EPFL

Thèse n° 7213

Cooperative Data Exchange and Private Information Retrieval

Présentée le 30 octobre 2020

à la Faculté informatique et communications
Laboratoire d'information dans les systèmes en réseaux
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Su LI

Acceptée sur proposition du jury

Prof. E. Telatar, président du jury
Prof. M. C. Gastpar, directeur de thèse
Prof. A. Sprintson, rapporteur
Prof. C. Hollanti, rapporteuse
Dr O. Lévêque, rapporteur

École polytechnique fédérale de Lausanne

2020

Acknowledgments

I would like to express my greatest gratitude to my supervisor, Prof. Michael Gastpar. Without him, I would never be able to finish my doctoral study and this thesis could not have been accomplished. For study, he gives me total freedom to choose the research topics which I am interested in and always encourages me to try different things. One typical scenario is that I come to him with a not very well thought idea and he probably doubts it. Instead of telling me it would not work, he gives me time to work on it and normally he would say "I will be curious to see if it works". He also has a very unique perspective to think about the potential value of the research problems and gives me insightful advises. For daily life, he keeps the perfect distance with students. He likes to share his experience and stories with us in our coffee break. Although he cares about everyone in the group and is willing to provide help, he never steps into my personal life unless I ask for his advise or help. He is probably the best supervisor that one can hope. He is not only my supervisor, but more like 师父 , someone I can always learn from and trust but still be awed by.

I am also thankful to my thesis committee, Prof. Emre Telatar, Dr. Lévêque Olivier, Prof. Alex Sprintson and Prof. Camilla Hollanti for their helpful comments and suggestions on my thesis.

I would like to particularly thank our secretary, France Faille. She is so thoughtful and without her, my life in EPFL would be much harder. She helps me with almost anything that I could imagine, including organizing trips to conferences, renting rooms, learning French, health insurance, tax refund and so on.

I am very fortunate to join LINX, with very nice labmates: Jingge Zhu, Giel Op't Veld, Saeid Sahraei, Erixhen Sula, and Amedeo Esposito. My first connection with LINX was doing semester project during my master study with Jingge and he gave me many good advises. Later in the first time I travel to the US alone to attend a conference, Jingge was very nice to drive me around and it was a enjoyable memory. Giel is a good photographer and shared many experience on traveling with me. Saeid has a great sense of humor. It always takes me sometime to realize how funny his jokes are. Erixhen and I traveled together several times and shared many funny moments. It was a pity that we could not travel to NYC and LA as we planned due to the pandemic. Amedeo has great cooking skills and it was fun that we spent one day learning skiing together in the Children's place. Yanina Shkel and Aditya Pradeep recently joined the lab and make our online coffee break more pleasant. I also had many

memorable moments with IPG members and appreciate the help from them.

I am also grateful to my friends met in Switzerland who make my life here much more colorful and delightful.

Finally, I want to thank my family for their consistent support. This thesis is dedicated to them.

Lausanne, July 16, 2020 Su Li.

Abstract

Coding techniques have been well studied and used for improving communication quality by combating noise and mitigating interference. Recently, it has been shown that the same coding techniques can also be exploited to further improve communication performance and provide specific communication features even when the communication channel is ideal. In this thesis, we study two problems where coding techniques are used for improving communications in distributed systems and protecting the privacy of the client from untrusted servers, respectively.

The first part of this thesis studies the cooperative data exchange problem for fully connected networks. While many previous studies have shown that the problem can be solved by algorithms based on submodular function minimization, we tackle this problem via a concept we refer to as "conditioning basis", which is closely linked to linear coding schemes with particular additional properties. We show that such special linear coding schemes are optimal for the cooperative data exchange problem. Hence, by searching the existence of such a conditioning basis and special linear coding schemes, we can solve this problem with lower complexity. We propose a deterministic algorithm for this problem and briefly show how to construct the optimal linear coding schemes starting from a Vandermonde matrix. Moreover, we show that our new method can be used to solve two generalized problems, which are cooperative data exchange with weighted cost and successive local omniscience problems.

The second part of this thesis investigates the problem of private information retrieval with side information. Specifically, three different extensions are studied: multi-message, multi-server, and multi-user, respectively. For each problem, we provide a proof of the converse for the download rate as well as propose efficient approaches to construct optimal coding schemes. For multi-message and multi-server cases, we give closed-form expressions for the download rates and introduce two useful tools, *conditioning answer string* and *virtual private information*, to analyze the problem. For multi-user cases, we show that the optimal download rate can be obtained by solving an optimization problem over all partitions of the total number of messages and propose a novel algorithm based on dynamic programming to solve the optimization problem.

Keywords: cooperative data exchange, maximum distance separable codes, linear codes, private information retrieval, information-theoretic privacy, multi-message, multi-server,

multi-user.

Résumé

Les techniques de codage ont été bien étudiées et utilisées pour améliorer la qualité de la communication en luttant contre le bruit et en atténuant les interférences. Récemment, il a été démontré que les mêmes techniques de codage peuvent également être exploitées pour améliorer encore les performances de communication et fournir des fonctionnalités de communication spécifiques même lorsque le canal de communication est idéal. Dans cette thèse, nous étudions deux problèmes où les techniques de codage sont utilisées pour améliorer les communications dans les systèmes distribués et protéger la confidentialité du client contre les serveurs non fiables, respectivement.

La première partie de cette thèse étudie le problème de l'échange coopératif de données pour des réseaux entièrement connectés. Alors que de nombreuses études précédentes ont montré que le problème peut être résolu par des algorithmes basés sur la minimisation de la fonction submodulaire, nous abordons ce problème via un concept que nous appelons "base de conditionnement", qui est étroitement lié aux schémas de codage linéaire avec des propriétés supplémentaires particulières. Nous montrons que de tels schémas spéciaux de codage linéaire sont optimaux pour le problème d'échange de données coopératif. Par conséquent, en recherchant l'existence d'une telle base de conditionnement et de schémas de codage linéaire spéciaux, nous pouvons résoudre ce problème avec une complexité moindre. Nous proposons un algorithme déterministe pour ce problème et montrons brièvement comment construire les schémas de codage linéaire optimaux à partir d'une matrice de Vandermonde. De plus, nous montrons que notre nouvelle méthode peut être utilisée pour résoudre deux problèmes généralisés, qui sont l'échange de données coopératif à coût pondéré et les problèmes d'omniscience locale successifs.

La deuxième partie de cette thèse examine le problème de la recherche d'informations privées avec des informations secondaires. Plus précisément, trois extensions différentes sont étudiées : multi-message, multi-serveur et multi-utilisateur, respectivement. Pour chaque problème, nous fournissons une preuve de l'inverse du taux de téléchargement et proposons des approches efficaces pour construire des schémas de codage optimaux. Pour les cas multi-messages et multi-serveurs, nous donnons des expressions de forme fermée pour les taux de téléchargement et introduisons deux outils utiles, it conditioning answer string et it virtual private information, pour analyser le problème. Pour les cas multi-utilisateurs, nous montrons que le taux de téléchargement optimal peut être obtenu en résolvant un problème d'optimisa-

tion sur toutes les partitions du nombre total de messages et proposons un nouvel algorithme basé sur une programmation dynamique pour résoudre le problème d'optimisation.
Mots clefs: échange de données coopératif, codes séparables à distance maximale, codes linéaires, récupération d'informations privées, confidentialité de la théorie de l'information, multi-message, multi-serveur, multi-utilisateur.

Contents

1 Introduction

After the fundamental limits of the point-to-point communication were established by Shannon [1], over the last seven decades, communication technologies have developed expeditiously and changed our daily life comprehensively. We have witnessed a proliferation of researches and studies about making communication faster (higher data rate) and more reliable (better noise tolerance). As the number of participants in communication is increasing in many applications, the potential of cooperative communications has attracted hefty attention. In particular, for distributed storage systems, the servers have to periodically synchronize their local data with each other. An efficient communication protocol for such synchronization requires the servers to generate transmissions cooperatively. And this is the first problem we study in this thesis, Cooperative Data Exchange. Besides the speed and reliability of communication, safety and privacy in communication have become increasingly important than ever before. In the current data era, many data analysis techniques are invented and improved, which conversely grows the demand for secure and private communication. The second problem investigated in this thesis, Private Information Retrieval, is about protecting the privacy of the client(s) from the untrusted servers.

1.1 Cooperative Data Exchange

Consider a fully connected network composed of N nodes that all want to recover a file consisting of K packets. Each node initially only has a subset of the packets. Each node can generate coded packets by using its locally available packets and transmit them to other nodes through a lossless broadcast channel, i.e. all other nodes receive the coded packets. The goal is for each node to assemble the full file. The key questions are: (1) What is the minimum number of required transmissions? (2) What should individual nodes transmit? This problem was introduced by El Rouayheb *et al.* in [2] and is referred to as *Cooperative Data Exchange* (CDE) for the fully connected network. The data exchange problem is also related to the problem of secret key generation introduced by Csiszár *et al.* in [3]. Concerning the minimum number of required transmissions, upper and lower bounds were established in [2]. A deterministic algorithm was proposed to produce a coding scheme which achieves universal recovery

using at most twice the minimum number of required transmissions. The CDE problem can be formulated as an Integer Linear Program (ILP) with the Slepian-Wolf constraints on all proper subsets of the nodes' available packet information. A randomized algorithm [4] and a deterministic algorithm [5] were proposed to give an approximate solution and optimal solution to the CDE problem. We note that the number of constraints in the ILP at hand grows exponentially with the problem size. Nevertheless, exact polynomial-time algorithms were found in [5, 6, 7, 8] based on minimizing submodular functions and subgradient optimization. It was also shown that the total number of transmissions can be reduced if each packet is split into sub-packets (resulting in non-integer rates), but that splitting into $N-1$ sub-packets is sufficient to attain optimal performance [7, 9]. Therefore, we may simply consider the sub-packets to be our packets and it is unnecessary to explicitly discuss the case of split packets.

The CDE problem was extended to general network topologies, and it was shown that linear codes are sufficient to optimally solve the CDE problem in [7, 10]. However, the same work also revealed that for arbitrarily connected networks, the CDE problem is NP-hard and cannot be solved exactly with polynomial-time algorithms. Many extensions of the CDE problem have also been studied. In [11], the nodes are divided into two classes, high and low priority. The resulting CDE problem with priorities was formulated as a multi-objective integer linear program. Assuming a uniformly random packet distribution and restricting to the limit as the number of packets tends to infinity, a closed-form expression for the minimal number of required transmissions was derived. In [12], successive omniscience is studied, where subsets of users first recover packets within each subset and then recover packets of users in other subsets. In [6, 13], transmissions sent by different nodes are considered to have different costs. Instead of minimizing the total number of transmissions, the goal becomes minimizing the total cost, *i.e.,* a weighted sum of the transmissions. To solve the CDE problem with weighted cost, a deterministic polynomial algorithm based on submodular function minimization was proposed in [6], while a randomized greedy algorithm was proposed in [13]. In [7, 9], it is assumed that each packet can be split into the same number of smaller chunks and the optimization goal is minimizing the normalized total number of transmissions. Intuitively, the larger the number of chunks we split each packet into, the smaller the normalized total number of transmissions that can be achieved, and it has been proved that it is sufficient to split each packet into $N-1$ chunks. In [14], the nodes are divided into two classes, reliable and unreliable. For unreliable nodes, the initially available packets are unknown (but it is known how many packets they have) and the packet transmissions are subject to arbitrary erasures. A closed-form expression for the minimal number of transmissions for the case of only a single unreliable node was derived with probability approaching 1 as the number of packets tends to infinity. For more than one unreliable node, an approximate solution was provided.

The CDE problem for the fully connected network is also related to the secret key generation problem, which was introduced in [3] and was formulated as a maximization problem over all partitions of the set of nodes. Tyagi *et al.* [15] leveraged this to derive an algorithm that achieves local omniscience in each step and outputs a solution for secret key generation. Courtade *et*

al. studied the CDE problem with the goal of generating secret key in [16]. The weakly secure data exchange problem was introduced in [17, 18]. The goal is to achieve universal recovery while revealing as little information as possible. In contrast to the coding scheme in [17] in which each transmission is a linear combination of as many packets as possible, our scheme considers a fixed number of packets for every transmission.

Contributions

In the cooperative data exchange part of this thesis, we study the CDE problem for the fully connected network from a new perspective. Our main contributions can be summarized as follows:

1. We present a new deterministic algorithm to compute the minimal number of required transmissions. It is based on searching for the existence of certain conditional bases of the packet distribution matrix. The complexity is bounded by $\mathcal{O}(N^3K^3\log(K))$, significantly lower than the complexity of the best known existing algorithms proposed in [6] based on minimizing submodular functions $\mathcal{O}((N^6K^3+N^7)\log(K))$ and based on subgradient methods $\mathcal{O}((N^4\log(N)+N^4K^3)K^2\log(K))$.

2. We establish the existence of coding schemes with the special property that each transmission is a linear combination of exactly $d+1$ packets, for any $0 \le d < K$. The scheme involves a total of $K-d$ transmissions and enables all nodes with at least d packets to recover their missing packets (irrespective of which d packets they had to begin with). The proposed scheme works if coding occurs over a finite field of large enough size, and is related to distributed Reed-Solomon codes. Using a standard approach, we briefly show that the scheme can be constructed deterministically from Vandermonde matrices. Note that the equally important question concerning the existence of coding schemes restricted to computations over *small* finite fields is left open.

3. We extend our approach to the CDE problem with a weighted cost objective function and to the successive local omniscience problem. For the former, the minimal number of required transmissions can be found in complexity no larger than $\mathcal{O}(N^3K^3\log(K))$, which is the same as the CDE problem without weighted cost in Item (1) above. For the successive local omniscience problem with M priority groups, our method has complexity bounded by $\mathcal{O}(N^3K^3\log(K))$. For both problems, the coding schemes are constructed by analogy to the basic CDE problem, and the consideration is again limited to the case of computations over sufficiently large finite fields.

1.2 Private Information Retrieval

In the original Privet Information Retrieval (PIR) problem, one user wants to download one message from a database which is stored at a single server, while keeping the index of the

desired message private from the server. The user generates and sends queries to the server and the server replies coded messages as the user requests. PIR requires that the server is not able to infer any information about the index of the message which the user wants to download. To find the optimal solutions to the PIR problem, we need to find the minimum number of required download bits which should be sent to the user by the server and construct the optimal coding schemes which can be used by the user to decode the demand message and reveal no information about the index of demand message to the servers.

The PIR problem was first introduced from the perspective of computational complexity [19, 20]. In recent years, we have witnessed an escalation of studies of the PIR problem from an information-theoretic point of view [21, 22, 23]. To achieve information-theoretic privacy in the original PIR problem, the user has to download all messages of the database. If one assumes that the database is stored in *multiple* servers, the problem becomes more interesting and has attracted considerable attention. By exploiting the advantages of replications of the database in multiple non-colluding servers, private information retrieval can be achieved without downloading all messages and the information-theoretic capacity of this problem is characterized in [22]. Many variations of PIR have been studied by ensuing work, including databases coded by erasure codes [24, 25, 26, 27, 28, 29, 30, 31, 32], partially colluding or adversarial servers [23, 24, 33, 34, 35, 36, 37, 38], symmetric PIR [39, 40, 41], side information messages available at users [42, 43, 44, 45, 46, 47, 48, 49], cache aided side information [50, 51, 52], multiple messages [53, 54, 55, 56], multi-user [44, 57], and private function computation [58, 59].

The problem of PIR with side information was first studied in [42], where the user wants to download one message from a single server while it already has some messages as side information. Two types of privacy were defined: (i) W-Privacy: only the index of demand message should be private and (ii) (W,S)-Privacy: both indices of demand and side information messages should be private, which is also referred to as private side information in other works [43, 45]. The minimum number of required transmissions and the optimal coding scheme for single-server cases for both W-Privacy and (W,S)-Privacy were found in [42]. For the multi-server extension of PIR with side information, (W,S)-Privacy problem was solved in [43, 60]. For the W-Privacy problem, a novel PIR coding scheme based on *super-messages* was proposed in [42], though without proof of optimality. In PIR with side information, the user is assumed to have complete messages as side information, which was later extended to the cases where linearly coded messages can be the side information [46, 60]. Recently, the single-server PIR with side information has been shown to be closely related to locally recoverable codes [61].

In [53], Banawan and Ulukus consider the problem that the user wants to download multiple messages from multiple servers, but there is no side information at the user. In [54], Shariatpanahi *et al.* study the multi-message PIR problem with side information and the user wants to protect both the privacy of the indices of demand messages and of the side information messages. In our problem, the user is only interested in protecting the privacy of the indices of

the demand messages, which is a more challenging problem than protecting both the indices of the demand and side information messages. The single-server multi-message PIR with side information problem is studied concurrently in [56].

Contributions

In the private information retrieval part of this thesis, we study three different extensions of PIR with side information, which are single-server multi-message, multi-server single-message, and single-server multi-user. Our main contributions can be summarized as follows:

1. **Single-server Multi-message PIR with Side Information:** We present a closed-form expression for the minimum number of required download bits and propose a novel method to construct optimal coding schemes. Hence, we establish the capacity for the single-server multi-message PIR with side information problem. We also show that the trivial MDS coding scheme with $K-M$ normalized number of download bits is optimal when $N > M$ or $N^2+N \geq K-M$[1]. We introduced a novel conception, *conditional answer string*, which captures the special property of PIR with side information and is used in deriving the converse bound.

2. **Multi-server Single-message PIR with Side Information:** We prove an information-theoretic converse bound for the capacity of multi-server single-message PIR with side information and W-Privacy. The coding scheme proposed in [42] matches our converse bound. Thus, our work establishes the capacity of this problem. When the number of servers equals 1, our result matches the capacity of single-server PIR with side information and W-Privacy characterized in [42]. When the number of side information message equals 0, our result matches the capacity of multi-server PIR without side information characterized in [22]. We introduce a novel conception that we refer to as *virtual side information*, which represent the multi-server effect in PIR with side information and is used in the proof of the converse bound.

3. **Single-server Multi-user PIR with Side Information:** We derive necessary conditions for linear coding schemes that satisfy the privacy condition of the PIR problem for the single-server multi-user cases. Based on these necessary conditions, we give an achievable lower bound on the number of required transmissions for generating linear coding schemes. We present a novel method to construct linear coding schemes that satisfy the requirements of PIR and use the minimal number of transmissions. Our proof method has two steps: we first partition the messages into several subsets and then generate linear combinations of messages within each subset. We show that the search over all message partitions can be carried out via a dynamic programming algorithm of complexity $\mathcal{O}(K^2)$.

[1] K is the total number of messages, M is the number of side information messages, and N is the number of demand messages.

1.3 Notations

For any vector X, the i-th entry of X is denoted as X_i. For two integer $i < j$, the notation $i : j$ denotes the integer set $\{i, i+1, \ldots, j\}$. We denote random variables and their realizations by bold-face and regular letters, respectively. We denote probability, conditional probability, (Shannon) Entropy, conditional entropy and mutual information by $\Pr(\cdot)$, $\Pr(\cdot|\cdot)$, $H(\cdot)$, $H(\cdot|\cdot)$ and $I(\cdot|\cdot)$. For any integer $i \leq j$, let $W_i^j = \bigcup_{l=i}^{j} \{W_l\}$. The ceiling and floor operator are denoted by $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$, respectively. More definitions for locally used notations are given in each section.

2 Preliminaries

Cooperative data exchange and private information retrieval are very fresh topics in network information theory. To solve them, we require knowledge from both information theory and coding theory. In this chapter, we give some necessary background information on both topics as well as the maximum distance separable codes, which are closely related to both cooperative data exchange and private information retrieval.

2.1 Maximum Distance Separable Matrix

In coding theory, the Singleton bound [62] indicates that for any linear code $\mathscr{C}$ over a finite field $\mathbb{F}_q$, with block length n, dimension k and minimum distance d, the maximum number of codewords satisfies

$$q^k \leq q^{n-d+1}. \tag{2.1}$$

Equivalently, the minimum distance d satisfies

$$d \leq n-k+1. \tag{2.2}$$

The linear codes which achieve equality in Equation (2.2) are referred to as the Maximum Distance Separable (MDS) codes. The most widely used examples of MDS codes are the Reed-Solomon codes. Let $\mathbf{G}$ denote the generator matrix of any MDS code. Then one of the many useful properties of $\mathbf{G}$ is that any k columns are linear independent. Let $X = [X_1, X_2, \ldots, X_k]^\mathsf{T} \in \mathbb{F}_q^k$ denote the message and the codewords for X, denoted by $\mathscr{C}(X) = [\mathscr{C}(X)_1, \mathscr{C}(X)_2, \ldots, \mathscr{C}(X)_n]^\mathsf{T}$,

can be expressed as follows

$$\begin{bmatrix} \mathscr{C}(X)_1 \\ \mathscr{C}(X)_2 \\ \vdots \\ \mathscr{C}(X)_{n-1} \\ \mathscr{C}(X)_n \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{G}_{1,1} & \mathbf{G}_{1,2} & \dots & \mathbf{G}_{1,k} & \dots & \mathbf{G}_{1,n} \\ \mathbf{G}_{2,1} & \mathbf{G}_{2,2} & \dots & \mathbf{G}_{2,k} & \dots & \mathbf{G}_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{G}_{k,1} & \mathbf{G}_{k,2} & \dots & \mathbf{G}_{k,k} & \dots & \mathbf{G}_{k,n} \end{bmatrix}^{\mathsf{T}}}_{\mathbf{G}} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}. \tag{2.3}$$

If we know any $n-k$ symbols in the message X, then the other symbols can be fully recovered from the codeword $\mathscr{C}(X)$. The matrix $\mathbf{G}$ is also called the MDS matrix, which will be used in both cooperative data exchange and private information retrieval problems.

