N-1} = q_{0:N-1}$ with a non-zero probability,

$$H\left(W_{1:K-1}\cdot\left[G^{(q_0)}_{0|0}, G^{(q_1)}_{1|0}, \ldots, G^{(q_{N-1})}_{N-1|0}\right]\middle|W_{1:K-1}\cdot G^{(q_n)}_{n|0}\right) = 0, \quad n \in \{0, 1, \ldots, N-1\}. \qquad (81)$$

This indeed implies that $W_{1:K-1}\cdot G^{(q_n)}_{n|0}$ can determine any $W_{1:K-1}\cdot G^{(q_{n'})}_{n'|0}$, for $n, n' \in \{0, 1, \ldots, N-1\}$. Since the query can be other than for the message W_0 (by taking a different permutation $\pi(\cdot)$ in the lemma), it follows that the deterministic property ***P2*** indeed holds.

Next consider (d), particularly for $k = K-1$ and the summand for $n = N-1$. For decomposable codes, the inequality being equality implies (82), shown at the top of this page, which further implies the independence between the random variables $W_{K-1}\cdot G^{(q_{N-1})}_{N-1|0:K-2}$ and $W_{K-1}\cdot\left[G^{(q_0)}_{0|0:K-2}, G^{(q_0)}_{1|0:K-2}, \cdot, G^{(q_{N-2})}_{N-2|0:K-2}\right]$. Since in the above argument, we can choose any value k in (d), and take any other order in the summation on both sides of (d), indeed the stated independence property ***P3*** holds. The proof is now complete. ■

References

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

[2] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[3] Q. Wang and M. Skoglund, "Symmetric private information retrieval for MDS coded distributed storage," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[4] R. Tandon, "The capacity of cache aided private information retrieval," Jun. 2017, *arXiv:1706.07035*. [Online]. Available: https://arxiv.org/abs/1706.07035

[5] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nov. 2015, pp. 2842–2846.

[6] S. R. Blackburn, T. Etzion, and M. B. Paterson, "PIR schemes with small download complexity and low storage requirements," Sep. 2016, *arXiv:1609.07027*. [Online]. Available: https://arxiv.org/abs/1609.07027

[7] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[8] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM J. Appl. Algebra Geometry*, vol. 1, no. 1, pp. 647–664, Nov. 2017.

[9] Y. Zhang, X. Wang, H. Wei, and G. Ge, "On private information retrieval array codes," Sep. 2016, *arXiv:1609.09167*. [Online]. Available: https://arxiv.org/abs/1609.09167

[10] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 856–860.

[11] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nov. 2015, pp. 2852–2856.

[12] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

[13] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[14] S. Kumar, E. Rosnes, and A. G. I. Amat, "Private information retrieval in distributed storage systems using an arbitrary linear code," Dec. 2016, *arXiv:1612.07084*. [Online]. Available: https://arxiv.org/abs/1612.07084

[15] K. Banawan and S. Ulukus, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[16] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[17] K. Banawan and S. Ulukus, "Asymmetry hurts: Private information retrieval under asymmetric traffic constraints," Jan. 2018, *arXiv:1801.03079*. [Online]. Available: https://arxiv.org/abs/1801.03079

[18] Q. Wang and M. Skoglund, "Secure private information retrieval from colluding databases with eavesdroppers," Oct. 2017, *arXiv:1710.01190*. [Online]. Available: https://arxiv.org/abs/1710.01190

[19] H.-Y. Lin, S. Kumar, E. Rosnes, and A. G. I. Amat, "An MDS-PIR capacity-achieving protocol for distributed storage using non-MDS linear codes," Jan. 2018, *arXiv:1801.04923*. [Online]. Available: https://arxiv.org/abs/1801.04923

[20] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. E. Rouayheb, "Private information retrieval schemes for coded data with arbitrary collusion patterns," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1908–1912.

[21] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2920–2932, Dec. 2017.

[22] C. Tian, H. Sun, and J. Chen, "A Shannon-theoretic approach to the storage-retrieval tradeoff in PIR systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1904–1908.

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

[24] A. Ambainis, "Upper bound on the communication complexity of private information retrieval," in *Proc. Int. Colloq. Automata, Lang., Program.*, Jul. 1997, pp. 401–407.

[25] A. Beimel, Y. Ishai, E. Kushilevitz, and J.-F. Raymond, "Breaking the $O(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval," in *Proc. 43rd Annu. IEEE Symp. Found. Comput. Sci.*, Nov. 2002, pp. 261–270.

[26] K. Efremenko, "3-query locally decodable codes of subexponential length," *SIAM J. Comput.*, vol. 41, no. 6, pp. 1694–1703, Dec. 2012.

[27] S. Yekhanin, "Towards 3-query locally decodable codes of subexponential length," *J. ACM*, vol. 55, no. 1, pp. 1–16, Feb. 2008.

[28] Z. Dvir and S. Gopi, "2-server PIR with sub-polynomial communication," in *Proc. 47th Annu. ACM Symp. Theory Comput. (STOC)*, Jun. 2015, pp. 577–584.

[29] Y. Ishai and E. Kushilevitz, "On the hardness of information-theoretic multiparty computation," in *Advances in Cryptology–EUROCRYPT*. Berlin, Germany: Springer-Verlag, 2004, pp. 439–455.