2.2 CDE based on Submodular Function Minimization

In the cooperative data exchange for the fully connected network problem, the N clients want to cooperatively recover a common file consisting of K packets. Each client initially has some packets as side information, and generates coded messages and sends them to other clients by ideal broadcast channel. At the end of the communication, all clients should successfully recover the file, which is referred to as *universal recovery*. For any $i \in \{1,\dots,N\}$, let $\mathbf{X}_i$ denote the set of packets locally available at client i. Let $\mathbf{r} = [r_1, r_2, \dots, r_N]$ denote the rate vector. Then for any subset of clients, $\mathscr{I} \subseteq \{1,\dots,N\}$, the number of collectively available packets is $|\cup_{i\in\mathscr{I}} \mathbf{X}_i|$. Hence for clients in $\mathscr{I}$ to fully recover the common file, they should receive at least $K - |\cup_{i\in\mathscr{I}} \mathbf{X}_i|$ packets from the other clients. A sufficient and necessary condition for universal recovery is $\forall \mathscr{I} \subsetneq \{1,\dots,N\}$:

$$\sum_{i\in\{1,\dots,N\}\setminus\mathscr{I}} r_i \geq K - |\cup_{i\in\mathscr{I}} \mathbf{X}_i| = |\cup_{i\in\mathscr{I}} \mathbf{X}_i^c| \tag{2.4}$$

Since for $\mathscr{I} = \emptyset$ and $\mathscr{I} = \{1,\dots,N\}$, the numbers of required transmissions are K and 0, respectively, the number of constraints is actually $2^N - 2$.

Definition 2.1 (Submodular Function)**.** *A set of function $f : 2^N \to \mathbb{Z}$ is submodular if $\forall U, V \subseteq \{1,\dots,N\} s.t. U \cap V \neq \emptyset$:*

$$f(U) + f(V) \geq f(U \cup V) + f(U \cap V) \tag{2.5}$$

Let $f_\beta(\mathscr{I}) = \beta - |\cup_{i\in\mathscr{I}} \mathbf{X}_i^c|$. It can be verified that the function $f_\beta(\mathscr{I})$ is submodular. In [6], it has been shown that by using the Submodular Function Minimization (SFM), it is not necessary to check all $2^N - 2$ constraints. The algorithm based on SFM can be used to check whether any sum rate $R = r_1 + \cdots + r_N$ is feasible for universal recovery. Hence, the complexity of such an approach depends on the complexity of SFM. The complexity of the currently best SFM algorithm proposed by Orlin in [63] is $\mathscr{O}(N^5K^3 + N^6)$. As the SFM needs to be used for all clients and the binary searching method is used for finding the optimal sum rate, the overall

complexity of the algorithm proposed in [6] is $\mathcal{O}(N(N^5K^3+N^6)\log(K))$. Our (d,K)-Basis method does not rely on SFM and has low complexity since we exploit the special structure of the linear codes. We also note that the SFM is also an ongoing research topic, whose complexity may be further reduced in the future.

2.3 Private Information Retrieval

While computational PIR can be achieved for single server scenarios by utilizing one-way functions, the information-theoretical PIR for single server cases requires the download of the full database. To make the problem nontrivial, two directions are considered: multiple non-colluding or partially colluding servers PIR and single-server PIR with side information.

The notations widely used in PIR are not very trivial. Normally, we use $Q^{[W]}$ and $A^{[W]}$ to denote the query and answer string generated for demand index W, respectively. Similarly, in PIR with side information, we use $Q^{[W,S]}$ and $A^{[W,S]}$ to denote the query and answer string generated for demand index W and side information indices S, respectively. Since $Q^{[W]}$ (or $Q^{[W,S]}$) is typically a stochastic function of W (or W,S), given W (or W,S), the user has to randomly choose one query from multiple candidate queries. Thus, given W (or W,S), we usually use $\mathbf{Q}^{[W]}$ (or $\mathbf{Q}^{[W,S]}$) to denote the random variable of the query. Meanwhile, the answer string $A^{[W]}$ (or $A^{[W,S]}$) is a deterministic function of query $Q^{[W]}$ (or $Q^{[W,S]}$) and all messages $X_1,\ldots,X_K$. Since the messages are also random variables, denoted by $\mathbf{X}_1,\ldots,\mathbf{X}_K$, only given the query $\mathbf{Q}^{[W]}=Q^{[W]}$ (or $\mathbf{Q}^{[W,S]}=Q^{[W,S]}$), the answer string is still a random variable, which is denoted by $\mathbf{A}^{[W]}$ (or $\mathbf{A}^{[W,S]}$).

The multi-server PIR requires each server individually should not be able to infer any information about the demand index from the query and answer string. In multi-server PIR, for Server j, the query $Q_j^{[W]}$answer string $A_j^{[W]}$ generated for W should also be possible generated for another index W'. Hence, $Q_j^{[W]}$ and $Q_j^{[W']}$ are actually indistinguishable from the Server j's perspective. Hence, in the derivation of the converse bounds for the capacity, the key step is to replace $Q_j^{[W]}$ and $A_j^{[W]}$ with $Q_j^{[W']}$ and $A_j^{[W']}$, respectively. Therefore, $\mathbf{Q}^{[W]}$ and $\mathbf{A}^{[W]}$ are identically distributed as $\mathbf{Q}^{[W']}$ and $\mathbf{A}^{[W']}$, respectively.

The single-server PIR with side information requires the server should not be able to infer any information about the demand index. But the joint distribution of W and S is not needed to be private. Hence, for any query $Q^{[W,S]}$ generated for W and S, for any other index W', there must exist a corresponding S' such that $Q^{[W',S']}$ is indistinguishable from the server's perspective. The key step in the derivation of the converse bounds for the capacity for single-server PIR is to replace $Q^{[W,S]}$ and $A^{[W,S]}$ with $Q^{[W',S']}$ and $A^{[W',S']}$, respectively. Unlike the multi-server PIR without side information cases, we note that $\mathbf{Q}^{[W,S]}$ and $\mathbf{A}^{[W,S]}$ are not necessarily identically distributed as $\mathbf{Q}^{[W',S']}$ and $\mathbf{A}^{[W',S']}$, which is the main difference between the multi-server PIR without side information and single-server PIR with side information.

In [49], the authors proved the converse for the capacity of single-server single-message PIR

with side information by using the maximum acyclic induced graph, which is very clear and nice. However, we want to mention that the converse can also be proved by using a similar approach which is used in Chapter 4. For any specific query realization Q, suppose given $\mathbf{X}_{S_0}$, message $\mathbf{X}_{W_0}$ can be decoded from the answer string generated according to Q and additionally messages in set $\mathbf{X}_{U_0}$ can also be decoded, i.e.,

$$H(\mathbf{X}_{W_0\cup U_0}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{S_0})=0. \tag{2.6}$$

We use the notation $W_0 \cup U_0$ to denote the union set $\{W_0\} \cup U_0$ for ease of notations. The number of download bits (D) can be computed as follows.

$$\begin{aligned}
D =& H(\mathbf{A}|\mathbf{Q}=Q) && (2.7)\\
=& H(\mathbf{X}_{W_0\cup U_0\cup S_0},\mathbf{A}|\mathbf{Q}=Q) - H(\mathbf{X}_{W_0\cup U_0\cup S_0}|\mathbf{A},\mathbf{Q}=Q) && (2.8)\\
=& H(\mathbf{X}_{W_0\cup U_0\cup S_0}|\mathbf{Q}=Q) + H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{W_0\cup U_0\cup S_0}|\mathbf{A},\mathbf{Q}=Q) && (2.9)\\
=& H(\mathbf{X}_{W_0\cup U_0\cup S_0}) + H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{S_0}|\mathbf{A},\mathbf{Q}=Q) && (2.10)
\end{aligned}$$

According to the *privacy condition*, the server should not be able to infer the information of the demand index. Hence for any index $W_1 \in \{1,\dots,K\} \setminus (W_0 \cup U_0 \cup S_0)$, there must exist $S_1 \subseteq \{1,\dots,K\} \setminus \{W_1\}$ such that $H(\mathbf{X}_{W_1}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{S_1})=0$, otherwise the server knows W_1 is not the demand index, which violates the *privacy condition* [49, Proposition 1]. Thus, we can find corresponding W_1, U_1 and S_1 such that the download bits can be expressed as follows.

$$\begin{aligned}
D =& H(\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{S_0}|\mathbf{A},\mathbf{Q}=Q)\\
&+ H(\mathbf{X}_{W_1\cup U_1\cup S_1},\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{W_1\cup U_1\cup S_1}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0\cup U_0\cup S_0}) && (2.11)\\
=& H(\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{S_0}|\mathbf{A},\mathbf{Q}=Q)\\
&+ H(\mathbf{X}_{W_1\cup U_1\cup S_1}|\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{S_1}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^1\cup U_0^1\cup S_0})\\
&+ H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{W_0^1\cup U_0^1\cup S_0^1}) && (2.12)
\end{aligned}$$

Similarly, we can find W_i, U_i and S_i such that after T iterations, we have

$$W_0^T \cup U_0^T \cup S_0^T = \{1,\dots,K\}, \tag{2.13}$$

which implies that

$$H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{W_0^T\cup U_0^T\cup S_0^T}) = H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{1,\dots,K}) = 0. \tag{2.14}$$

Then, the number of download bits now can be written as

$$\begin{aligned}
D =& H(\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{S_0}|\mathbf{A},\mathbf{Q}=Q) \\
&+ H(\mathbf{X}_{W_1\cup U_1\cup S_1}|\mathbf{X}_{W_0\cup U_0\cup S_0}) - H(\mathbf{X}_{S_1}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^1\cup U_0^1\cup S_0}) \\
&+ \dots, \\
&+ H(\mathbf{X}_{W_T\cup U_T\cup S_T}|\mathbf{X}_{W_0^{T-1}\cup U_0^{T-1}\cup S_0^{T-1}}) - H(\mathbf{X}_{S_T}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^T\cup U_0^T\cup S_0^{T-1}}) \\
&+ H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{W_0^T\cup U_0^T\cup S_0^T}) && (2.15)\\
=& H(\mathbf{X}_{W_0^T\cup U_0^T\cup S_0^T}) - \sum_{i=0}^{T} H(\mathbf{X}_{S_i}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^i\cup U_0^i\cup S_0^{i-1}}) && (2.16)\\
=& KL - \sum_{i=0}^{T} H(\mathbf{X}_{S_i}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^i\cup U_0^i\cup S_0^{i-1}}) && (2.17)
\end{aligned}$$

By taking minimization over all choices of U_0^T and S_0^T, we can get

$$\begin{aligned}
D \geq& \min_{U_0^T,S_0^T} KL - \sum_{i=0}^{T} H(\mathbf{X}_{S_i}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^i\cup U_0^i\cup S_0^{i-1}}) && (2.18)\\
=& KL - \max_{U_0^T,S_0^T} \sum_{i=0}^{T} H(\mathbf{X}_{S_i}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^i\cup U_0^i\cup S_0^{i-1}}) && (2.19)
\end{aligned}$$

Apparently, the optimal choices for U_0^T is $\emptyset$[1]. This gives us that

$$\begin{aligned}
D \geq& KL - \max_{S_0^T} \sum_{i=0}^{T} H(\mathbf{X}_{S_i}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{W_0^i\cup S_0^{i-1}}) && (2.20)\\
\geq& KL - \max_{S_0^T} \sum_{i=0}^{T} H(\mathbf{X}_{S_i}) && (2.21)
\end{aligned}$$

Each $|S_i|$ for $i \in \{1,\dots,T\}$ is upper bounded by the number of side information M. According to Equation (2.13), we have

$$K = \sum_{i=0}^{T} |W_i| + |S_i| = T + 1 + \sum_{i=0}^{T} |S_i|, \tag{2.22}$$

In order to maximize $\sum_{i=0}^{T} H(\mathbf{X}_{S_i})$, one of the optimal choices could be

$$\begin{aligned}
T+1 =& \left\lceil \frac{K}{M+1} \right\rceil, && (2.23)\\
|S_0| =& |S_1| = \dots = |S_{T-1}| = M, && (2.24)\\
|S_T| =& K - (T-1)(M+1) - 1. && (2.25)
\end{aligned}$$

[1] This is only true for single-message cases. For multi-message cases, the optimal choices for U_0^T is not $\emptyset$.

The number of download bits can be lower-bounded by

$$D \geq KL - ((T-1)M + K - (T-1)(M+1) - 1)L \tag{2.26}$$

$$= (T-1)L \tag{2.27}$$

$$= \left\lceil \frac{K}{M+1} \right\rceil L \tag{2.28}$$

which is the same as the converse bound derived by the maximum acyclic induced graph in [49].

3 Cooperative Data Exchange based on MDS Codes

In the cooperative data exchange problem, multiple nodes want to recover a common file by communicating with each other. Each node is assumed to initially possess some part of the common file as side information. In this chapter, we study the cooperative data exchange problem for the fully connected network. We prove that the cooperative data exchange problem for the fully connected network can be solved by searching the existence of a conditional basis. We also proposed a novel algorithm to solve the problem with polynomial-time complexity. We present an approach to generate the optimal coding scheme, which has a particular feature that each of the transmission is a linear combination of the same number of packets. When coding occurs over a sufficiently large finite field, we also show how the coefficient of these linear combinations can be chosen by leveraging a connection to maximum distance separable codes. Moreover, we show that our method can be used to solve two extended versions of cooperative data exchange problems, which are cooperative data exchange with weighted cost and the so-called successive local omniscience problem.

3.1 Problem Statement

Consider a fully connected network which has N nodes and one file composed of K independent packets. Let $\mathscr{N} = \{1, \ldots, N\}$ and $\mathscr{P} = \{P_1, \ldots, P_K\}$ denote the set of nodes and set of packets, respectively. We assume that each packet $P_i \in \mathbb{F}$, where $\mathbb{F}$ is some finite field with large enough size. Without loss of generality, we assume that every packet is initially available at least at two nodes and at most at $N-1$ nodes[1]. The set of the packets initially available at node i ($i \in \mathscr{N}$) is denoted by $\mathbf{X}_i$ ($\mathbf{X}_i \subseteq \mathscr{P}$). The union set of the packets initially available at a subset of nodes $\mathscr{I} \subseteq \mathscr{N}$ is denoted by $\mathbf{X}_\mathscr{I} = \bigcup_{i\in\mathscr{I}} \mathbf{X}_i$. We assume that all the nodes collectively have all packets, which means $\mathbf{X}_\mathscr{N} = \mathscr{P}$. The notation $\mathbf{X}^c_\mathscr{I} = \mathscr{P} \setminus \mathbf{X}_\mathscr{I}$ denotes the jointly missing packets at nodes in set $\mathscr{I}$. Let $\mathscr{M} = \min_{i\in\mathscr{N}} |\mathbf{X}_i|$ be the minimum number of initially available packets at any single node.

[1] If there is a packet that is only initially available at one node, the optimal strategy is just letting that node send the uncoded packet to the others. If there is a packet that is available at all nodes, then no one needs to recover it.

Definition 3.1. *Define the packet distribution matrix E as the $N \times K$ matrix with entry at i^{th} row j^{th} column:*

$$E_{ij} = \begin{cases} 1, & P_j \in \mathbf{X}_i, \\ 0, & otherwise. \end{cases} \tag{3.1}$$

We refer to the K-dimensional binary (row) vector e_i, the i^{th} row of E, as the Packet Distribution Vector (PDV) of node i.

Let $\mathbf{T} = \{T_1, \ldots, T_R\}$ denote a linear[2] coding scheme with R transmissions, which means that each transmission T_i is a linear combination of packets available at the sender node. Let $\mathbf{r} = [r_1, \ldots, r_N]^\mathsf{T}$ denote the rate vector where each r_i is the number of transmissions made by node i. Hence, the total number of transmissions can be expressed as

$$R = \sum_{i=1}^{N} r_i. \tag{3.2}$$

Let R^* denote the minimum number of required transmissions.

Define the coefficient matrix A with entries $a_{ij} (\forall i \in \{1, \ldots, R\}, j \in \{1, \ldots, K\})$, and denote by $\alpha_i = [a_{i1}, \ldots, a_{iK}]$ and $\beta_j = [a_{1j}, \ldots, a_{Rj}]^\mathsf{T}$ the i^{th} row and j^{th} column vectors of A, respectively. Then we have:

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_R \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1K} \\ a_{21} & a_{22} & \ldots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{R1} & a_{R2} & \ldots & a_{RK} \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix} \tag{3.3}$$

$$= \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_R \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix} \tag{3.4}$$

$$= \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_K \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix} \tag{3.5}$$

It has been shown that any rate vector $\mathbf{r}$ which achieves universal recovery should satisfy the

[2] Only linear coding schemes are considered since it has been proved that they are sufficient to optimally solve the cooperative data exchange problem [7, 10].

following Slepian-Wolf constraints [3, 64]:

$$\sum_{i\in\{1,\ldots,N\}\setminus\mathscr{I}} r_i \geq \left|\mathbf{X}^c_{\mathscr{I}}\right|, \forall \mathscr{I} \subsetneq \{1,\ldots,N\} \tag{3.6}$$

Let $\Omega = \{\mathbf{r} = [r_1,\ldots,r_N]^\mathsf{T} : \sum_{i\in\mathscr{N}\setminus\mathscr{I}} r_i \geq \left|\mathbf{X}^c_{\mathscr{I}}\right|, \forall \mathscr{I} \subsetneq \mathscr{N}\}$ denote the set of all rate vectors $\mathbf{r}$ which satisfy (3.6). The minimum number of required transmissions for achieving universal recovery can be computed by solving the following integer linear program:

$$R^* = \min_{\mathbf{r}\in\Omega} \sum_{i=1}^{N} r_i. \tag{3.7}$$

Example 3.1. *Consider a cooperative data exchange problem for the fully connected network with $N = 4$ nodes and $K = 9$ packets. The packet distribution matrix is as follows:*

$$E = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{3.8}$$

The number of non-empty proper subsets of nodes is 14. *Thus, we can write down* 14 *linear constraints and solve the inequalities. For example, for $\mathscr{I} = \{1\}$, the constraint for total number of transmissions made by nodes $\{2,3,4\}$ is*

$$\sum_{i=\{2,3,4\}} r_i \geq |\mathbf{X}^c_1| = 3. \tag{3.9}$$

By using the methods proposed in [6, 7], based on minimizing submodular functions, the integer linear program can be solved in polynomial time and the minimum number of required transmissions should be 5. *After knowing the minimal number of transmissions, generating the coding scheme is a multicast network code construction problem and can be solved by polynomial-time algorithms proposed in [65]. One feasible coding scheme could be:*

- *Node 1 sends $T_1 = P_1 + P_5$ and $T_2 = P_2 + P_6$.*
- *Node 2 sends $T_3 = P_3 + P_7$.*
- *Node 3 sends $T_4 = P_4 + P_8$ and $T_5 = P_9$.*

In general, there are multiple different optimal coding schemes that achieve universal recovery. Although not all nodes have to make transmissions, the existing algorithms which solve the integer linear program (3.7) have to consider constraints introduced by all non-empty proper subset of nodes. In Example 3.1, the optimal coding scheme does not require node 4 to make any transmission, but the algorithms still have to consider the constraints related to node 4. However, we will show that without knowing the exact packet distribution information at some

nodes (in this example, node 4), but only knowing the number of initially available packets at them, it is still possible to compute the minimum number of required transmissions and construct the optimal coding scheme which achieves universal recovery with the minimum number of transmissions.

Definition 3.2 (Hamming Weight)**.** *For any vector v, define the Hamming Weight $w_H(v)$ as the number of non-zero entries of v.*

For subset of vectors $\mathbf{S} \subseteq \mathbf{V}$, let $v_{\mathbf{S}}$ denote the entry-wise **OR** result of all vectors in $\mathbf{S}$.

Definition 3.3 ((d,K)-Basis)**.** *Let $0 \le d \le K-1$. A set of K-dimensional binary linearly independent vectors ($\mathbf{V} = \{v_i : i \in \{1,\dots,K-d\}\}$) is called a (d,K)-Basis if*

$$w_H(v_{\mathbf{S}}) \ge |\mathbf{S}| + d, \qquad \forall \emptyset \ne \mathbf{S} \subseteq \mathbf{V}. \tag{3.10}$$

Definition 3.4 (Balanced (d,K)-Basis)**.** *A (d,K)-Basis ($\mathbf{V} = \{v_i : i \in \{1,\dots,K-d\}\}, 0 \le d \le K-1$) is called a balanced (d,K)-Basis if*

$$w_H(v_i) = d+1, \qquad \forall i \in \{1,\dots,K-d\}. \tag{3.11}$$

Condition (3.10) requires that $w_H(v_{\mathbf{S}})$, the number of dimensions spanned by vectors in $\mathbf{S}$, is no less than the number of vectors plus d. Hence, the number of vectors in each subspace of the K-dimensional space is limited.

Definition 3.5. *A binary vector (u) can generate another binary vector (v) if u and v have the same dimension and*

$$\{m : v_m = 1\} \subseteq \{n : u_n = 1\}. \tag{3.12}$$

Moreover, let $\mathcal{G}(u)$ denote the set of all binary vectors that can be generated by u. Define $\mathcal{G}(\mathbf{S}) = \cup_{u \in \mathbf{S}} \mathcal{G}(u)$ and $\mathcal{G}(u,d) = \{v : v \in \mathcal{G}(u), w_H(v) = d+1\}$.

Definition 3.6. *A set of K-dimensional binary vectors $\mathbf{U} = \{u_1,\dots,u_N\}$ is able to generate a (d,K)-Basis $\{v_i : i \in \{1,\dots,K-d\}\}$ if $\forall\, i \in \{1,\dots,K-d\}$, $v_i \in \mathcal{G}(\mathbf{U})$. Let $d^*(\mathbf{U})$ denote the maximum value of d such that a (d,K)-Basis can be generated by $\mathbf{U}$. For ease of notation, we use d^* instead of $d^*(\mathbf{U})$ when no ambiguity exists.*

Lemma 3.1. *If a set of K-dimensional binary vectors is able to generate a (d_1,K)-Basis, then it is also able to generate a (d_2,K)-Basis for any $d_2 \le d_1$.*

Proof. Consider a set of binary vectors $\{u_1,\dots,u_N\}$ that is able to generate a (d_1,K)-Basis $\mathbf{V} = \{v_1,\dots,v_{K-d_1}\}$. Then

$$\forall\, i \in \{1,\dots,K-d_1\}, \exists\, j \in \{1,\dots,N\} : \{m : v_{im}\} \subseteq \{n : u_{jn}\} \tag{3.13}$$

Hence any vector generated by v_i should also be able to be generated by the corresponding u_j. Thus, to prove this lemma, it suffices to show that $\forall\, d_2 \le d_1$, there exists a (d_2,K)-Basis

$\mathbf{Q} = \{q_1, \ldots, q_{K-d_2}\}$ that can be generated by $\{v_1, \ldots, v_{K-d_1}\}$. Since $\mathbf{V}$ is a (d_1, K)-Basis and $d_2 \leq d_1$, $\forall \mathbf{S} \subseteq \mathbf{V}$, we have

$$w_H(v_{\mathbf{S}}) \geq |\mathbf{S}| + d_1 \geq |\mathbf{S}| + d_2 \tag{3.14}$$

Thus all vectors in $\{v_1, \ldots, v_{K-d_1}\}$ satisfy the constraints for vectors of (d_2, K)-Basis. We can choose $q_i = v_i$, $\forall i = \{1, \ldots, K-d_1\}$. Moreover, $\forall j \in \{K-d_1+1, \ldots, K-d_2\}$, we choose $q_j = v_1$ to be the repeated vector. Then, $\forall \hat{\mathbf{S}} \subseteq \mathbf{Q}$:

$$w_H(q_{\hat{\mathbf{S}}}) \geq |\hat{\mathbf{S}}| + d_1 - c \geq |\hat{\mathbf{S}}| + d_2 \tag{3.15}$$

where $c = |\hat{\mathbf{S}} \cap \{q_j : j \in \{K-d_1+1, \ldots, K-d_2\}\}| \leq d_1 - d_2$ is the number of the repeated vectors. Hence $\mathbf{Q} = \{q_1, \ldots, q_{K-d_2}\}$ is a (d_2, K)-Basis. □

3.2 Cooperative Data Exchange and (d,K)-Basis

In this section, we present the relationship between a (d,K)-Basis and a coding scheme that can enable nodes with at least d packets to recover all missing packets, which is revealed by the following theorem.

Theorem 3.1. *If for some subset of nodes $\mathscr{I} \subseteq \mathscr{N}$ there exists a (d,K)-Basis $\mathbf{V} \subseteq \mathscr{G}(\{e_i, i \in \mathscr{I}\}, d)$, then the nodes of $\mathscr{I}$ can generate a coding scheme $\mathbf{T} = \{T_1, \ldots, T_R\}$ with $R = K - d$ such that $\forall i \in \mathscr{N}$, $w_H(e_i) \geq d$, node i can recover all packets.*

Proof. In our coding scheme, each transmission T_i is a linear combination (with appropriate coefficients) of the packets indexed by the non-zero entries in v_i. Since the vectors v_i's are a subset of the vectors generated by the PDVs of the nodes in $\mathscr{I}$, there is one node in $\mathscr{I}$ for each v_i that can locally produce and transmit said linear combination. The overall code can thus be characterized by a matrix A as in Equation (3.3) where in row i, only the elements indexed by v_i are non-zero.

For any $\mathscr{C} \subset \{1, \ldots, K\}$ with $|\mathscr{C}| = R$, let $A(\mathscr{C})$ denote the submatrix of A consisting of the R columns indexed by $\mathscr{C}$. Due to constraint (3.10), $\forall \emptyset \neq \mathbf{S} \subseteq \mathbf{V}$, we have

$$w_H(v_{\mathbf{S}}) \geq |\mathbf{S}| + d. \tag{3.16}$$

Denote the i^{th} of row of $A(\mathscr{C})$ by $\alpha_i(\mathscr{C})$. Then $\forall \hat{\mathbf{S}} \subseteq \{\alpha_1(\mathscr{C}), \cdots, \alpha_R(\mathscr{C})\}$, we have

$$w_H(\alpha_{\hat{\mathbf{S}}}(\mathscr{C})) \geq w_H(v_{\hat{\mathbf{S}}}) - d \geq |\hat{\mathbf{S}}|. \tag{3.17}$$

Let $\mathbb{G}(A(\mathscr{C}))$ denote the bipartite graph corresponding to $A(\mathscr{C})$, where there is an edge between i^{th} left vertex and j^{th} right vertex if and only if $A(\mathscr{C})_{ij} \neq 0$. Since Equation (3.17) satisfies the condition of Hall's marriage theorem[66], there exists a perfect matching in $\mathbb{G}(A(\mathscr{C}))$. According to Edmond's Theorem [67], the existence of perfect matching in bipartite graph

$\mathbb{G}(A(\mathscr{C}))$ implies that $\det(A(\mathscr{C})) \not\equiv 0$.

The product of determinants of all submatrices with R columns, denoted by $\prod_{\mathscr{C}} \det(A(\mathscr{C}))$, is a multivariate polynomial of non-zero entries of A. For a large enough finite field, there always exists a good choice of non-zero entries of A such that $\prod_{\mathscr{C}} \det(A(\mathscr{C})) \not\equiv 0$ [68]. For such choices, any R columns of A can be linearly independent at the same time. In other words, given any d packets, the other R missing packets can be recovered from our coding scheme. □

Remark 3.1. *In Theorem 3.1, we proved that if the PDVs of nodes are able to generate a (d,K)-Basis, they can also generate a coding scheme such that nodes with at least d packets can recover all missing packets from the coding scheme. The coefficient matrix used in the proposed coding scheme can be associated with a constrained generator matrix for an MDS code [69, 70]. We will introduce an efficient way to construct it in sufficiently large finite fields by performing elementary row operations on a Vandermonde matrix in Section 3.4.*

Theorem 3.1 characterizes a certain class of coding schemes. Their common feature is that each transmission is a (judiciously chosen) linear combination of exactly the same number of pure packets, namely, $d+1$. Initially, this last feature may appear to be too restrictive to attain optimal performance. However, in the sequel, we will establish in two steps that there always exists an optimal scheme with this special property. Nonetheless, let us recall that in general, the optimal data exchange scheme is not unique, so there may be alternative schemes attaining the same (optimal) number of transmissions while not satisfying the special property. To establish the existence of an optimal scheme with the special property, we will next establish that if a (linear) scheme enabling universal recovery exists, then the nodes are also able to generate a corresponding basis (and hence, by Theorem 3.1, a scheme with the special property must exist). More precisely, we have the following theorem:

Theorem 3.2. *If a subset of nodes is able to generate a linear coding scheme with R ($R = K-d$) transmissions which achieves universal recovery, then the PDVs of the nodes can generate a (d,K)-Basis $\mathbf{V} = \{v_1 \ldots, v_R\}$.*

Proof. We assume that a subset of nodes $\mathscr{I}$ can generate R linearly independent transmissions $\hat{\mathbf{T}} = \{\hat{T}_1, \ldots, \hat{T}_R\}$ which achieves universal recovery. The code can be characterized by a matrix $\hat{A}$ as in Equation (3.3) with rows $\hat{\alpha}_i$'s and columns $\hat{\beta}_j$'s. Let $\hat{\mathbf{V}} = \{\hat{v}_1, \ldots, \hat{v}_R\}$ where each $v_i = supp(\hat{\alpha}_i)$. That means in row i of $\hat{A}$, only the elements indexed by $\hat{v}_i$ are non-zero. We would like to show that if $\hat{\mathbf{V}} = \{\hat{v}_1, \ldots, \hat{v}_R\}$ does not satisfy Constraint (3.10) of the (d,K)-Basis, then the nodes which generate the corresponding transmissions are able to add more packets into the linear combinations until the Constraint (3.10) is satisfied.

For each non-empty subset $\mathbf{S} \subseteq \{\hat{\alpha}_1, \ldots, \hat{\alpha}_R\}$ such that $w_H(\hat{\alpha}_{\mathbf{S}}) < |\mathbf{S}| + d$, we have

$$K - w_H(\hat{\alpha}_{\mathbf{S}}) > K - |\mathbf{S}| - d \geq R - |\mathbf{S}| + 1 \tag{3.18}$$

For the row vectors in $\mathbf{S}$, at least $R-|\mathbf{S}|+1$ columns are all zeros. Hence, there must exist a subset of columns $\mathscr{C} \subset \{1,\ldots,K\}$ and corresponding subset of column vectors $\mathbf{C} \subseteq \{\hat{\beta}_1,\ldots,\hat{\beta}_K\}$ such that

$$|\mathscr{C}| = |\mathbf{C}| = R-|\mathbf{S}|+1 \tag{3.19}$$

$$R - w_H(\hat{\beta}_{\mathbf{C}}) \geq |\mathbf{S}| \Rightarrow w_H(\hat{\beta}_{\mathbf{C}}) \leq R-|\mathbf{S}| < |\mathbf{C}| \tag{3.20}$$

Let $\hat{A}(\mathscr{C})$ denote the submatrix which is composed of the columns indicated by the subset of column vectors $\mathscr{C}$. Then submatrix $\hat{A}(\mathscr{C})$ is rank deficient. Let $P_{\mathscr{C}} \doteq \{P_i : i \in \mathscr{C}\}$ denote the set of packets indexed by $\mathscr{C}$. If the set $\mathscr{N}'$ of nodes that generate transmissions $\{\hat{T}_i : \alpha_i \in \mathbf{S}\}$ cannot add any more packets into the linear combination for their transmissions, they have no more extra available packets in $P_{\mathscr{C}}$ and each transmission is a linear combination of all its sender node's available packets. This assumption leads to a contradiction that nodes in $\mathscr{N}'$ cannot recover all missing packets. Thus, nodes in $\mathscr{N}'$ must have more packets in $P_{\mathscr{C}}$ and can add them into the linear combination to generate new transmissions $\{T_i : \alpha_i \in \mathbf{S}\}$ such that $w_H(\alpha_{\mathbf{S}}) = |\mathbf{S}|+d$, where α_i denotes the coefficient vector of transmission T_i. By replacing $\{\hat{T}_i : \alpha_i \in \mathbf{S}\}$ with $\{T_i : \alpha_i \in \mathbf{S}\}$, we have a new coding scheme $\mathbf{T}$ such that the set of corresponding support vectors $\mathbf{V} = \{v_i,\ldots,v_R\}$ forms a (d,K)-Basis. For each transmission T_i and the corresponding $\hat{T}_i$, we have $\hat{v}_i \in \mathscr{G}(v_i)$. Given that $\hat{\mathbf{T}}$ can achieve universal recovery, $\mathbf{T}$ can also achieve universal recovery. □

Lemma 3.2. *If a subset of nodes can generate a linear coding scheme based on a (d,K)-Basis which enables nodes with at least d packets to recover all packets, they also can generate an equivalent linear coding scheme based on a balanced (d,K)-Basis.*

Proof. For any linear coding scheme $\mathbf{T} = \{T_1,\ldots,T_{K-d}\}$ based on a (d,K)-Basis $\mathbf{V} = \{v_1,\ldots,v_{K-d}\}$, let A denote the coefficient matrix of $\mathbf{T}$ and α_i denote the i^{th} row of A. For each T_i with $w_H(\alpha_i) > d+1$, we show that it can be reduced to a linear combination of $d+1$ packets. $\forall \tilde{\mathbf{S}} \subseteq \{\alpha_j : j \neq i\}$, $w_H(\alpha_{\tilde{\mathbf{S}}}) \geq |\tilde{\mathbf{S}}|+d$. The linear combination of $\{T_j : j \neq i\}$ can provide $K-d-1$ degrees of freedom among the used packets. Hence, by subtracting a proper linear combination of $\{T_j : j \neq i\}$ from T_i, we can get $\bar{T}_i$ with $w_H(\bar{\alpha}_i) = d+1$. Thus the corresponding $\bar{\mathbf{V}}$ is a balanced (d,K)-Basis. □

Example 3.2. *For CDE problem in Example 3.1, we already know a coding scheme with* 5 *transmissions that can achieve universal recovery. But each coded packet for transmission is a linear combination of two packets or just one pure packet. According to Theorem 3.2 and Lemma 3.2, there must exist another coding scheme in which every coded packet for transmission is a linear combination of* 5 *packets. It is easy to verify that the coding scheme with the following coefficient matrix (over finite field $GF(2^4)$ with primitive polynomial $\alpha^4+\alpha+1$) also achieves*

universal recovery.

$$A = \begin{bmatrix} 5 & 4 & 4 & 1 & 1 & 0 & 0 & 0 & 0 \\ 15 & 11 & 14 & 14 & 0 & 1 & 0 & 0 & 0 \\ 3 & 6 & 13 & 0 & 0 & 0 & 15 & 14 & 0 \\ 9 & 12 & 7 & 0 & 0 & 0 & 15 & 0 & 14 \\ 0 & 0 & 0 & 10 & 14 & 6 & 9 & 8 & 0 \end{bmatrix} \tag{3.21}$$

Each transmission is a linear combination of 5 *packets. Define binary matrix* V *such that*

$$V_{ij} = \begin{cases} 1, & A_{ij} \neq 0, \\ 0, & A_{ij} = 0. \end{cases} \tag{3.22}$$

Then we have

$$V = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}. \tag{3.23}$$

The row vectors of V *actually form a balanced* $(4,9)$*-Basis. As mentioned in Theorem 3.1, given any* 4 *packets, the other* 5 *packets can be recovered from the coding scheme based on the coefficient matrix A. Hence, in this example, the detailed information of available packets at node* 4 *is not necessary. As long as it initially has* 4 *packets, it can always recover the other packets by receiving these coded packets.*

Now we have the connection between the optimal coding schemes with the minimum number of required transmissions and the balanced (d,K)-Bases. Thus, we can search balanced (d,K)-Bases to get achievable (upper) bounds on the minimum number of required transmissions. Extending the search over all values of d (and using Theorem 3.2 and Lemma 3.2) then establishes optimal performance. More precisely, we have the following theorem:

Theorem 3.3. *For the cooperative data exchange in the fully connected network, the minimal number of required transmissions* R^* *satisfies:*

$$R^* = K - \min\{\mathscr{M}, d^*\} \tag{3.24}$$

where the (d^*,K)*-Basis is the largest* (d,K)*-Basis that can be generated by the PDVs of nodes and* $\mathscr{M} = \min_{i\in\mathscr{N}} |\mathbf{X}_i|$ *is the minimum number of initially available packets at any single node.*

Proof. By assumption, d^* is the largest value of d for which a (d,K)-Basis can be generated by the PDVs. But then, by Theorem 3.2, there does not exist any linear coding scheme that can achieve universal recovery by using fewer than $K-d^*$ transmissions.

Suppose that $\mathscr{M} \geq d^*$. Then every node has at least d^* packets. Since a (d^*,K)-Basis can be generated by the PDVs, according to Theorem 3.1, there is a linear coding scheme with $K-d^*$

transmissions such that every node with at least d^* packets can recover all missing packets.

Now suppose that $\mathscr{M} < d^*$. According to Lemma 3.1, the PDVs can also generate a (d,K)-Basis with $d = \mathscr{M}$. According to Theorem 3.1, there is a linear coding scheme with $K - \mathscr{M}$ transmissions such that every node with at least $\mathscr{M}$ packets can recover all missing packets.

Hence, the minimum number of required transmissions satisfies $R^* = K - \min\{\mathscr{M}, d^*\}$. □

3.3 Algorithms

According to Theorem 3.3, to solve the cooperative data exchange problem for the fully connected network, we need to find the largest value of d such that a balanced (d,K)-Basis that can be generated by the PDVs of nodes. We denote this optimal value of d by d^*. This problem can be decomposed into two subproblems:

(1) Given a fixed d, determine whether any balanced (d,K)-Basis can be generated by the PDVs of nodes or not.

(2) Find the maximum value of d such that the PDVs of nodes can generate one balanced (d,K)-Basis.

In this section, we propose two algorithms to solve the two subproblems, respectively.

3.3.1 Existence of (d,K)-Basis

Given the packet distribution matrix E and a specific parameter d, we propose Algorithm 1 to check whether any balanced (d,K)-Basis can be generated by the PDVs of nodes or not. Due to constraint (3.10), only nodes with at least $d+1$ packets can generate the (d,K)-Basis vectors. Hence, it is sufficient that we only consider the PDVs with $w_H(e_i) > d$ as the candidates to generate basis vectors.

Definition 3.7. *For any binary vector u with $w_H(u) > d$, let $\mathscr{J}(u) = \{j_1, \ldots, j_{w_H(u)}\}$ denote the set of indices of the non-zero entries of u. Define the set $\mathscr{B}(u,d) = \{b_i : supp(b_i) = \{j_1, \ldots, j_d\} \cup \{j_{d+i}\}, \forall\, i \in 1, \ldots, w_H(u) - d\}$.*

The set $\mathscr{B}(u,d)$ is a particular set of binary vectors that are generated by u. Specifically, each of the vectors in the set has weight $d+1$ and it can be verified that the vectors satisfy the Constraint (3.10) in the definition of the (d,K)-Basis. Therefore, they are basis vector candidates for the balanced (d,K)-Basis.

Example 3.3. *Given $e_1 = [1,1,1,1,1,1,0,0,0]$, we can assign the $\mathscr{B}(e_1,4) = \{b_1, b_2\}$ with*

$$b_1 = [1,1,1,1,1,0,0,0,0], \tag{3.25}$$

$$b_2 = [1,1,1,1,0,1,0,0,0]. \tag{3.26}$$

Lemma 3.3. *For any binary vector $v \in \mathscr{G}(u,d) \setminus \mathscr{B}(u,d)$, let $\mathbf{S} = \mathscr{B}(u,d) \cup \{v\}$. We have $w_H(b_{\mathbf{S}}) < |\mathbf{S}| + d$.*

Proof. Since $\mathscr{B}(u,d) \subset \mathscr{G}(u,d)$, $w_H(b_{\mathscr{B}(u,d)}) \leq w_H(u)$. Also, according to the definition of $\mathscr{B}(u,d)$, we have

$$w_H(b_{\mathscr{B}(u,d)}) \geq |\mathscr{B}(u,d)| + d = w_H(u). \tag{3.27}$$

Hence, $w_H(b_{\mathscr{B}(u,d)}) = w_H(u)$. For any $v \in \mathscr{G}(u,d) \setminus \mathscr{B}(u,d)$, $|\mathbf{S}| = |\mathscr{B}| + 1 = w_H(u) - d + 1$, thus

$$w_H(b_{\mathbf{S}}) = w_H(u) < w_H(u) - d + 1 + d = |\mathbf{S}| + d. \tag{3.28}$$

□

Thus, any vector $v \in \mathscr{G}(u,d) \setminus \mathscr{B}(u,d)$ is not compatible with $\mathscr{B}(u,d)$ in terms of the Constraint (3.10).