[30] J. Connelly and K. Zeger, "Linear network coding over rings—Part I: Scalar codes and commutative alphabets," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 274–291, Jan. 2018.

[31] J. Connelly and K. Zeger, "Linear network coding over rings—Part II: Vector codes and non-commutative alphabets," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 292–308, Jan. 2018.

[32] S. Lin and D. J. Costello, *Error Control Coding*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2004.

[33] B. Sasidharan, M. Vajha, and P. V. Kumar, "An explicit, coupled-layer construction of a high-rate MSR code with low sub-packetization level, small field size and all-node repair," Jul. 2016, *arXiv:1607.07335*. [Online]. Available: https://arxiv.org/abs/1607.07335

[34] M. Ye and A. Barg, "Explicit constructions of optimal-access MDS codes with nearly optimal sub-packetization," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6307–6317, Oct. 2017.

[35] S. Goparaju, I. Tamo, and R. Calderbank, "An improved sub-packetization bound for minimum storage regenerating codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2770–2779, May 2014.

[36] J. Li, X. Tang, and C. Tian, "A generic transformation to enable optimal repair in MDS codes for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6257–6267, Sep. 2018.

[37] S. B. Balaji and P. V. Kumar, "A tight lower bound on the sub-packetization level of optimal-access MSR and MDS codes," Oct. 2017, *arXiv:1710.05876*. [Online]. Available: https://arxiv.org/abs/1710.05876

[38] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, Oct. 2016.

[39] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5821–5833, Sep. 2017.

[40] L. Tang and A. Ramamoorthy, "Coded caching schemes with reduced subpacketization from linear block codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 3099–3120, Apr. 2018.

[41] C. Tian, J. Chen, S. N. Diggavi, and S. Shamai (Shitz), "Matched multiuser Gaussian source channel communications via uncoded schemes," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4155–4171, Jul. 2017.

[42] C. Tian, "Characterizing the rate region of the (4, 3, 3) exact-repair regenerating codes," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 5, pp. 967–975, May 2014.

[43] C. Tian, "Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching," *Entropy*, vol. 20, no. 8, pp. 603.1–603.43, Aug. 2018.

[44] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," Mar. 2019, *arXiv:1903.08229*. [Online]. Available: https://arxiv.org/abs/1903.08229

Chao Tian (S'00–M'05–SM'12) received the B.E. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2000 and the M.S. and Ph. D. degrees in Electrical and Computer Engineering from Cornell University, Ithaca, NY in 2003 and 2005, respectively. Dr. Tian was a postdoctoral researcher at Ecole Polytechnique Federale de Lausanne (EPFL) from 2005 to 2007, a member of technical staff–research at AT&T Labs–Research in New Jersey from 2007 to 2014, and an Associate Professor in the Department of Electrical Engineering and Computer Science at the University of Tennessee Knoxville from 2014 to 2017. He joined the Department of Electrical and Computer Engineering at Texas A&M University in 2017. His research interests include data storage systems, multi-user information theory, joint source-channel coding, signal processing, and compute algorithms.

Dr. Tian received the Liu Memorial Award at Cornell University in 2004, AT&T Key Contributor Award in 2010, 2011 and 2013. His authored and co-authored papers received the 2014 IEEE ComSoc DSTC Data Storage Best Paper Award and the 2017 IEEE Jack Keil Wolf ISIT Student Paper Award. He was an Associate Editor for the IEEE Signal Processing Letters from 2012 to 2014, and is currently an Editor for the IEEE Transactions on Communications.

Hua Sun (S'12–M'17) received his B.E. in Communications Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2011, M.S. in Electrical and Computer Engineering from University of California Irvine, USA, in 2013, and Ph.D. in Electrical Engineering from University of California Irvine, USA, in 2017. He is an Assistant Professor in the Department of Electrical Engineering at the University of North Texas, USA. His research interests include information theory and its applications to communications, privacy, networking, and storage.

Dr. Sun received the IEEE Jack Keil Wolf ISIT Student Paper Award in 2016, an IEEE GLOBECOM Best Paper Award in 2016, and the University of California Irvine CPCC Fellowship for the year 2011-2012.

Jun Chen (S'03–M'06–SM'16) received the B.E. in communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001 and the M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University, Ithaca, NY, in 2004 and 2006, respectively.

He was a Postdoctoral Research Associate in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign, Urbana, IL, from September 2005 to July 2006, and a Postdoctoral Fellow at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, from July 2006 to August 2007. Since September 2007 he has been with the Department of Electrical and Computer Engineering at McMaster University, Hamilton, ON, Canada, where he is currently an Associate Professor. His research interests include information theory, machine learning, wireless communications, and signal processing.

He held the title of the Barber-Gennum Chair in Information Technology from 2008 to 2013 and the Joseph Ip Distinguished Engineering Fellow from 2016 to 2018. He was a recipient of the Josef Raviv Memorial Postdoctoral Fellowship (2006), the Early Researcher Award from the Province of Ontario (2010), and the IBM Faculty Award (2010). He served as an Associate Editor for the IEEE Transcations on Information Theory from 2014 to 2016.