Corollary 3.1. *For each PDV e_i, it is sufficient to check vectors of $\mathscr{B}(e_i,d)$ instead of all vectors of $\mathscr{G}(e_i,d)$.*

Proof. Suppose that all vectors in $\mathscr{B}(e_i,d)$ are selected to be the (d,K)-Basis vectors, then according to Lemma 3.3, the vectors in $\mathscr{G}(e_i,d) \setminus \mathscr{B}(e_i,d)$ are not compatible with $\mathscr{B}(e_i,d)$.

If not all vectors in $\mathscr{B}(e_i,d)$ are selected to be the (d,K)-Basis vectors, for the PDV e_j which is in $\mathscr{B}(e_i,d)$ but not selected as one of the (d,K)-Basis vectors, there must exist a binary vector (denoted by q) in $\mathbf{Q}$ that can generate e_j according to the 6-th line of Algorithm 1. As q is the bitwise OR result of a subset of the basis vectors generated by previous PDVs, which satisfy Inequality (3.34), the number of (d,K)-Basis vectors that can be generated by q achieves the maximum. Hence, any other vectors which can be generated by q are not compatible with those already selected basis vectors. Furthermore, since e_j can be generated by q and e_j has d ones at the common positions with other vectors in $\mathscr{B}(e_i,d)$, q must have d ones at the common positions with other vectors in $\mathscr{B}(e_i,d)$. Therefore, the other vectors which are in $\mathscr{B}(e_i,d)$ and have been selected to be the (d,K)-Basis vectors should be merged with q. Let us denote the merged vector by q'. And vectors in $\mathscr{G}(e_i,d) \setminus \mathscr{B}(e_i,d)$ can all be generated by q and are not compatible with the basis vectors already selected before.

Hence, it is sufficient to check vectors in $\mathscr{B}(e_i,d)$ and ignore vectors in $\mathscr{G}(e_i,d) \setminus \mathscr{B}(e_i,d)$. □

Although for each PDV e_i, there are as many as $\binom{w_H(e_i)}{d+1}$ balanced (d,K)-Basis vectors that can be generated, we can select any $\mathscr{B}(e_i,d)$ and only consider them as the candidate basis vectors. Any other $v \in \mathscr{G}(e_i,d) \setminus \mathscr{B}(e_i,d)$ can be ignored.

Lemma 3.4. *Let $\mathbf{S} = \{v_1, \ldots, v_{|\mathbf{S}|}\}$ denote a set of binary vectors with weight $w_H(v_i) = d+1, \forall\, v_i \in \mathbf{S}$ and $v_{\mathbf{S}}$ denote the bitwise **OR** result of all vectors in $\mathbf{S}$. For any vector $v \in \mathscr{G}(v_{\mathbf{S}},d) \setminus \mathbf{S}$, let*

$\hat{\mathbf{S}} = \mathbf{S} \cup \{v\}$, *we have* $w_H(v_{\hat{\mathbf{S}}}) < |\hat{\mathbf{S}}| + d$ *if*

$$w_H(v_{\mathbf{S}}) \leq \sum_{i \in \mathbf{S}} w_H(v_i) - (|\mathbf{S}| - 1)d. \tag{3.29}$$

Proof. Since v and all vectors in $\mathbf{S}$ can be generated by $v_{\mathbf{S}}$, we have

$$\begin{aligned} w_H(v_{\hat{\mathbf{S}}}) = w_H(v_{\mathbf{S}}) &\leq \sum_{i \in \mathbf{S}} w_H(v_i) - (|\mathbf{S}| - 1)d && (3.30)\\ &= \sum_{i \in \mathbf{S}} (w_H(v_i) - d) + d && (3.31)\\ &= |\mathbf{S}| + d && (3.32)\\ &< |\hat{\mathbf{S}}| + d. && (3.33) \end{aligned}$$

□

Thus, any vector $v \in \mathscr{G}(v_{\mathbf{S}}, d) \setminus \mathbf{S}$ is not compatible with $\mathbf{S}$ in terms of the Constraint (3.10) if Inequality (3.29) holds. Hence, once we find any set of basis vectors which satisfy Inequality (3.29), any vector that can be generated by the merged vector should not be considered.

Remark 3.2. *Binary vector v_m which has weight larger than $d+1$ can be treated as a merged vector of $\mathscr{B}(v_m, d)$. Therefore, $w_H(v_m) - d = |\mathscr{B}(v_m, d)|$. Inequality* (3.29) *also works for the cases where some of the vectors have weights larger than $d+1$.*

We use set $\mathbf{V}$ to store the balanced (d,K)-Basis vectors that have been generated by checked PDVs. We use set $\mathbf{Q}$ to store merged (d,K)-Basis vectors. Any set of vectors which satisfy Inequality (3.29) will be merged as one vector and stored in $\mathbf{Q}$. Only $b \in \mathscr{B}(e_i, d)$ that cannot be generated by any vector in $\mathbf{Q}$ can be selected as the basis vectors. After all vectors in $\mathscr{B}(e_i, d)$ have been checked, there must exist e_i or a vector that can generate e_i in $\mathbf{Q}$.

In the subspace spanned by any two vectors in $\mathbf{Q}$, there must exist at least one vector that should be added to form the (d,K)-Basis. Instead of checking every subset of $\mathbf{Q}$ for merging, it is sufficient to only check the newly added vector with any subset $\mathbf{S} \subseteq \mathbf{Q}$ with $|\mathbf{S}| \leq 2$ and treat the merged vector as the newly added vector for further merging until no merging possibility.

In the end, if $K-d$ such vectors are found, the PDVs of nodes are able to generate a (d,K)-Basis which is stored by $\mathbf{V}$ and the algorithm returns **True** and the corresponding basis $\mathbf{V}$. Otherwise, return **False**.

Theorem 3.4 (Correctness of Algorithm 1)**.** *Algorithm 1 can output the valid (d,K)-Basis if there exists.*

Proof. We first prove that a set of vectors, $\mathbf{V} = \{v_1, \ldots, v_R\}$ with $R = K-d$, output by Algorithm 1 is always a balanced (d,K)-Basis. Since all vectors $v \in \mathbf{V}$ are binary vectors belonging to

Algorithm 1 Search balanced (d,K)-Basis (SdB)

Input: $E=[e_1,\ldots,e_N]^{\mathsf{T}}$ and d.
Output: True, $\mathbf{r}$, $\mathbf{V}$ or **False**.
Initialization: $\mathbf{Q}=\emptyset$, $\mathbf{V}=\emptyset$, $\mathbf{r}=[r_1,\ldots,r_N]^{\mathsf{T}}=\mathbf{0}_{1\times N}$.
for $i: i\in\{1,\ldots,N\}$ **do**
  for $b\in\mathscr{B}(e_i,d)$ **do**
    if $b\notin\mathscr{G}(\mathbf{Q},d)$ **then**
      $r_i=r_i+1$
      $\mathbf{V}=\mathbf{V}\cup\{b\}$
      while $\exists\mathbf{S}\subseteq\mathbf{Q},|\mathbf{S}|\leq 2:$ (3.34) *holds* **do**

$$w_H(q_{\mathbf{S}}\vee b)\leq\sum_{q_i\in\mathbf{S}}w_H(q_i)+w_H(b)-|\mathbf{S}|d \tag{3.34}$$

        $b=b\vee q_{\mathbf{S}}$, $\mathbf{Q}=\mathbf{Q}\setminus\mathbf{S}$
      end while
      $\mathbf{Q}=\mathbf{Q}\cup\{b\}$
    end if
    if $|\mathbf{V}|=K-d$ **then**
      return **True**, $\mathbf{r}$ and $\mathbf{V}$
    end if
  end for
end for
return **False**

$\mathscr{B}(e_i,d)$, each vector has exactly $d+1$ ones. And each newly added vector is compatible with all previously selected vectors in terms of the condition (3.10) according to Lemma 3.4. Thus, $\mathbf{V}=\{v_1,\ldots,v_R\}$ is a valid (d,K)-Basis.

Secondly, we prove that Algorithm 1 is always able to find one (d,K)-Basis if there exists some (d,K)-Basis which can be generated by PDVs of nodes. Since every valid balanced (d,K)-Basis is a subset of binary vectors with $d+1$ ones which can be generated by all PDVs of nodes. According to Corollary 3.1, it is sufficient to only check $\mathscr{B}(e_i,d)$ for all $i\in\{1,\ldots,N\}$. Hence, Algorithm 1 searches (d,K)-Basis from all possible candidates and can output a (d,K)-Basis if there exists one. □

3.3.2 Searching for d^*

We propose Algorithm 2 which uses binary search method to find the (d^*,K)-Basis that can be generated by PDVs of nodes. Let e^* be the PDV of the node which has the largest number of available packets initially, i.e.,

$$e^*=\arg\max_{e_i}w_H(e_i). \tag{3.35}$$

According to Theorem 3.3, if the PDVs of nodes can generate any (d,K)-Basis such that $d \geq \mathscr{M}$, we do not have to check for any larger d. Also, the (d,K)-Basis with the largest d that can be generated should always be no larger than $w_H(e^*)-1$. Therefore, we start from $d_{max} = \min\{\mathscr{M}, w_H(e^*)-1\}$ instead of K.

Algorithm 2 Minimal Number of Required Transmissions and d-Basis

Input: $E_{N\times K} = [e_1,\ldots,e_N]^\mathsf{T}$.
Output: $R^*, \mathbf{V}^*$
Initialization: $d_{min} = 1$, $d_{max} = \min\{\mathscr{M}, w_H(e^*)-1\}$.
$(F,\mathbf{r},\mathbf{V}) = SdB(E, d_{max})$
if F is **True then**
    $d^* = d_{max}, \mathbf{V}^* = \mathbf{V}$
else
    $(F,\mathbf{r},\mathbf{V}) = SdB(E, d_{min})$
    if $\neg F$ **then**
        $d^* = 0, \mathbf{V}^* = I_K$
    else
        while $d_{max} - d_{min} > 1$ **do**
            $d = \lfloor \frac{d_{min}+d_{max}}{2} \rfloor$
            $(F,\mathbf{r},\mathbf{V}) = SdB(E, d)$
            if F **then**
                $d_{min} = d, \mathbf{V}^* = \mathbf{V}$
            else
                $d_{max} = d$
            end if
        end while
        $d^* = d_{min}$
    end if
end if
$R^* = K - d*$

3.3.3 Complexity

In Algorithm 2, the binary search method is used to find the (d^*,K)-Basis that can be generated by the PDVs of nodes. The complexity of the algorithm is bounded by $\log(K)$. For each specific d, Algorithm 1 is used to search the existence of (d,K)-Basis. Let $M(d)$ denote the number of nodes that have at least $d+1$ packets. The first *For* loop has at most $M(d)$ iterations. For the i^{th} candidate PDV e_i, the size of set $\mathscr{B}(e_i,d)$ satisfies $|\mathscr{B}(e_i,d)| = w_H(e_i)-d$. Hence, the second *For* loop has at most $w_H(e_i)-d$ iterations. The number of subsets of vectors in $\mathbf{Q}$ with size 1 and 2 are $|\mathbf{Q}|$ and $\binom{|\mathbf{Q}|}{2}$, respectively. For the i^{th} checked node, $|\mathbf{Q}| \leq i$, because basis vectors generated by the same PDV can always be merged to one vector and basis vectors generated by different PDVs may still be merged. The number of possible merging iteration for each candidate basis vector is less than the size of (d,K)-Basis vector which is $K-d$. Then, the *While* loop has at most $(i+\binom{i}{2})(K-d)$ iterations for the i^{th} PDV. Hence the

complexity[3] of Algorithm 1 is bounded by $\sum_{i=1}^{M(d)}(i+\binom{i}{2})(w_H(e_i)-d)(K-d)K$. Since $M(d) \leq N$ and $w_H(e_i) \leq K$, we have the overall complexity is bounded by $\mathcal{O}(N^3K^3\log(K))$, which is much lower than the complexity of existing algorithms proposed in [6] based on minimizing a submodular function $\mathcal{O}((N^6K^3+N^7)\log(K))$ and algorithm based on subgradient methods $\mathcal{O}((N^4\log(N)+N^4K^3)K^2\log(K))$.

Example 3.4. *Apply our algorithms on Example 3.1. Node 4 and node 1 initially have the smallest and the largest number of packets respectively, which means* $\mathscr{M}=4$ *and* $w_H(e^*)=6$. *Therefore we have* $d_{max}=4$. *Algorithm 1 will first check whether it is possible to generate any* $(4,9)$*-Basis from* $\{e_1,e_2,e_3\}$ *by* $SdB(E,4)$. *The PDV of the* 4^{th} *node,* e_4*, will not be considered as the candidate since* $w_H(e_4)=4$ *and it can not generate any binary vector with* 5 *ones. In this example,* $SdB(E,4)$ *returns* ***True****. The minimum number of required transmissions is* 5. *For the general case, if a* (d_{max},K)*-Basis cannot be generated, a binary search methods would be used to find* d^*.

Now we investigate the detail of algorithm $Sdb(E,4)$. *The first* ***For*** *loop only runs for* $\{e_1,e_2,e_3\}$.

- *For* e_1, $\mathscr{B}(e_1,4)=\{b_{11},b_{12}\}$ *where* $b_{11}=[1,1,1,1,1,0,0,0,0]$ *and* $b_{12}=[1,1,1,1,0,1,0,0,0]$.
- *For* e_2, $\mathscr{B}(e_2,4)=\{b_{21},b_{22}\}$ *where* $b_{21}=[1,1,1,0,0,0,1,1,0]$ *and* $b_{22}=[1,1,1,0,0,0,1,0,1]$.
- *For* e_3, $\mathscr{B}(e_3,4)=\{b_{31},b_{32}\}$ *where* $b_{31}=[0,0,0,1,1,1,1,1,0]$ *and* $b_{32}=[0,0,0,1,1,1,1,0,1]$.

The second ***For*** *loop runs for each* $b_{ij}\in\mathscr{B}(e_i,4)$ *for all* $i\in\{1,2,3\}$.

- *For* b_{11}*, since currently* $\mathbf{Q}=\emptyset$, b_{11} *will be added into* $\mathbf{V}$ *as* v_1 *and* $\mathbf{Q}$ *as* q_1 *directly.*
- *For* b_{12}*, since it cannot be generated by* q_1, b_{12} *will be added into* $\mathbf{V}$ *as* v_2. *Now,* $\mathbf{Q}$ *is not empty anymore and has* q_1. *We have to check whether* b_{12} *should be merged with* q_1 *as one vector or not. Since* $w_H(b_{12}\vee q_1)\leq w_H(b_{12})+w_H(q_1)-d$ *satisfies Inequality* (3.34). *We should merge them and update as* $q_1=[1,1,1,1,1,1,0,0,0]$ [4].
- *For* b_{21}*, since it cannot be generated by* q_1, b_{21} *will be added into* $\mathbf{V}$ *as* v_3. *The merging possibility between* b_{21} *and* q_1 *will be checked and it turns out that they should not be merged. Hence* b_{21} *will be added into* $\mathbf{Q}$ *as* q_2.
- *For* b_{22}*, since it cannot be generated by* q_1 *or* q_2, b_{22} *will be added into* $\mathbf{V}$ *as* v_4. *It can be verified that* b_{22} *should be merged with* q_2 *but not with* q_1. *Hence* q_2 *is updated as* $q_2=[1,1,1,0,0,0,1,1,1]$.

[3] Computing bitwise **AND** or **OR** of two K-dimensional binary vector has complexity of K basic operations. In step 6 of Algorithm 1, we compute bitwise **OR** between b and each vector in $\mathbf{Q}$ and this results are also used in merging checking. Hence complexity of step 6 is not considered.

[4] In fact, b_{12} and q_1 can be merged without checking Condition (3.34), since $q_1=b_{11}$ and b_{12} are generated by the same PDV, e_1.

- *For b_{31}, since it cannot be generated by q_1 or q_2, b_{31} will be added into* $\mathbf{V}$ *as* v_5. *Now we have enough* $(4,9)$*-Basis vectors. The algorithm* $SdB(E,4)$ *will return* ***True*** *and corresponding* $\mathbf{V}$ *shown as Equation (3.23).*

Actually, if we check the merging possibility between b_{31} and $\{q_1, q_2\}$, we will find that b_{31} should not be merged with q_1 or q_2 individually, but should be merged with them together. And when we have the complete (d,K)-Basis, we can always merge all vectors in $\mathbf{Q}$ *into one vector with K ones. Although b_{32} is in $\mathscr{B}(e_3,4)$, it is not used, because we have found enough basis vectors before its iteration.*

3.4 Code Construction

In previous sections, we presented one algorithm to compute the minimum number of required transmissions. To completely solve the cooperative data exchange problem, we still need to construct the coding scheme which achieves universal recovery by using the minimum number of transmissions. In this section, we briefly show how to deterministically construct the optimal coding scheme in sufficiently large finite fields. For exponentially large enough finite field, it is well known that it is possible to deterministically set the coefficients. It has been shown in [71] that the $[n,k]$ generalized Reed-Solomon codes with sparsest and balanced generator matrices exist over finite field $q \geq n + \lceil \frac{k(k-1)}{n} \rceil$. For finite fields with small size, it is not known how to deterministically set the coefficients and a randomized method was presented in [18].

After knowing the number of transmissions which should be made by each node, designing the coding scheme can be formulated as a multicast network code construction problem. Methods based on the mixed matrix completion algorithm [72] and the Jaggi *et al.* algorithm [65] are presented in [6]. However, those methods have to take all packet distribution information into consideration and generate a coding scheme that may only work for this particular setting. As Theorem 3.1 points out, it is possible to construct a coding scheme that enables universal recovery at all nodes with at least $K - R^*$ packets. Packet distribution information of nodes that do not send anything is not necessary for constructing the code and can be ignored. This class of codes is based on MDS codes. It is well-known that the existence of such MDS codes with constrained generator matrices in finite fields with small size is an open problem and a corresponding conjecture was proposed in [70]. A randomized method was presented in [18] for finite fields with small size. If the size of the finite field is allowed to be (exponentially) large enough, the coefficient matrix of the coding scheme for cooperative data exchange problem can be efficiently constructed by starting from Vandermonde matrices.

Consider an $R \times K$ Vandermonde matrix over a finite field $\mathbb{F}_q$, where $R = K - d$:

$$\mathcal{V} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ \theta_1 & \theta_2 & \theta_3 & \dots & \theta_{K-1} & \theta_K \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \theta_1^{R-1} & \theta_2^{R-1} & \theta_3^{R-1} & \dots & \theta_{K-1}^{R-1} & \theta_K^{R-1} \end{bmatrix} \tag{3.36}$$

For large enough q, there exist $\{\theta_1, \dots, \theta_K\}$ (e.g., $\{0, \dots, K-1\}$) such that any m ($m \leq R$) columns of $\mathcal{V}$ are linearly independent. Hence, $\mathcal{V}$ is the generator matrix of an MDS code. However, the coefficient matrix A cannot simply be set equal to $\mathcal{V}$, since the number of non-zero entries of each row cannot be larger than the number of available packets at the node which generates this transmission. Nevertheless, by performing elementary transformations on $\mathcal{V}$, we can transform it into a coefficient matrix A with the property that each row has $K - R + 1$ non-zero entries.

Lemma 3.5. *For any $R \times K$ ($R \leq K$) Vandermonde matrix $\mathcal{V}$, by performing elementary row operations on $\mathcal{V}$, it is possible to get a matrix A with row vectors $\{\alpha_1, \dots, \alpha_R\}$ such that*

$$w_H(\alpha_i) = K - R + 1 \qquad \forall i \in \{1, \dots, K-R\} \tag{3.37}$$

$$w_H(\alpha_{\mathbf{S}}) \geq |\mathbf{S}| + K - R \qquad \emptyset \neq \mathbf{S} \subseteq \{\alpha_1, \dots, \alpha_R\} \tag{3.38}$$

Proof. Suppose we have a $R \times K$ Vandermonde matrix depicted as Equation (3.36). We use $\mathcal{V}_l$ and $\mathcal{V}_r$ to denote the first R columns submatrix and the last $K - R$ columns submatrix of $\mathcal{V}$, respectively. Then, $\mathcal{V} = \begin{bmatrix} \mathcal{V}_l & \mathcal{V}_r \end{bmatrix}$. Since any R columns of $\mathcal{V}$ are linearly independent, $\mathcal{V}_l$ is always a full rank matrix and invertible. Performing elementary row operations on $\mathcal{V}$ is equivalent to left multiplying a $R \times R$ matrix to $\mathcal{V}$. Let D denote a $R \times R$ matrix and $D = \mathcal{V}_l^{-1}$.

$$D\mathcal{V} = D \begin{bmatrix} \mathcal{V}_l & \mathcal{V}_r \end{bmatrix} = \begin{bmatrix} I_R & D\mathcal{V}_r \end{bmatrix}, \tag{3.39}$$

where I_R is the $R \times R$ identity matrix. Since D is invertible and is a full rank matrix, we have

$$\text{rank}(D\mathcal{V}_r) = \text{rank}(\mathcal{V}_r) = \min\{R, K - R\}. \tag{3.40}$$

If $R \geq \frac{K}{2}$, equivalently $R \geq K - R$, we have $\text{rank}(D\mathcal{V}_r) = K - R$. Then, $D\mathcal{V}_r$ is a column full rank matrix and each row can have $K - R$ non-zero entries. If $R < \frac{K}{2}$, equivalently $R < K - R$, we have $\text{rank}(D\mathcal{V}_r) = R$. Then $D\mathcal{V}_r$ is a row full rank matrix. However, since any R columns of $\mathcal{V}_r$ are linearly independent, we can further divide $\mathcal{V}_r$ into submatrices $\mathcal{V}_{r_1}, \mathcal{V}_{r_2}, \dots, \mathcal{V}_{r_t}$ such that the number of columns of each submatrix is no more than R. Thus, $D\mathcal{V}_r$ can be expressed as follows.

$$D\mathcal{V}_r = \begin{bmatrix} D\mathcal{V}_{r_1} & D\mathcal{V}_{r_2} & \dots & D\mathcal{V}_{r_t} \end{bmatrix} \tag{3.41}$$

For $i \in \{1, \dots, t\}$, each submatrix $D\mathcal{V}_{r_i}$ is a column full rank matrix and total number of non-zero

entries in each row of $D\mathcal{V}_r$ can be $K-R$. Let $A = D\mathcal{V}$, then we have row vectors of A that satisfy

$$w_H(\alpha_i) = 1 + K - R \qquad \forall i \in \{1, \ldots, K-R\} \tag{3.42}$$

$$w_H(\alpha_{\mathbf{S}}) \geq |\mathbf{S}| + K - R \qquad \emptyset \neq \mathbf{S} \subseteq \{\alpha_1, \ldots, \alpha_R\} \tag{3.43}$$

□

The matrix $D\mathcal{V}$ with $D = \mathcal{V}_l^{-1}$ satisfies both conditions of the balanced (d,K)-Basis with $d = K-R$. Hence it can be a coefficient matrix for the coding scheme based on the (d,K)-Basis. Normally, the places of non-zero entries of the (d,K)-Basis generated by the PDVs of nodes are different from matrix $D\mathcal{V}$. However, since any row vector with $d+1$ ones is in the space spanned by row vectors of $D\mathcal{V}$, further elementary row operations can be performed on $D\mathcal{V}$ to get the coefficient matrix with non-zero entries at the same places as the (d,K)-Basis generated by the PDVs of nodes.

Example 3.5. *Now we show how to use a Vandermonde matrix to construct the linear coding scheme for Example 3.1. We know that $R^* = 5$ and there exists a $(4,9)$-Basis V. Consider the Vandermonde matrix $\mathcal{V}$ over the finite field $GF(2^4)$ with primitive polynomial $\alpha^4 + \alpha + 1$.*

$$\mathcal{V} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 4 & 5 & 3 & 2 & 7 & 6 & 12 & 13 \\ 1 & 8 & 15 & 12 & 10 & 1 & 1 & 10 & 15 \\ 1 & 3 & 2 & 5 & 4 & 6 & 7 & 15 & 14 \end{bmatrix} \tag{3.44}$$

By elementary row transformations and Gaussian eliminations, we can get the coefficient matrix A shown as (3.21)*. Given any four packets, the other packets can be recovered from transmissions based on A. Suppose there is another node with PDV $e_5 = [1,0,1,0,1,0,0,1,0]$. It can also recover all its missing packets by receiving transmissions based on A. The detail of its packet distribution information is not used for either computing the minimal number of required transmissions or designing the coding scheme. Although in this example the coefficient matrix of our method looks more complicated than that of methods based on Jaggi et al.'s algorithm, in general cases, the complexity of constructing the coefficient matrix via our method is much lower.*

3.5 Cooperative Data Exchange with Weight Cost

In the basic cooperative data exchange problem, every transmission incurs the same cost, irrespective of the transmitting node. However, in more general cases, it is intuitive to consider that the (transmit) costs for different nodes are different. Let $\mathbf{w} = [w_1, \ldots, w_N]^\mathsf{T}$ denote the weight vector where each w_i is the cost for node i to make one transmission. For any coding

scheme with rate vector $\mathbf{r} = [r_1, \ldots, r_N]^\mathsf{T}$, the weighted cost is defined by

$$\mathscr{C}(\mathbf{r}) = \mathbf{w}^\mathsf{T} \cdot \mathbf{r} = \sum_{i=1}^{N} w_i r_i. \tag{3.45}$$

Instead of minimizing the total number of transmissions (sum rate), the goal of the cooperative data exchange problem with weighted cost is to achieve universal recovery by a coding scheme with a rate vector which has the minimum weighted cost. We note that once the optimal rate vector is found, a corresponding optimal transmission scheme can be developed exactly along the lines of the unweighted case discussed in previous sections.

The minimum weighted cost for the cooperative data exchange problem with weighted cost can be computed as

$$\mathscr{C}^* = \min_{\mathbf{r}\in\Omega} \mathscr{C}(\mathbf{r}) = \min_{\mathbf{r}\in\Omega} \sum_{i=1}^{N} w_i r_i. \tag{3.46}$$

Although the optimization should be taken over all vectors in Ω, we can actually decompose this optimization problem into two sub-optimization problems.

- We first find the optimal rate vector under the condition that the sum rate is fixed.
- Then, further optimization should only be over the optimal rate vectors for different fixed sum rates.

Definition 3.8. *Let $\mathscr{K}(R)$ denote the minimum weighted cost of all rate vectors that can achieve universal recovery and has sum-rate equal to R.*

$$\mathscr{K}(R) = \min_{\mathbf{r}\in\Omega, \mathscr{S}(\mathbf{r})=R} \mathscr{C}(\mathbf{r}) = \min_{\mathbf{r}\in\Omega, \mathscr{S}(\mathbf{r})=R} \sum_{i=1}^{N} w_i r_i, \tag{3.47}$$

where $\mathscr{S}(\mathbf{r}) = \sum_{i=1}^{N} r_i$ is the sum-rate.

Let R_{min} denote the minimum sum rate such that a corresponding rate vector can achieve universal recovery[5]. Only rate vectors with sum rate between R_{min} and K should be considered. The minimum weighted cost can also be computed as

$$\begin{aligned} \mathscr{C}^* &= \min_{R\in\{R_{min},\ldots,K\}} \mathscr{K}(R) \\ &= \min_{R\in\{R_{min},\ldots,K\}} \min_{\mathbf{r}\in\Omega, \mathscr{S}(\mathbf{r})=R} \sum_{i=1}^{N} w_i r_i. \end{aligned} \tag{3.48}$$

Example 3.6. *Consider a cooperative data exchange problem for the fully connected network with 5 nodes and 9 packets with the goal of minimizing the weighted cost of transmissions. The*

[5] In previous sections, for the basic cooperative data exchange problem, we use R^* to denote the minimum sum rate such that universal recovery can be achieved. However, in cooperative data exchange problems with weighted cost, the optimal rate vector may not have minimum sum rate.

packet distribution matrix (PDM) is as follows:

$$E = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

The weights of nodes are as follows:

Node(i)	*1*	*2*	*3*	*4*	*5*
w_i	*2*	*3*	*6*	*8*	*10*

Here we assume that the weights of the nodes are in non-decreasing order. If this is not the case, we rearrange the nodes of the PDM such that the weights are in non-decreasing order. By using the methods proposed in [6, 13], we can find that the optimal rate vector is $\mathbf{r}^* = [3,3,1,0,0]^\mathsf{T}$ *and the minimum weighted cost is* 21. *However, for the basic cooperative data problem (unweighted case) with the same packet distribution matrix, the optimal rate vector is* $\mathbf{r} = [1,1,1,1,1]^\mathsf{T}$.

Remark 3.3. *By using algorithms in [6, 7, 73], we can show that the minimum sum rate* R_{min} *for Example 3.6 is* 5. *But for the cooperative data exchange problem with weighted cost, the optimal rate vector has a sum rate of* 7, *which is larger than the minimum required sum rate. That is why we need to have the second optimization over* $\mathscr{K}(R)$, *where* $R \in \{R_{min}, \ldots, K\}$. *Thus, only finding rate vectors with sum rate* R_{min} *is not enough, we have to optimize* $\mathscr{K}(R)$ *over all* $R \in \{R_{min}, \ldots, K\}$. *However, we show that it is not necessary to compute* $\mathscr{K}(R)$ *for all* $R \in \{R_{min}, \ldots, K\}$. *By exploiting the convexity of the function* $\mathscr{K}(R)$, *we can search the optimal R and rate vector by the binary search method.*

We propose an efficient deterministic algorithm based on (d,K)-Basis to solve the optimization problem (3.47). For any given fixed number of transmissions R, Algorithm 3 searches the existence of a corresponding (d,K)-Basis where $d = K - R$.

Theorem 3.5. *For any* $R \in \{R_{min}, \ldots, K\}$ *and* $d = K - R$, *let* $\mathbf{r} = [r_1, \ldots, r_N]^\mathsf{T}$ *denote the output rate vector of Algorithm 3 with input PDM E and d, then* $\mathscr{K}(R) = \sum_{i=1}^{N} w_i r_i$.

The details of the proof of Theorem 3.5 are given in Appendix 3.8.1, but for a brief outline, we may observe that for any other rate vector which has the same sum rate as the rate vector $\mathbf{r}$ output by Algorithm 3, we must have either (1) if it can achieve universal recovery, it has equal or larger weighted cost than $\mathbf{r}$; or (2) it cannot achieve universal recovery, hence it should not be considered.

Remark 3.4. *In words, Theorem 3.5 says that the output rate vector of Algorithm 3 is the optimal rate vector which has the minimum weighted cost among all the rate vectors which have sum rate R and can achieve universal recovery.*

Algorithm 3 Search (d,K)-Basis (SdB)

Input: $E=[e_1,\ldots,e_N]^{\mathsf{T}}$ ($w_i \le w_j \ \forall i \le j$) and d.
Output: True, $\mathbf{r}$, $\mathbf{V}$ or **False**.
Initialization: $\mathbf{Q}=\emptyset$, $\mathbf{V}=\emptyset$, $\mathbf{r}=[r_1,\ldots,r_N]^{\mathsf{T}}=\mathbf{0}_{1\times N}$.
for $i : i \in \{1,\ldots,N\}$ **do**
  for $b \in \mathscr{B}(e_i,d)$ **do**
    if $b \notin \mathscr{G}(\mathbf{Q},d)$ **then**
      $r_i = r_i + 1$
      $\mathbf{V} = \mathbf{V} \cup \{b\}$
      while $\exists \mathbf{S} \subseteq \mathbf{Q}, |\mathbf{S}| \le 2 :$ (3.49) *holds* **do**

$$w_H(q_{\mathbf{S}} \vee b) \le \sum_{q_i \in \mathbf{S}} w_H(q_i) + w_H(b) - |\mathbf{S}|d \tag{3.49}$$

        $b = b \vee q_{\mathbf{S}}$, $\mathbf{Q} = \mathbf{Q} \setminus \mathbf{S}$
      end while
      $\mathbf{Q} = \mathbf{Q} \cup \{b\}$
    end if
    if $|\mathbf{V}| = K - d$ **then**
      return **True**, $\mathbf{r}$ and $\mathbf{V}$
    end if
  end for
end for
return **False**

Comparing to Algorithm 1 which checks the existence of any balanced (d,K)-Basis for the basic cooperative data exchange problem and outputs the corresponding (d,K)-Basis vectors if they exist, Algorithm 3 requires that the input PDVs be ordered according to their weights. The nodes with smaller weights have smaller indices. The node with the smallest weight would be selected to generate as many (d,K)-Basis vectors as it can. Then, the nodes with larger weights would be selected to generate (d,K)-Basis vectors that can not be generated by previous nodes. We show that by ordering the input PDVs in ascending order of their weights, Algorithm 3 can find the optimal rate vector and corresponding (d,K)-Basis vectors which can achieve universal recovery by using $K-d$ transmissions and has the minimum overall weighted cost. The ordering of the PDVs according to their weights can be done before the start of Algorithm 3 and only requires complexity $\mathcal{O}(N\log(N))$. As compared to the complexity of searching the existence of a (d,K)-Basis, which is $\mathcal{O}(N^3K^3)$, the complexity of pre-ordering nodes can be ignored.

Now we have a method to get the optimal solution to the sub-optimization problem (3.47). In order to get the globally optimal solution to the optimization problem (3.48), it is sufficient to only consider the rate vectors that are output by Algorithm 3 with different values of input parameter d ($d = K-R$). However, it is not necessary to run Algorithm 3 with all possible $R \in \{R_{min},\ldots,K\}$, by leveraging convexity of the function $\mathscr{K}(R)$ which is stated by the following theorem. Hence the optimal weighted cost and rate vector can be found by a binary search style method.

Theorem 3.6. *For $R_{min} \le R \le K$, the function defined by* (3.47)*: $\mathscr{K}(R) = \min_{\mathbf{r}\in\Omega,\mathscr{S}(\mathbf{r})=R} \sum_{i=1}^{N} w_i r_i$ is convex.*

The proof is given in Appendix 3.8.2. To prove Theorem 3.6, it is sufficient to only consider coding schemes with rate vectors output by Algorithm 3, since they are the conditionally optimal solution for fixed sum rate R. In particular, we exploit some properties of the rate vector output by Algorithm 3 to show that the second order difference of $\mathscr{K}(R)$ is non-negative, i.e. $\mathscr{K}(R+2)+\mathscr{K}(R)-2\mathscr{K}(R+1) \ge 0$. By induction, we prove that $\mathscr{K}(R)$ is a convex function of R.

Remark 3.5. *In [6], it has been proved that the function $\mathscr{K}(R)$ defined in* (3.47) *is convex for $R_{min} \le R \le K$ for a relaxed condition where each entry of $\mathbf{r} = [r_1,\ldots,r_N]^\mathsf{T}$ can be non-integer rate vector. However, the entries of the rate vector should always be integers for the cooperative data exchange problem. The improvement of our theorem is that we prove that for integer rate vectors, the function $\mathscr{K}(R)$ defined in* (3.47) *is still convex for $R_{min} \le R \le K$.*

Since the function $\mathscr{K}(R)$ is a convex function, it is not necessary to search all possible R to get the optimal solution to optimization problem (3.48). We propose Algorithm 4 to compute the minimum weighted cost by using a binary search method.

The complexity of the binary search of Algorithm 4 is approximately $\mathcal{O}(\log(K))$. Hence, the overall complexity of our two algorithms is $\mathcal{O}(N^3K^3\log(K))$ which is the same complexity as

Algorithm 4 Finding $\mathbf{r}^*$ and $\mathscr{C}^*$ using Binary Search Algorithm

Input: $E = [e_1, \ldots, e_N]^\mathsf{T}$, K and $\mathbf{w} = [w_1, \ldots, w_N]^\mathsf{T}$ such that $(w_i \leq w_j \ \forall i \leq j)$
Output: $\mathbf{r}^*$ and $\mathscr{C}^*$
Initialization: $d_{start} = 0$, $d_{end} = \mathscr{M}$
while $d_{start} < d_{end}$ **do**
    $d = \max\{\lfloor \frac{d_{start}+d_{end}}{2} \rfloor, d_{start}+1\}$
    $(F, \mathbf{r}, \mathbf{V}) = SdB(E, d)$
    if F is **False then**
        $d_{end} = d$
    else
        $\hat{d} = d - 1$
        $(\hat{F}, \hat{\mathbf{r}}, \hat{\mathbf{V}}) = SdB(E, \hat{d})$
        if $\mathbf{w}^\mathsf{T} \cdot \mathbf{r} > \mathbf{w}^\mathsf{T} \cdot \hat{\mathbf{r}}$ **then**
            $d_{end} = \hat{d}$, $\mathbf{r}^* = \hat{\mathbf{r}}$
        else
            $d_{start} = d$, $\mathbf{r}^* = \mathbf{r}$
        end if
    end if
end while
$R^* = K - d$, $\mathscr{C}^* = \mathbf{w}^\mathsf{T} \cdot \mathbf{r}$

the complexity of algorithms for the basic cooperative data exchange problem[6].

Example 3.7. *On applying Algorithm 3 on Example 3.6 for $d = \{0,1,2,3,4\}$, we can get the results as shown in Table 3.1. As can be seen from the table, the minimum cost is achieved by a*

Table 3.1 – Sum rate, optimal weighted cost and rate vector

d	R=K-d	$\mathscr{K}(R)$	r_1	r_2	r_3	r_4	r_5
4	5	29	1	1	1	1	1
3	6	22	2	2	2	0	0
2	7	21	3	3	1	0	0
1	8	23	4	3	1	0	0
0	9	25	5	3	1	0	0

coding scheme that uses 7 transmissions, which is larger than the minimum number of required transmissions ($R_{min} = 5$) for achieving universal recovery. Additionally, if we plot the function $\mathscr{K}(R)$ vs R for example 3.6 and connect the points, it is easy to see the convexity in Fig. 3.1.

[6] If we include the pre-ordering process for the nodes, the overall complexity should be $\mathscr{O}(N^3K^3\log(K) + N\log(N))$. But as we mentioned, the complexity of pre-ordering can be ignored compared to the complexity of other parts.

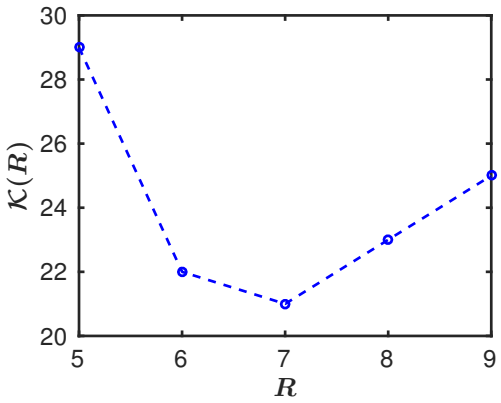

Figure 3.1 – Optimal weighted cost ($\mathscr{K}(R)$) vs Sum rate (R) for Example 3.6.

3.6 Successive Omniscience

In the basic cooperative data exchange problems, all nodes have the same priority and should be able to recover all packets at the end of the communication phase. In this section, we consider a generalized problem called *Successive Local Omniscience (SLO)* [12, 74], where nodes have different priorities. Specifically, let $\mathbf{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_M\}$ be a partition of nodes $\{1, \ldots, N\}$ In the SLO problem, communication occurs in M rounds, numbered from 1 to M and taking place in this order, as follows:

- In round i, only the nodes in the set $\mathbf{G}_{[i]} \stackrel{\text{def}}{=} \cup_{j=1}^{i} \mathbf{G}_j$ are allowed to transmit.
- After round i, all nodes in the set $\mathbf{G}_{[i]}$ must be able to recover all packets that were initially present at all the nodes in the set $\mathbf{G}_{[i]}$.

In this sense, if $i < k$, then nodes in G_i can be thought of as having priority over nodes in G_k (although in the general case, no node is guaranteed to attain full omniscience of *all* packets before the end of the last round).

Let $\mathbf{r}^i = [r_1^i, \ldots, r_N^i]^\mathsf{T}$ denote the accumulated rate vector up to and including the i^{th} round, where each r_j^i denotes the total number of transmissions made by node j from the first round to the i^{th} round. The corresponding entries of rate vectors $\mathbf{r}^i$ and $\mathbf{r}^{i+1}$ satisfy $r_j^i \leq r_j^{i+1}$ for every node $j \in \{1, \ldots, N\}$. Let $\Omega(\mathbf{G}_{[i]})$ be the set of rate vectors up to and including the i^{th} communication round satisfying

$$\sum_{j \in \mathbf{G}_{[i]} \setminus \mathscr{J}} r_j^i \geq \left| \mathbf{X}_{\mathbf{G}_{[i]}} \setminus \mathbf{X}_{\mathscr{J}} \right|, \forall \mathscr{J} \subsetneq \mathbf{G}_{[i]} \tag{3.50}$$

Then we have the following lemma characterizing solutions to the SLO problem:

Lemma 3.6. *Any solution to the SLO problem is also a solution to the following multi-objective*

linear program:

$$\min_{\mathbf{r}^i \in \Omega(\mathbf{G}_{[i]})} \sum_{j=1}^{N} r_j^i, \forall i \in \{1,\ldots,M\} \tag{3.51}$$

Proof. For any $i \in \{1,\ldots,M\}$, rate vectors $\mathbf{r}^i \in \Omega(\mathbf{G}_{[i]})$ satisfy the Slepian-Wolf constraints for achieving local omniscience and only nodes in $\mathbf{G}_{[i]}$ are allowed to make transmissions. The minimization gives the minimum sum rate. Thus, for M communication rounds, the overall optimal solutions achieve successive local omniscience. □

In this section, we present an efficient solution of the SLO problem via the (d,K)-Basis method. Let $E_{\mathbf{G}_{[i]}}$ denote the packet distribution matrix of the nodes in $\mathbf{G}_{[i]}$. If we run Algorithm 2 with $E_{\mathbf{G}_{[i]}}$ as input in the subspace indexed by the collectively available packets of $\mathbf{G}_{[i]}$, it will return the minimum number of required transmissions for achieving local omniscience as well as the corresponding (d,K_i)-Basis vectors. Algorithm 2 can be called for every $E_{\mathbf{G}_{[i]}}$, $\forall i \in \{1,\ldots,M\}$ and we can get the d_i-Basis vectors for local omniscience achieved by each $\mathbf{G}_{[i]}$. If $d_i \geq d_{i+1}$, the (d_i,K_i)-Basis vectors can also be used to generate (d_{i+1},K_{i+1})-Basis vectors by adding 0's to the dimensions that are added by packets in $\mathbf{X}_{\mathbf{G}_{[i+1]}} \setminus \mathbf{X}_{\mathbf{G}_{[i]}}$. If $d_i < d_{i+1}$, the (d_i,K_i)-Basis vectors cannot be used to generate (d_{i+1},K_{i+1})-Basis vectors. Hence, the optimal strategy is to use the coding scheme based on (d_i,K_i)-Basis in the subspace indexed by packets of $\mathbf{X}_{\mathbf{G}_{[i+1]}}$ so that every transmission used in the previous round are useful in the current round.

Theorem 3.7. *For successive local omniscience problem with $\mathbf{G}_{[i]}$ and corresponding packet distribution submatrix $E_{\mathbf{G}_{[i]}}$, for $i \in \{1,\ldots,M\}$, the minimum number of required transmissions R_i^* for round i is*

$$R_i^* = K_i - \min\{\mathscr{M}_i, d_1^*, \ldots, d_i^*\}, \tag{3.52}$$

where $K_i = |\mathbf{X}_{\mathbf{G}_{[i]}}|$ denotes the number of packets collectively available at nodes in $\mathbf{G}_{[i]}$, $\mathscr{M}_i = \min_{j \in \mathbf{G}_{[i]}} |X_j|$ is the minimum number of available packets at any single node in $\mathbf{G}_{[i]}$ and d_i^* is the maximum (d,K_i)-Basis that can be generated by PDVs of nodes in $\mathbf{G}_{[i]}$.

Proof. For the first round, $R_1^* = K_1 - \min\{\mathscr{M}_1, d_1^*\}$, according to Theorem 3.3. For the i^{th} round, since $\mathbf{G}_{[j]} \subset \mathbf{G}_{[i]}$, $\forall j < i$, $\mathscr{M}_i \leq \mathscr{M}_j$. According to Theorem 3.3, nodes in $\mathbf{G}_{[i]}$ can generate a coding scheme which is based on the $\{\mathscr{M}_i, d_i^*\}$-Basis and can achieve the local omniscience. If $d_i^* = \min\{d_1^* \ldots, d_i^*\}$, and all transmissions used in previous rounds can also be used as the transmissions of coding schemes based on the $\{\mathscr{M}_i, d_i^*\}$-Basis. Thus, in the i^{th} round, only additional transmissions are required and the total minimum number of required transmissions for achieving local omniscience is $R_i^* = K_i - \min\{\mathscr{M}_i, d_i^*\}$. If $d_j^* = \min\{d_1^* \ldots, d_i^*\}$ and $j < i$, then transmissions generated in the j^{th} round cannot all be used for the coding scheme based on the $\{\mathscr{M}_i, d_i^*\}$-Basis. In order to make use of all previously generated transmissions, the coding scheme based on $\{\mathscr{M}_i, d_j^*\}$-Basis can be used to achieve

local omniscience for nodes in $\mathbf{G}_{[i]}$ and the total number of required transmissions is $R_i^* = K_i - \{\mathscr{M}_i, d_j^*\}$. Therefore, $R_i^* = K_i - \min\{\mathscr{M}_i, d_1^*, \ldots, d_i^*\}$. □

We propose Algorithm 5 to compute the minimum number of required transmissions (R_i^*) and the local optimal rate vector ($\mathbf{r}_i^*$) for nodes in each group with different priorities. Algorithm 5 iteratively calls Algorithm 1 to search for the existence of a (d, K_i)-Basis that can be generated for linear coding schemes to achieve local omniscience.

Algorithm 5 Successive Local Omniscience

Input: $E = [e_1, \ldots, e_N]^\mathsf{T}$ and $\mathbf{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_M\}$
Output: $R_1^*, \ldots, R_M^*$ and $\mathbf{r}_1^*, \ldots, \mathbf{r}_M^*$
Initialization: $d^* = K$
for $i = 1 \ldots M$ **do**
  $d_{min} = 1$, $d_{max} = \min\{\mathscr{M}_i, d^*\}$
  $(F, \mathbf{r}, \mathbf{V}) = SdB(E_{\mathbf{G}_{[i]}}, d_{end})$
  if F is **True then**
    $d_i^* = d_{max}$, $\mathbf{V}_i^* = \mathbf{V}$, $\mathbf{r}_i^* = \mathbf{r}$
  else
    $(F, \mathbf{r}, \mathbf{V}) = SdB(E_{\mathbf{G}_{[i]}}, d_{min})$
    if F is **False then**
      $d_i^* = 0$, $\mathbf{V}_i^* = I_{K_i}$
    else
      while $d_{max} - d_{min} > 1$ **do**
        $d = \lfloor \frac{d_{min} + d_{max}}{2} \rfloor$
        $(F, \mathbf{r}, \mathbf{V}) = SdB(E, d)$
        if F is **True then**
          $d_{min} = d$, $d_i^* = d$, $\mathbf{V}_i^* = \mathbf{V}$, $\mathbf{r}_i^* = \mathbf{r}$
        else
          $d_{max} = d$
        end if
      end while
    end if
  end if
  $d^* = d_i^*$, $R_i^* = K_i - d_i^*$
end for

Based on the (d, K_i)-Basis vectors $\mathbf{V}_i^*$ and local optimal rate vector r_i^*, the corresponding linear coding scheme can be generated to achieve local omniscience. Instead of generating linear coding schemes for each communication round individually, it is possible to globally generate a linear coding scheme in which the first R_i^* transmissions can achieve local omniscience.

Regarding the complexity of our approach, in each communication round, the minimum number of required transmissions and the accumulated rate vector is found by using binary search method and iteratively calling Algorithm 1. The total number of outer iterations is equal to the number of priority groups, M. The binary search method for the i^{th} round has

complexity bounded by $\mathcal{O}(\log(K_i))$. For the i^{th} round, Algorithm 1 has complexity bounded by $\mathcal{O}(|\mathbf{G}_i|^3 K_i^3)$, since the number of new nodes and packets considered in the i^{th} round are $|\mathbf{G}_i|$ and K_i, respectively. We note that for algorithm 1, the nodes in $\mathbf{G}_{[i-1]}$ have already been checked in previous interations and the basis vectors generated by them in previous iterations can be reused in current iteration. Hence, we do not need to check them again. Therefor, the total number of computation can be expressed as $\sum_{i=1}^{M} |\mathbf{G}_i|^3 K_i^3 \log(K_i)$. Since $\forall\, i$: $|\mathbf{G}_{[i]}| \leq N$, $K_i \leq K$ and $\sum_{i=1}^{M} |\mathbf{G}_i| = N$, we have $\sum_{i=1}^{M} |\mathbf{G}_i|^3 K_i^3 \log(K_i) \leq N^2 K^3 \log(K) \sum_{i=1}^{M} |\mathbf{G}_i| = N^3 K^3 \log(K)$. The overall complexity of our (d,K)-Basis method for solving SLO problem is bounded by $\mathcal{O}(N^3 K^3 \log(K))$.

Example 3.8. *Consider the successive local omniscience problem with the following packet distribution matrix*

$$E = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{3.53}$$

And the nodes are partitioned into three groups with decreasing priorities: $\mathbf{G}_1 = \{1,2\}$, $\mathbf{G}_2 = \{3,4\}$ *and* $\mathbf{G}_3 = \{5,6\}$. *Since nodes in* $\mathbf{G}_1$ *collectively only have packets* $P_1, \ldots, P_5$, *the optimization for the first communication round is equivalent to the basic CDE problem with packet distribution matrix* $E_{\mathbf{G}_1}$, *which is a submatrix of the first two rows and five columns of* E.

$$E_{\mathbf{G}_1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3.54}$$

It is apparent that only two transmissions are required to achieve local omniscience for $\mathbf{G}_1$. *Consider the following two coding schemes:*

- *Coding scheme 1: Node 1 sends* P_1 *and Node 2 sends* P_5.
- *Coding scheme 2: Node 1 sends* $P_1 + P_2 + P_3 + P_4$ *and Node 2 sends* $P_2 + P_3 + P_4 + P_5$.

In Coding scheme 1, each transmission is a linear combination of as few packets as possible, while in Coding scheme 2, each transmission is a linear combination of as many packets as possible. Both coding schemes can enable two nodes to fully recover packets that are collectively available at them. However, we will show that Coding scheme 1 is suboptimal but Coding scheme 2 is optimal. In the second communication round, the goal is to enable node in $\mathbf{G}_{[2]}$ *to recover packets which are collectively available at them. Similarly, we have the packet*

distribution matrix $E_{\mathbf{G}_{[2]}}$, which is a submatrix of the first four rows and seven columns of E.

$$E_{\mathbf{G}_{[2]}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{3.55}$$

If we treat this as a packet distribution matrix of a basic CDE problem, it is easy to find that the minimum number of required transmission is 5*, since the* $(2,7)$*-Basis is the* $(d,7)$*-Basis with largest d value that can be generated by row vectors of $E_{\mathbf{G}_{[2]}}$. And this implies that in the successive local omniscience problem, the total number of required transmissions is at least* 5*. If we choose Coding scheme 1 in the first transmission round, the packet distribution matrix becomes*

$$\hat{E}_{\mathbf{G}_{[2]}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \tag{3.56}$$

As the row vectors of $\hat{E}_{\mathbf{G}_{[2]}}$ can only generate a $(3,7)$*-Basis which has largest d value,* 4 *transmissions are required in the second communication round to achieve local omniscience for nodes in $\mathbf{G}_{[2]}$. Hence, the total number of transmissions for the first and second rounds is* $2+4=6$ *which is larger than the lower bound* 5*. However, if Coding Scheme* 2 *is chosen in the first round, it is possible to generate a coding scheme based on* $(2,7)$*-Basis in which the first two transmissions achieve local omniscience for nodes in $\mathbf{G}_1$. The desired* $(2,7)$*-Basis generated by $E_{\mathbf{G}_{[2]}}$ is*

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{3.57}$$

As one can see the first 5 *columns of v_1 and v_2 can actually form a* $(3,5)$*-Basis. And the coding scheme based on them can achieve local omniscience for nodes in $\mathbf{G}_1$. Similarly, we can show that* 2 *transmissions are required in the third communication round to achieve omniscience for nodes in $\mathbf{G}_{[3]}$. Instead of generating coefficients for linear combinations of packets for each round individually, we can deal with them together by constructing a linear coding scheme based on the final* (d,K)*-Basis we need, which is a* $(2,9)$*-Basis in this case. Given the rate vector in each round:*

$$\mathbf{r}_1 = [1,1,0,0,0,0]^\mathsf{T} \tag{3.58}$$

$$\mathbf{r}_2 = [0,1,1,1,0,0]^\mathsf{T} \tag{3.59}$$

$$\mathbf{r}_3 = [0,0,0,0,1,1]^\mathsf{T} \tag{3.60}$$

And the $(2,9)$*-Basis that generated by row vectors of* E

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{3.61}$$

By using the coding construction method based on MDS code in Section 3.4, we can get a coefficient matrix as follows, where all entries are over finite field $GF(2^4)$ *with primitive polynomial* $\alpha^4+\alpha+1$.

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \end{bmatrix} = \begin{bmatrix} 4 & 7 & 3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 12 & 3 & 2 & 0 & 0 & 0 & 0 \\ 0 & 13 & 13 & 2 & 2 & 0 & 0 & 0 & 0 \\ 15 & 8 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 4 & 5 & 0 & 0 & 10 & 0 & 0 \\ 10 & 0 & 9 & 9 & 5 & 5 & 5 & 5 & 0 \\ 9 & 4 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{3.62}$$

It can be verified that the first 2 *transmissions achieve local omniscience for nodes in* $\mathbf{G}_1$*, the first* 5 *transmissions achieve local omniscience for nodes in* $\mathbf{G}_{[2]}$*, and all transmissions together achieve omniscience for nodes in* $\mathbf{G}_{[3]}$ *(all nodes).*

3.7 Conclusion

In this chapter, we introduce the concept of (d,K)-Basis. We establish that the existence of such a basis is both a necessary and sufficient condition for the existence of coding schemes that can achieve universal recovery with $K-d$ transmissions for the fully connected network. We provide a polynomial-time deterministic algorithm based on the (d,K)-Basis construction which solves the cooperative data exchange problem. We show that we can efficiently construct the coefficients of an optimal linear coding scheme starting from a Vandermonde matrix by levering the connection between the (d,K)-Basis and maximum distance separable codes. Moreover, we demonstrate that our (d,K)-Basis construction method can also be used in solving generalized versions of the cooperative data exchange problem, including CDE with weighted cost and with successive local omniscience.

3.8 Appendix

3.8.1 Proof of Theorem 3.5

In order to prove Theorem 3.5, we first prove two useful Lemmas.

Lemma 3.7. *Let* $\mathbf{r}^* = [r_1^*, r_2^*, \ldots, r_N^*]^\mathsf{T}$ *denote the rate vector output by Algorithm 3. For any rate vector* $\mathbf{r} = [r_1, \ldots, r_N]^\mathsf{T}$ *such that* $\mathbf{r} \in \Omega$ *and* $\mathscr{S}(\mathbf{r}^*) = \mathscr{S}(\mathbf{r})$*, there does not exists any node pair* (i, j) *such that* $i < j$*,* $r_i > r_i^*$ *and* $r_j < r_j^*$*.*

Proof. If the coding scheme with rate vector $\mathbf{r}$ can achieve universal recovery and uses the same total number of transmissions, then the coding scheme can be implemented as a (d, K)-Basis based coding scheme which has the same d value as the coding scheme with rate vector $\mathbf{r}^*$. As Algorithm 3 guarantees that $\forall i \in \{1, \ldots, N\}$, if $r_i^* > 0$, then there must exist as many as $\sum_{j=i}^{N} r_j^*$ (d, K)-Basis vectors that cannot be generated by nodes in set $\{1, 2, \ldots, i-1\}$. If $\exists i < j$ such that, $r_i > r_i^*$ and $r_j < r_j^*$, then $\sum_{j=i}^{N} r_j < \sum_{j=i}^{N} r_j^*$ which is not possible as such vectors can only be generated by nodes in set $\{i, i+1 \ldots, N\}$. Hence, it is impossible that $\exists i < j$: $r_i > r_i^*$ and $r_j < r_j^*$. □

Lemma 3.8. *Let* $\mathbf{r}^* = [r_1^*, r_2^*, \ldots, r_N^*]^\mathsf{T}$ *denote the rate vector output by Algorithm 3. If there exists a coding scheme with rate vector* $\mathbf{r} = [r_1, \ldots, r_N]^\mathsf{T}$ *such that* $\mathbf{r} \in \Omega$*,* $\mathscr{S}(\mathbf{r}^*) = \mathscr{S}(\mathbf{r})$*, and there exists node pair (i,j) such that* $i < j$*,* $r_i < r_i^*$ *and* $r_j > r_j^*$*, then* $\mathscr{C}(\mathbf{r}) \geq \mathscr{C}(\mathbf{r}^*)$*.*

Proof. Let $\mathbf{S}_1 = \{i : r_i < r_i^*\}$, $\mathbf{S}_2 = \{j : r_j > r_j^*\}$ and $\mathbf{S}_3 = \{k : r_k = r_k^*\}$. Since $\mathscr{S}(\mathbf{r}^*) = \mathscr{S}(\mathbf{r})$, we have

$$0 = \sum_{i=0}^{N} (r_i - r_i^*) = \sum_{i \in \mathbf{S}_1} (r_i - r_i^*) + \sum_{j \in \mathbf{S}_2} (r_j - r_j^*) + \sum_{k \in \mathbf{S}_3} (r_k - r_k^*) \tag{3.63}$$

$$= \sum_{i \in \mathbf{S}_1} (r_i - r_i^*) + \sum_{j \in \mathbf{S}_2} (r_j - r_j^*) \tag{3.64}$$

Hence, for each $i \in \mathbf{S}_1$ that sends one less transmission, there must exist one corresponding $j \in \mathbf{S}_2$ which sends one more transmission. According to Lemma 3.7, if there exists such pair of (i, j), it must satisfy $i < j$ and $w_i < w_j$. Let $P = \sum_{i \in \mathbf{S}_1} (r_i^* - r_i) = \sum_{j \in \mathbf{S}_2} (r_j - r_j^*)$ denote the total number of such pairs and $\mathscr{P}$ denote the partition of such pairs. Therefore,

$$\mathscr{C}(\mathbf{r}) - \mathscr{C}(\mathbf{r}^*) = \sum_{i=0}^{N} w_i r_i - \sum_{i=0}^{N} w_i r_i^* \tag{3.65}$$

$$= \sum_{i \in \mathbf{S}_1} w_i (r_i - r_i^*) + \sum_{j \in \mathbf{S}_2} w_j (r_j - r_j^*) \tag{3.66}$$

$$= \sum_{(i,j) \in \mathscr{P}} (w_j - w_i) \tag{3.67}$$

$$\geq 0 \tag{3.68}$$

□

Now, we are ready to prove Theorem 3.5.

Proof of Theorem 3.5. If there exists any linear coding scheme that achieves universal recovery by using $K-d$ transmissions with rate vector $\mathbf{r}=[r_1,\ldots,r_N]^\mathsf{T}$ ($\sum_{i=i}^{N} r_i = K-d$), it is always possible to generate a corresponding linear coding scheme based on the (d,K)-Basis that have the same rate vectors [73]. Hence, they have the same weighted cost and we can only consider the coding schemes based on (d,K)-Basis. Let $r^*=[r_1^*,\ldots,r_N^*]^\mathsf{T}$ denote the rate vector output by Algorithm 3. According to Lemma 3.7, there does not exist any $i<j$ such that $r_j<r_j^*$. Additionally, since $\mathscr{S}(\mathbf{r}^*)=\mathscr{S}(\mathbf{r})$, if rate vector $\mathbf{r}$ is different from $\mathbf{r}^*$, the change can only be $\exists i<j: r_i<r_i^*$ and $r_j>r_j^*$. According to Lemma 3.8, $\mathscr{C}(\mathbf{r})\geq\mathscr{C}(\mathbf{r}^*)$. Therefore, the rate vector output by Algorithm 3 has the minimum weighted cost in all coding schemes which use $K-d$ transmissions and achieve universal recovery. □

3.8.2 Proof of Theorem 3.6

In order to prove Theorem 3.6, we first prove two useful Lemmas.

Lemma 3.9. *Let* $\mathbf{r}(l)$ *be the rate vector output by Algorithm 3 for input* E *and* $d=K-l$. *Thus,* $\mathbf{r}(l)$ *is the optimal rate vector with minimum weighted cost among all the rate vectors with* $\mathscr{S}(\mathbf{r})=l$. *For the coding schemes which have rate vectors* $\mathbf{r}(l)=[r_{(l,1)},\ldots,r_{(l,N)}]^\mathsf{T}$ *with* $l\in\{R_{min},\ldots,K\}$ *yielded by Algorithm 3,we have*

(1) $r_{(l+1,1)}=r_{(l,1)}+1$.

(2) $r_{(l+1,m)}\leq r_{(l,m)}+1$, $\forall 2\leq m\leq N$.

(3) If $r_{(l+1,m)}<r_{(l,m)}$, *then* $r_{(l+2,m)}\leq r_{(l+1,m)}$.

Proof. (1) Since in Algorithm 3, we always start the generation of basis vectors from the PDV of node 1 and there is no previously generated basis vector, then the number of basis vectors that should be generated by node 1 is

$$r_{(l,1)}=w_H(e_1)-d=w_H(e_1)-K+l \tag{3.69}$$

Since $R_{min}\leq l\leq K$ and $R_{min}=K-\min\{\mathscr{M},d^*\}$, we have $0\leq r_{(l,1)}\leq w_H(e_1)$. We note that $w_H(e_1)\geq\mathscr{M}\geq K-R_{min}$. Therefore, for any feasible l, we have $r_{(l+1,1)}=r_{(l,1)}+1$. This means the first node generates 1 more vector when the total number of transmissions increases by 1. When $r_{(l,1)}=|X_1|$, each transmissions is just a pure packet. In such cases, we have $d=0$ and $l=K$. Universal recovery can always be achieved when all packets have been sent individually. No coding scheme with more than K transmissions should be considered.

(2) Similarly, for any $2 \leq m \leq N$, the total number of feasible basis vectors that can be generated by node m is $w_H(e_m) - K + l$. However, some of them may not be compatible with basis vectors that have been generated by previous nodes. Hence we have

$$r_{(l,m)} \leq w_H(e_m) - K + l \tag{3.70}$$

And $r_{(l+1,m)} \leq r_{(l,m)} + 1$, $\forall 2 \leq m \leq N$. This means node m can generate at most 1 more basis vector when the total number of transmissions increases by 1.

(3) As the total number of transmissions (sum rate) goes from l to $l+1$, the corresponding basis change from $(K-l)$-Basis to $(K-l-1)$-Basis. Therefore, the number of packets that are used to generate each transmission decreases by 1. Note that $w_H(e_m) \geq \mathscr{M} \geq K - R_{min}$, $\forall m \in \{1,\ldots,N\}$. When $l = R_{min}$, nodes m with $w_H(e_m) = K - R_{min}$ are not considered to generate any basis vector, since every basis vector needs $K - R_{min} + 1$ ones. But when $l > R_{min}$, every node is considered to generate basis vectors. If node i is not used to generate any basis vector, that means all basis vectors that can be generated by node i are not compatible with the basis vectors generated by previous nodes. If $r_{(l+1,m)} < r_{(l,m)}$, that means besides the first node, there exists at least one node with lower weight than node m that generates more basis vector(s), i.e. $\exists n$ s.t. $n < m$ and $r_{(l+1,n)} > r_{(l,n)}$. The set of basis vectors that are generated to form the $(K-l-1)$-Basis by node m is a subset of $\mathscr{B}(e_m, K-l)$. Let $\mathscr{D}(m, l+1)$ denote vectors in $\mathscr{B}(e_m, K-l)$ but are not selected to form the $(K-l-1)$-Basis. Then every vector in $\mathscr{D}(m, l+1)$ is not compatible with the $(K-l-1)$-Basis vectors generated by previous nodes. Any vector in $\mathscr{B}(e_m, K-l-1)$ which can be generated by vectors in $\mathscr{D}(m, l+1)$ is also not compatible with the $(K-l-2)$-Basis vectors generated by previous nodes. Hence, the maximum number of basis vectors that can be generated by node m for the next round is upper-bounded by $r_{(l+1,m)}$. Therefore, If $r_{(l+1,m)} < r_{(l,m)}$, then $r_{(l+2,m)} \leq r_{(l+1,m)}$, $\forall 2 \leq m \leq N$.

□

Definition 3.9. *Let $\mathbf{S}_{(l,\uparrow)}$ denote the set of nodes which generate more number of transmissions when the sum rate increases from l to $l+1$. Let $\mathbf{S}_{(l,0)}$ denote the set of nodes which generate the same number of transmissions when the sum rate increases from l to $l+1$. Let $\mathbf{S}_{(l,\downarrow)}$ denote the multiset of nodes that generate fewer transmissions when the sum rate increases from l to $l+1$. The multiplicity of node i in $\mathbf{S}_{(l,\downarrow)}$ equals $r_{(l,i)} - r_{(l+1,i)}$.*

Lemma 3.10. *For $\forall R_{min} \leq l \leq K-1$, we have (1) $\mathbf{S}_{(l+1,\uparrow)} \subseteq \mathbf{S}_{(l,\uparrow)}$ and (2) Let W_{l+1}^i be the i^{th} largest $w \in \{w_j : j \in \mathbf{S}_{(l+1,\downarrow)}\}$ and W_l^i be the i^{th} largest $w \in \{w_j : j \in \mathbf{S}_{(l,\downarrow)}\}$. For any W_{l+1}^i, there exists W_l^i such that $W_{l+1}^i \leq W_l^i$.*

Proof. Let $\mathbf{r}(l) = [r_{(l,1)},\ldots,r_{(l,N)}]^\mathsf{T}$ and $\mathbf{r}(l+1) = [r_{(l,1)},\ldots,r_{(l+1,N)}]^\mathsf{T}$ denote the rate vectors output by Algorithm 3 for $d = K-l$ and $d = K-l-1$, respectively. According to Theorem 3.5, $\mathbf{r}(l)$ and $\mathbf{r}(l+1)$ are optimal rate vectors for fixed sum rate l and $l+1$, respectively.

(1) Assuming that $\mathbf{S}_{(l+1,\uparrow)} \nsubseteq \mathbf{S}_{(l,\uparrow)}$, then there must exist at least one node k, such that $k \in \mathbf{S}_{(l+1,\uparrow)}$

and $k \notin \mathbf{S}_{(l,\uparrow)}$. Hence, k must be in $\mathbf{S}_{(l,0)}$ or $\mathbf{S}_{(l,\downarrow)}$. It is apparent that $k \neq 1$, since the first node always increases the rate by 1 when the total sum-rate increases by 1. For $k \in \mathbf{S}_{(l+1,\uparrow)} \setminus \{1\}$, there must always exist a corresponding node $m \in \mathbf{S}_{(l+1,\downarrow)}$ such that $w_k < w_m$.

(i) If $k \in \mathbf{S}_{(l,0)}$, we know that $r_{(l+1,k)} = r_{(l,k)}$. Coding schemes with rate vector $\hat{\mathbf{r}}(l) = [\hat{r}_{(l,1)}, \ldots, \hat{r}_{(l,N)}]^{\mathsf{T}}$ such that

$$\hat{r}_{(l,k)} = r_{(l+1,k)} = r_{(l,k)} + 1 \tag{3.71}$$

$$\hat{r}_{(l,m)} = r_{(l+1,m)} = r_{(l,m)} - 1 \tag{3.72}$$

$$\hat{r}_{(l,i)} = r_{(l,i)}, \forall\, i \in \{1, \ldots, N\} \setminus \{k, m\} \tag{3.73}$$

can also achieve universal recovery. Moreover, coding scheme with rate vector $\hat{\mathbf{r}}(l)$ has lower cost than coding scheme with rate vector $\mathbf{r}(l)$. This contradicts that coding scheme with rate vector $\mathbf{r}(l)$ is optimal for all rate vector with sum rate l.

(ii) If $k \in \mathbf{S}_{(l,\downarrow)}$, we know that $r_{(l+1,k)} < r_{(l,k)}$. According to Lemma 3.9, $r_{(l+1,k)} \leq r_{(l,k)}$. This contradicts our assumption that $k \in \mathbf{S}_{(l+1,\uparrow)}$.

Therefore, we have $\mathbf{S}_{(l+1,\uparrow)} \subseteq \mathbf{S}_{(l,\uparrow)}$.

(2)We use the induction proof method to prove this part of lemma. For $i = 1$, let $W_{l+1}^1 = w_m$, $W_l^i = w_n$. We assume that $W_{l+1}^i > W_l^i$, then we have $w_m > w_n$ which implies that $m \notin \mathbf{S}_{(l,\downarrow)}$. Since $\mathbf{S}_{(l+1,\uparrow)} \subseteq \mathbf{S}_{(l,\uparrow)}$, coding scheme with rate vector $\hat{\mathbf{r}}(l) = [\hat{r}_{(l,1)}, \ldots, \hat{r}_{(l,N)}]^{\mathsf{T}}$ which satisfies

$$\hat{r}_{(l,m)} = r_{(l,m)} - 1 \tag{3.74}$$

$$\hat{r}_{(l,n)} = r_{(l,n)} + 1 \tag{3.75}$$

$$\hat{r}_{(l,j)} = r_{(l,j)}, \forall\, j \in \{1, \ldots, N\} \setminus \{m, n\} \tag{3.76}$$

can also achieve universal recovery with the same sum-rate and has lower weighted cost. This contradicts that coding scheme with rate vector $\mathbf{r}(l) = [r_{(l,1)}, \ldots, r_{(l,N)}]^{\mathsf{T}}$ is optimal for all rate vector with sum rate l. Thus, we have $W_{l+1}^1 \leq W_l^1$. For $i > 1$, assuming that $W_{l+1}^{i-1} \leq W_l^{i-1}$, we show that $W_{l+1}^i \leq W_l^i$. let $W_{l+1}^i = w_a$, $W_l^i = w_b$. If $W_{l+1}^{i-1} \leq w_b$, then it is straightforward that $w_a = W_{l+1}^i \leq W_{l+1}^{i-1} \leq w_b = W_l^i$. If $W_{l+1}^{i-1} > w_b$, and we assume that $w_a > w_b$. In such cases, $a \notin \mathbf{S}_{(l,\downarrow)}$, since $w_a \leq W_{l+1}^{i-1} \leq W_l^{i-1}$. By using a similar trick as we used for $i = 1$, it is able to show that there exists another coding scheme which achieves universal recovery and has a lower sum weighted cost. Hence the assumption $w_a > w_b$ can never be true. Therefore, $W_{l+1}^i \leq W_l^i$.

□

Now we are ready to prove Theorem 3.6.

Proof of Theorem 3.6. For any $R_{min} \leq l \leq K-2$, we show that the second order difference of

$\mathscr{K}(l)$ is non-negative, i.e. $\mathscr{F}(l+1)-\mathscr{F}(l)\geq 0$, where $\mathscr{F}(l)=\mathscr{K}(l+1)-\mathscr{K}(l)$. We compute the difference of the weighted cost of two coding schemes when sum-rate increases by 1.

$$\mathscr{F}(l+1)=\mathscr{K}(l+2)-\mathscr{K}(l+1) \tag{3.77}$$
$$=\sum_{i\in\mathscr{S}_{(l+1,\uparrow)}} w_i-\sum_{i\in\mathscr{S}_{(l+1,\downarrow)}} w_i \tag{3.78}$$
$$=w_1+\sum_{i\in\mathscr{S}_{(l+1,\uparrow)}\setminus\{1\}} w_i-\sum_{i\in\mathscr{S}_{(l+1,\downarrow)}} w_i. \tag{3.79}$$

According to Lemma 3.9, node 1 always generates 1 more transmission when the total number of transmissions increases by 1. And for other nodes, if their rate increases, the increment is 1, whereas if their rate decreases, the decrement can be more than 1. And the number of multiplications of the nodes in $\mathscr{S}_{(l+1,\downarrow)}$ is equal to the decrease in rate. Similarly, for sum-rate change from l to $l+1$, we have

$$\mathscr{F}(l)=\mathscr{K}(l+1)-\mathscr{K}(l)=w_1+\sum_{i\in\mathscr{S}_{(l,\uparrow)}\setminus\{1\}} w_i-\sum_{i\in\mathscr{S}_{(l,\downarrow)}} w_i. \tag{3.80}$$

The reason why node 1 is separated from other nodes is that the total number of transmissions only increases by 1, which implies that the total number of transmissions sent by other nodes, except node 1, remains the same. Hence

$$|\mathscr{S}_{(l,\uparrow)}\setminus\{1\}|=|\mathscr{S}_{(l,\downarrow)}| \tag{3.81}$$
$$|\mathscr{S}_{(l+1,\uparrow)}\setminus\{1\}|=|\mathscr{S}_{(l+1,\downarrow)}| \tag{3.82}$$

Therefore, $\forall i\in\mathscr{S}_{(l,\uparrow)}\setminus\{1\}$, $\exists j\in\mathscr{S}_{(l,\downarrow)}$ such that $w_i<w_j$. We can construct a partition of node pairs (i,j), where $i\in\mathscr{S}_{(l,\uparrow)}\setminus\{1\}$ and $j\in\mathscr{S}_{(l,\downarrow)}$ as follows

$$\mathscr{P}(l)=\{(i,j): i\in\mathscr{S}_{(l,\uparrow)}\setminus\{1\}, j\in\mathscr{S}_{(l,\downarrow)}, i<j\} \tag{3.83}$$

Note that the number of node pairs in $\mathscr{P}(l)$ is equal to $|\mathscr{S}_{(l,\uparrow)}\setminus\{1\}|$. Then we have

$$\mathscr{F}(l)=w_1+\sum_{(i,j)\in\mathscr{P}(l)}(w_i-w_j) \tag{3.84}$$

where every term of the summation (w_i-w_j) is negative.

We show that for each pair $(i,j)\in\mathscr{P}(l+1)$, there always exists a pair $(\hat{i},\hat{j})\in\mathscr{P}(l)$ such that

$$w_i-w_j-(w_{\hat{i}}-w_{\hat{j}})\geq 0 \tag{3.85}$$

Assuming that there exists a node pair $(i,j)\in\mathscr{P}(l+1)$ such that for all possible pairs $(\hat{i},\hat{j})\in\mathscr{P}(l)$:

$$w_i-w_j-(w_{\hat{i}}-w_{\hat{j}})<0 \tag{3.86}$$

Equivalently, we have

$$w_i - w_j < \max_{\hat{i}\in\mathscr{S}_{(l,\uparrow)},\hat{j}\in\mathscr{S}_{(l,\downarrow)}} (w_{\hat{i}} - w_{\hat{j}}) \tag{3.87}$$

If $i \in \mathscr{S}_{(l,\uparrow)}$, then $w_j > \max_{\hat{j}\in\mathscr{S}_{(l,\downarrow)}} w_{\hat{j}}$, which contradicts Lemma 3.10. If $i \notin \mathscr{S}_{(l,\uparrow)}$, consider another coding scheme with rate vector $\mathbf{r} = [r_1, r_2, \ldots, r_N]^\mathsf{T}$ such that

$$r_i = r_{(l+1,i)} + 1, r_j = r_{(l+1,j)} - 1 \tag{3.88}$$

$$r_{\hat{i}} = r_{(l+1,\hat{i})} - 1, r_{\hat{j}} = r_{(l+1,\hat{j})} + 1 \tag{3.89}$$

$$r_m = r_{(l+1,m)}, \forall m \notin \{i, j, \hat{i}, \hat{j}\} \tag{3.90}$$

It can be verified that this coding scheme can also achieve universal recovery with total $l+1$ transmissions. It has lower weighted cost than the coding scheme with rate vector $[r_{(l+1,1)}, \ldots, r_{(l+1,N)}]^\mathsf{T}$, which contradicts that coding scheme with rate vector $[r_{(l+1,1)}, \ldots, r_{(l+1,N)}]^\mathsf{T}$ has the minimum weighted cost over all coding schemes that achieve universal recovery with $l+1$ transmissions. Starting form the node pair (i, j) with largest j, we can apply this binding for every (i, j) and remove used $(\hat{i}, \hat{j})$ iteratively. And it is able to find $(\hat{i}, \hat{j}) \in \mathscr{P}(l)$ such that Eqn (3.85) is satisfied for every pair $(i, j) \in \mathscr{P}(l+1)$. Hence, we have

$$\begin{aligned}
&\mathscr{F}(l+1) - \mathscr{F}(l) \\
&= \sum_{(i,j)\in\mathscr{P}(l+1)} (w_i - w_j) - \sum_{(m,n)\in\mathscr{P}(l)} (w_m - w_n) && (3.91) \\
&= \sum_{(i,j)\in\mathscr{P}(l+1),(\hat{i},\hat{j})\in\mathscr{P}(l)} [(w_i - w_j) - (w_{\hat{i}} - w_{\hat{j}})] \\
&- \sum_{(m,n)\in\mathscr{P}(l)\setminus\{\mathbf{Q}\}} (w_m - w_n) && (3.92) \\
&\geq 0 && (3.93)
\end{aligned}$$

where every $(w_i - w_j) - (w_{\hat{i}} - w_{\hat{j}}) \geq 0$, every $w_m - w_n < 0$ and $\mathbf{Q}$ is the set of node pairs $(\hat{i}, \hat{j})$ that are used in the first summation. Hence, the function $\mathscr{K}(l) = \min_{\mathbf{r}\in\Omega,\mathscr{S}(\mathbf{r})=l} \sum_{i=1}^{N} w_i r_i$ is convex. □

4 Single-Server Multi-Message PIR with Side Information

In the information-theoretic private information retrieval problem, one user wishes to download one or multiple messages from a database, which is stored at one or multiple servers, and requires that the server(s) should not be able to infer any information about which message(s) the user wants to download. If the database is only stored at a single server (or equivalently multiple colluding servers) and no side information is available, the user has to download the whole database to achieve the information-theoretic privacy. However, it has been shown that the information-theoretic privacy can be achieved without downloading the whole database under either of the following conditions: (1) the database is stored in multiple non-colluding servers [22]; (2) the user possesses some messages as side information [49]. In this chapter, we investigate the extended case for private information retrieval with side information, which is multi-message single-server private information retrieval with side information. We establish the capacity for this problem by presenting the proof for the converse and proposing an achievability coding scheme.

4.1 Problem Statement

In the single-server multi-message private information retrieval with side information problem, it is assumed that there exists a database consisting of K messages, denoted by $X_{1:K} = \{X_1, \dots, X_K\}$. The database is only stored at a single server. The random variable of the messages, $\mathbf{X}_i$'s for all $i \in \{1, \dots, K\}$, are assumed to be independent from each other and consists of L bits, i.e.,

$$H(\mathbf{X}_1) = \cdots = H(\mathbf{X}_K) = L, \tag{4.1}$$

$$H(\mathbf{X}_1, \dots, \mathbf{X}_K) = H(\mathbf{X}_1) + \cdots + H(\mathbf{X}_K). \tag{4.2}$$

Let $W_{1:N} = \{W_1, \dots, W_N\} \subseteq \{1, \dots, K\}$ denote the set of indices of the demand messages and let $S_{1:M} = \{S_1, \dots, S_M\} \subseteq \{1, \dots, K\} \setminus W_{1:N}$ denote the set of indices of the side information messages. The user initially possesses M messages, denoted by $X_{S_{1:M}} = \{X_{S_1}, \dots, X_{S_M}\}$, as side information messages and wants to download N messages, denoted by $X_{W_{1:N}} = \{X_{W_1}, \dots, X_{W_N}\}$, from the

server. We assume that the server only knows the number (M) of side information that the user has but does not know the set of indices $S_{1:M}$ of those side information messages. For each $n \in \{1,\dots,N\}$, let $\mathbf{W}_n$ denote the random variable for demand index W_n. For each $m \in \{1,\dots,M\}$, let $\mathbf{S}_m$ denote the random variable for side information index S_m. Let $\mathbf{W}_{1:N} = \{\mathbf{W}_1,\dots,\mathbf{W}_N\}$ denote the random variable for the set of the random variables for the demand indices and $\mathbf{S}_{1:M} = \{\mathbf{S}_1,\dots,\mathbf{S}_M\}$ denote the random variable for the set of the random variables for the side information indices. We assume that $\mathbf{W}_{1:N}$ is uniformly distributed over all subsets of $\{1,\dots,K\}$ with size N, i.e.,

$$\Pr(\mathbf{W}_{1:N} = W_{1:N}) = \frac{1}{\binom{K}{N}}, \qquad \forall W_{1:N} \subseteq \{1,\dots,K\}, |W_{1:N}| = N. \tag{4.3}$$

Additionally, we assume that $\mathbf{S}_{1:M}$ is conditionally uniformly distributed over all subsets of $\{1,\dots,K\} \setminus \mathbf{W}_{1:N}$, i.e.,

$$\Pr(\mathbf{S}_{1:M} = S_{1:M} | \mathbf{W}_{1:N} = W_{1:N}) = \begin{cases} \frac{1}{\binom{K-N}{M}}, & \forall S_{1:M} \subseteq \{1,\dots,K\} \setminus W_{1:N}, |S_{1:M}| = M, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

The goal of the user is to retrieve the demand messages $X_{W_{1:N}}$ from the server and still keeps the indices $W_{1:N}$ private from the server. To achieve this goal, the user generates and sends a query to the server and the server. Let $Q^{[W_{1:N},S_{1:M}]}$ denote a query generated for demand indices $W_{1:N}$ and side information indices $S_{1:M}$ and let $\mathbf{Q}^{[W_{1:N},S_{1:M}]}$ denote the random variable for $Q^{[W_{1:N},S_{1:M}]}$. We assume that the query is generated from a (stochastic) function of the indices $W_{1:N}$ and $S_{1:M}$ and is independent of all messages, i.e.,

$$H(\mathbf{Q}^{[W_{1:N},S_{1:M}]} | \mathbf{X}_{1:K}) = H(\mathbf{Q}^{[W_{1:N},S_{1:M}]}) \tag{4.5}$$

We also use the notation Q to denote a query realization that is generated without specifying the demand side information indices. When the server receives query realization $Q^{[W_{1:N},S_{1:M}]}$, it generates and sends back the answer string $A^{[W_{1:N},S_{1:M}]}$, which is a deterministic function of $Q^{[W_{1:N},S_{1:M}]}$ and all messages. Let $\mathbf{A}^{[W_{1:N},S_{1:M}]}$ denote the random variable of $A^{[W_{1:N},S_{1:M}]}$, which should satisfy

$$H(\mathbf{A}^{[W_{1:N},S_{1:M}]} | \mathbf{Q}^{[W_{1:N},S_{1:M}]}, \mathbf{X}_{1:K}) = 0. \tag{4.6}$$

We note that given the query realization Q, the random variable of answer string, $\mathbf{A}$, is still not determined since the messages are random variables. The query $\mathbf{Q}^{[W_{1:N},S_{1:M}]}$ is from an alphabet $\mathscr{Q}$ and the answer string $\mathbf{A}^{[W_{1:N},S_{1:M}]}$ is from an alphabet $\mathscr{A}$. The PIR scheme is the set of all queries and answer strings.

Let D denote the number of download bits from the server for any coding scheme satisfying

above requirements., which can be computed as

$$\begin{aligned} D =& H(\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}|\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{W}_{1:N},\mathbf{S}_{1:M}) && (4.7)\\ =& \mathbb{E}_{(W_{1:N},S_{1:M})\sim(\mathbf{W}_{1:N},\mathbf{S}_{1:M})}\left[H(\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}|\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{W}_{1:N}=W_{1:N},\mathbf{S}_{1:M}=S_{1:M})\right] && (4.8)\\ =& H(\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}|\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{W}_{1:N}=W_{1:N},\mathbf{S}_{1:M}=S_{1:M}) && (4.9)\end{aligned}$$

where Equation (4.9) is because the number of download bits is independent of the realizations of demand and side information indices.

The rate of such coding scheme is defined as follows.

$$R=\lim_{L\to\infty}\frac{L}{D}, \tag{4.10}$$

where L is the number of bits per message. The capacity is the supreme of all achievable rates. We use $C(K.M,N)$ to denote the capacity for single-server multi-message private information retrieval with side information problem, which has K messages, M side information messages and N demand messages.

$$C(K,M,N)=\sup\lim_{L\to\infty}\frac{L}{D}. \tag{4.11}$$

4.1.1 Retrieval and Privacy Conditions

For any demand indices $W_{1:N}$ and side information indices $S_{1:M}$, any generated query $Q^{[W_{1:N},S_{1:M}]}$ and corresponding answer string $A^{[W_{1:N},S_{1:M}]}$ should permit decoding of the demand messages $X_{W_{1:N}}$ with side information messages $X_{S_{1:M}}$. Hence, the random variables of query and answer string, $\mathbf{Q}^{[W_{1:N},S_{1:M}]}$ and $\mathbf{A}^{[W_{1:N},S_{1:M}]}$, should satisfy:

$$H(\mathbf{X}_{W_{1:N}}|\mathbf{A}^{[W_{1:N},S_{1:M}]},\mathbf{Q}^{[W_{1:N},S_{1:M}]},\mathbf{X}_{S_{1:M}})=0,\forall W_{1:N}\subseteq\{1,\dots,K\},\forall S_{1:M}\subseteq\{1,\dots,K\}\setminus W_{1:N}. \tag{4.12}$$

We refer to Condition (4.12) as the *retrieval condition* for single-server multi-message PIR with side information.

Besides, private information retrieval requires that the server should not be able to infer any information about the indices of the demand messages, which requires the query to satisfy:

$$I(\mathbf{W}_{1:N};\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]})=0. \tag{4.13}$$

By using the chain rule of mutual information, we can get

$$\begin{aligned} I(\mathbf{W}_{1:N};\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{X}_{1:K}) =& I(\mathbf{W}_{1:N};\mathbf{X}_{1:K})+I(\mathbf{W}_{1:N};\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}|\mathbf{X}_{1:K})\\ &+I(\mathbf{W}_{1:N};\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}|\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{X}_{1:K}) && (4.14)\\ =& I(\mathbf{W}_{1:N};\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}). && (4.15)\end{aligned}$$

Equation (4.15) is because random variables of demand indices $\mathbf{W}_{1:N}$ are independent of the random variable of messages $\mathbf{X}_{1:K}$, i.e., $I(\mathbf{W}_{1:N};\mathbf{X}_{1:K}) = 0$, the answer string is deterministic given query and all messages, i.e., $H(\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]}|\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{X}_{1:K}) = 0$. Hence, the answer string, query and all messages must satisfy:

$$I(\mathbf{W}_{1:N};\mathbf{A}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{Q}^{[\mathbf{W}_{1:N},\mathbf{S}_{1:M}]},\mathbf{X}_{1:K}) = 0. \tag{4.16}$$

We refer to Condition (4.16) as the *privacy condition* for the single-server multi-message PIR with side information.

For any fixed $W_{1:N} \subseteq \{1,\dots,K\}$ and $S_{1:M} \subseteq \{1,\dots,K\}\setminus W_{1:N}$, the probability of choosing query $Q \in \mathscr{Q}$ and corresponding answer string $A \in \mathscr{A}$, denoted by $\Pr(\mathbf{A} = A, \mathbf{Q} = Q|\mathbf{W}_{1:N} = W_{1:N}, \mathbf{S}_{1:M} = S_{1:M})$, is known by the server. Hence, the server can verify *privacy condition* (4.16) for any PIR coding scheme. For each query Q_j and corresponding answer string $\mathbf{A}$ individually, we cannot determine whether it satisfies the *privacy condition* or not. But the following necessary condition can be derived from the *privacy condition*.

Definition 4.1 (Necessary Condition)**.** *The query realization Q and corresponding answer string $\mathbf{A}$ generated for K messages, M side information messages and N demand messages satisfy the necessary condition, if for any $W_{1:N'} \subseteq \{1,\dots,K\}$ with $|W_{1:N'}| = N' \le N$, there exists at least one corresponding $S_{1:M'} \subseteq \{1,\dots,K\}\setminus W_{1:N'}$ with $|S_{1:M'}| = M' \le M$ such that*

$$H(\mathbf{X}_{W_{1:N'}}|\mathbf{A},\mathbf{Q} = Q,\mathbf{X}_{S_{1:M'}}) = 0. \tag{4.17}$$

.

We note that, according to Definition 4.1, the *necessary condition* is defined for any single query and its corresponding answer string. The *privacy condition*, which is defined for the coding scheme, is closely related to the *necessary condition*.

Lemma 4.1. *For any query realization Q and corresponding answer string $\mathbf{A}$ from a single-server multi-message PIR with side information coding scheme which satisfies the privacy condition, Q and $\mathbf{A}$ also satisfy the necessary condition defined in Definition 4.1.*

Proof. We need to show that for any $W_{1:N'} \subseteq \{1,\dots,K\}$ with $|W_{1:N'}| = N' \le N$, there exists at least one corresponding $S_{1:M'} \subseteq \{1,\dots,K\}\setminus W_{1:N'}$ with $|S_{1:M'}| = M' \le M$ such that Equation (4.17) is satisfied. The proof is by contradiction. Suppose there exists $W_{1:N'} \subseteq \{1,\dots,K\}$ with $|W_{1:N'}| = N' \le N$, all $S_{1:M'} \subseteq \{1,\dots,K\}\setminus W_{1:N'}$ with $|S_{1:M'}| = M' \le M$ satisfy

$$H(\mathbf{X}_{W_{1:N'}}|\mathbf{A},\mathbf{Q} = Q,\mathbf{X}_{S_{1:M'}}) > 0. \tag{4.18}$$

Then, $\mathbf{X}_{W_{1:N'}}$ cannot be decoded from $\mathbf{A}$ given any M side information messages, which implies that $\mathbf{X}_{W_{1:N'}}$ cannot be the demand messages. This contradicts the requirement of the *privacy condition*. □

This necessary condition for single demand message cases has been mentioned in [49]. We have the following remarks for Lemma 4.1.

- Although the *privacy condition* cannot be verified for any particular query Q and corresponding answer string $\mathbf{A}$, the necessary condition in Lemma 4.1 can be verified easily by checking the existence of $S_{1:M}$ which satisfies Equation (4.18) for every $W_{1:N}$.

- For the special cases, where $N' = 1$, for any single message X_{W_1}, there always exists at least one $S_{1:M} \subseteq \{1,\dots,K\} \setminus \{W_1\}$ such that

$$H(\mathbf{X}_{W_1}|\mathbf{A},\mathbf{Q} = Q,\mathbf{X}_{S_{1:M}}) = 0. \tag{4.19}$$

- Since every $\mathbf{A}$ and $\mathbf{Q}$ can be generated for any $W_{1:N}$ with a proper $S_{1:M}$, we just use $\mathbf{A}$ and $\mathbf{Q}$ instead of $\mathbf{A}^{[W_{1:N},S_{1:M}]}$ and $\mathbf{Q}^{[W_{1:N},S_{1:N}]}$.

Example 4.1. *Consider a single-server multi-message PIR with side information problem which has parameters $K = 9$, $M = 2$, $N = 2$. Suppose there exists one query realization Q_1 and corresponding answer string $\mathbf{A}_1$ which satisfies*

$$\mathbf{A}_1 = \left\{ \begin{array}{l} \mathbf{X}_1 + \mathbf{X}_2 \\ \mathbf{X}_2 + \mathbf{X}_3 \\ \mathbf{X}_4 + \mathbf{X}_5 \\ \mathbf{X}_5 + \mathbf{X}_6 \\ \mathbf{X}_7 + \mathbf{X}_8 \\ \mathbf{X}_8 + \mathbf{X}_9 \end{array} \right\} \tag{4.20}$$

It is easy to verify that Q_1 and $\mathbf{A}_1$ satisfy the necessary condition in Lemma 4.1. Hence, there must exist a coding scheme that satisfies the privacy condition and retrieval condition. And $\mathbf{A}_1$ is one of its answer strings. The construction of coding scheme from $\mathbf{A}_1$ is present in Section 4.1.2. Consider another query realization Q_2 and corresponding answer string $\mathbf{A}_2$ which satisfies

$$\mathbf{A}_2 = \left\{ \begin{array}{l} \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 \\ \mathbf{X}_4 + \mathbf{X}_5 + \mathbf{X}_6 \\ \mathbf{X}_7 + \mathbf{X}_8 + \mathbf{X}_9 \end{array} \right\} \tag{4.21}$$

It can be verified that Q_2 and $\mathbf{A}_2$ do not satisfy the necessary condition in Lemma 4.1, since decoding $\mathbf{X}_1$ and $\mathbf{X}_4$ requires 4 messages, which are $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_5$, and $\mathbf{X}_6$. However, the number of side information messages is only 2, which implies that $\mathbf{X}_1$ and $\mathbf{X}_4$ are not the demand messages. Thus, we can conclude that $\mathbf{A}_2$ can not be one answer string from any coding scheme which satisfies both privacy condition and retrieval condition.

4.1.2 Coding Scheme based on One Answer String

Given only one answer string $\mathbf{A}$ from a single-server PIR with side information coding scheme, it is not enough to verify whether the coding scheme satisfies the *privacy condition* or not. However, if the given answer string $\mathbf{A}$ satisfies the necessary condition of privacy condition defined in Lemma 4.1, it is possible to construct a single-server PIR with side information coding scheme based on answer string $\mathbf{A}$.

Without loss of generality, we can express the answer string $\mathbf{A}$ as the function of all messages, i.e.,

$$\mathbf{A} = f(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K) \tag{4.22}$$

The coding scheme for messages $\mathbf{X}'_1, \ldots, \mathbf{X}'_K$ can be constructed by the following steps.

1. The user randomly generates a mapping from the demand messages $\mathbf{X}'_{W_{1:N}}$ to messages $\mathbf{X}_{W'_{1:N}}$ such that

$$\mathbf{X}_{W'_{1:N}} = \mathbf{X}'_{W_{1:N}}, \tag{4.23}$$

   with the same probability $\forall\, W'_{1:N} \subset \{1, \ldots, K\}$ and $|W'_{1:N}| = N$:

$$\Pr(\mathbf{W}'_{1:N} = W'_{1:N}) = \frac{1}{\binom{K}{N}}. \tag{4.24}$$

2. Let $S'_{1:M}$ denote the indices of messages which are required to decode $\mathbf{X}_{W'_{1:N}}$ from $\mathbf{A}$. The user generates the mapping from the side information messages $\mathbf{X}'_{S_{1:M}}$ to them such that

$$\mathbf{X}_{S'_{1:M}} = \mathbf{X}'_{S_{1:M}}. \tag{4.25}$$

3. The user randomly maps the messages $\mathbf{X}'_{1:K} \setminus \mathbf{X}'_{W_{1:N} \cup S_{1:M}}$ to $\mathbf{X}_{1:K} \setminus \mathbf{X}_{W'_{1:N} \cup S'_{1:M}}$.

$$\mathbf{X}'_{1:K} \setminus \mathbf{X}'_{W_{1:N} \cup S_{1:M}} = \mathbf{X}_{1:K} \setminus \mathbf{X}_{W'_{1:N} \cup S'_{1:M}}. \tag{4.26}$$

4. The query Q' for answer string $\mathbf{A}'$ can be generated by mapping the messages $\mathbf{X}_{1:K}$ to $\mathbf{X}'_{1:K}$ according to previous steps,

$$\mathbf{A}' = f(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K), \tag{4.27}$$

   where $\mathbf{X}_{W'_{1:N}} = \mathbf{X}'_{W_{1:N}}$, $\mathbf{X}_{S'_{1:M}} = \mathbf{X}'_{S_{1:M}}$ and $\mathbf{X}'_{1:K} \setminus \mathbf{X}'_{W_{1:N} \cup S_{1:M}} = \mathbf{X}_{1:K} \setminus \mathbf{X}_{W'_{1:N} \cup S'_{1:M}}$.

Lemma 4.2. *The constructed coding scheme satisfies both privacy condition and retrieval condition.*

Proof. According to Step 2, messages $\mathbf{X}_{W'_{1:N}}$ can be decoded from $\mathbf{A}$ given $\mathbf{X}_{S'_{1:M}}$ as side infor-

mation, i.e.,

$$H(\mathbf{X}_{W'_{1:N}}|\mathbf{A},\mathbf{X}_{S'_{1:M}})=0. \tag{4.28}$$

Replacing all $\mathbf{X}_{1:K}$ with $\mathbf{X}'_{1:K}$ according to the selected random mapping, we can get

$$H(\mathbf{X}'_{W_{1:N}}|\mathbf{A}',\mathbf{X}'_{S_{1:M}})=0. \tag{4.29}$$

Hence, messages $\mathbf{X}'_{W_{1:N}}$ can be decoded from $\mathbf{A}'$ given $\mathbf{X}'_{S_{1:M}}$ as side information, which satisfies the *retrieval condition*.

According to Step 1, the probability for any N messages $\mathbf{X}'_{W_{1:N}}$ to be the demand messages is the same as the probability of choosing $W'_{1:N}$ to be the mapping indices for $W_{1:N}$, which is the same for any subset of indices in $\{1,\dots,K\}$ of size N. Therefore, the *privacy condition* is also satisfied. □

Definition 4.2 (Valid Answer String)**.** *For any answer string which satisfies the necessary condition defined in Definition 4.1, we call it a valid answer string, since based on it we can generate a PIR coding scheme which satisfies both privacy condition and retrieval condition.*

Example 4.2. *Let us revisit the Example 4.1 and show how to design a PIR coding scheme based on answer string* $\mathbf{A}_1$*. Suppose we want to construct the query* Q_3 *for demand messages* $\mathbf{X}'_1$ *and* $\mathbf{X}'_5$ *with side information messages* $\mathbf{X}'_4$ *and* $\mathbf{X}'_6$*. We first randomly map* $\mathbf{X}'_1$ *and* $\mathbf{X}'_5$ *to any two messages with equal probability. Suppose we choose* $\mathbf{X}_3$ *and* $\mathbf{X}_7$ *and generate the mapping*

$$\mathbf{X}_3=\mathbf{X}'_1,\mathbf{X}_7=\mathbf{X}'_5 \tag{4.30}$$

To decode $\mathbf{X}_3$ *and* $\mathbf{X}_7$ *from* $\mathbf{A}_1$*, we need to know one message of* $\{\mathbf{X}_1,\mathbf{X}_2\}$ *and one message of* $\{\mathbf{X}_8,\mathbf{X}_9\}$*. Suppose we choose* $\mathbf{X}_2$ *and* $\mathbf{X}_8$ *and generate the mapping*

$$\mathbf{X}_2=\mathbf{X}'_4,\mathbf{X}_8=\mathbf{X}'_6 \tag{4.31}$$

And we randomly map the other messages as follows

$$\mathbf{X}_1=\mathbf{X}'_2,\mathbf{X}_4=\mathbf{X}'_3,\mathbf{X}_5=\mathbf{X}'_7,\mathbf{X}_6=\mathbf{X}'_8,\mathbf{X}_9=\mathbf{X}'_9 \tag{4.32}$$

Then the answer string $\mathbf{A}_3$ *for query* Q_3 *can be expressed as*

$$\mathbf{A}_3=f(\mathbf{X}'_2,\mathbf{X}'_4,\mathbf{X}'_1,\mathbf{X}'_3,\mathbf{X}'_7,\mathbf{X}'_8,\mathbf{X}'_5,\mathbf{X}'_6,\mathbf{X}'_9)=\left\{\begin{matrix}\mathbf{X}'_2+\mathbf{X}'_4\\ \mathbf{X}'_4+\mathbf{X}'_1\\ \mathbf{X}'_3+\mathbf{X}'_7\\ \mathbf{X}'_7+\mathbf{X}'_8\\ \mathbf{X}'_5+\mathbf{X}'_6\\ \mathbf{X}'_6+\mathbf{X}'_9\end{matrix}\right\}. \tag{4.33}$$

It can be verified that given the side information messages $\mathbf{X}'_4$ and $\mathbf{X}'_6$, the demand messages $\mathbf{X}'_1$ and $\mathbf{X}'_5$ can be successfully decoded from $\mathbf{A}_3$. We note that besides the demand messages, $\mathbf{X}'_2$ and $\mathbf{X}'_9$ can also be decoded as by-product. Intuitively, we do not want many of these non-demand messages to be decodable since each of them cost L extra download bits. However, we will show that it is inevitable in multi-message private information retrieval with side information.

4.1.3 Conditional Answer String

Given the query realization Q, the answer string $\mathbf{A}$ is a deterministic function of all messages $\mathbf{X}_{1:K}$, which can be expressed as

$$\mathbf{A} = f_Q(\mathbf{X}_1, \ldots, \mathbf{X}_K) \tag{4.34}$$

We note that when the messages $\mathbf{X}_1, \ldots, \mathbf{X}_K$ as the input of the function f_Q are random variables, the answer string $\mathbf{A}$ as the output should also be a random variable.

Definition 4.3. *(Conditional Answer String) For each answer string $\mathbf{A}$ and any subset of messages $\mathbf{X}_{\mathcal{K}}$, $\forall \mathcal{K} \subseteq \{1, \ldots, K\}$, define the conditional answer string of $\mathbf{A}$ given $\mathbf{X}_{\mathcal{K}}$ as*

$$\mathbf{A} \| \mathbf{X}_{\mathcal{K}} = f_Q(\mathbf{X}_1, \ldots, \mathbf{X}_K | \mathbf{X}_i = c, \forall i \in \mathcal{K}), \tag{4.35}$$

where c is any constant value that the messages can take.

Remark 4.1. *Setting the random variables of messages into constant value in the function can remove the randomness of those random variables. Since the constant value is known to both user and server, there is no difference between using one constant value or different constant values for the conditional messages.*

Lemma 4.3. *For any answer string $\mathbf{A}$ that satisfies the necessary condition, i.e., $\forall W_{1:N'} \subseteq \{1, \ldots, K\}$ with $|W_{1:N'}| = N' \leq N$, there exists at least one $S_{1:M'} \subseteq \{1, \ldots, K\} \setminus W_{1:N'}$ with $|S_{1:M'}| = M' \leq M$ such that*

$$H(\mathbf{X}_{W_{1:N'}} | \mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}}) = 0, \tag{4.36}$$

for any subset of messages $\mathbf{X}_{\mathcal{K}} \subset \mathbf{X}_{1:K}$, the corresponding conditional answer string $\mathbf{A} \| \mathbf{X}_{\mathcal{K}}$ also satisfies the necessary condition for the messages $\mathbf{X}_{1:K} \setminus \mathbf{X}_{\mathcal{K}}$, i.e., $\forall W_{1:N'} \subseteq \{1, \ldots, K\} \setminus \mathcal{K}$ with $|W_{1:N'}| = N' \leq N$, there exists at least one $S_{1:M'} \subseteq \{1, \ldots, K\} \setminus (W_{1:N'} \cup \mathcal{K})$ with $|S_{1:M'}| = M' \leq M$ such that

$$H(\mathbf{X}_{W_{1:N'}} | \mathbf{A} \| \mathbf{X}_{\mathcal{K}}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}}) = 0. \tag{4.37}$$

Proof. Since the answer string $\mathbf{A}$ satisfies the necessary condition in Lemma 4.1, then for any $W_{1:N'} \subseteq \{1, \ldots, K\}$ with $N' \leq N$, there must exist at least one $S_{1:M'} \subseteq \{1, \ldots, K\} \setminus W_{1:N'}$ with $M' \leq M$ such that $H(\mathbf{X}_{W_{1:N'}} | \mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}}) = 0$. Then for any $W_{1:N'} \subseteq \{1, \ldots, K\} \setminus \mathcal{K}$, there are the following two cases.

1. If $S_{1:M'} \cap \mathcal{K} = \emptyset$, then there must exist $S_{1:M'} \subseteq \{1,\dots,K\} \setminus (W_{1:N'} \cup \mathcal{K})$ such that

$$H(\mathbf{X}_{1:N'}|\mathbf{A}\|\mathbf{X}_{\mathcal{K}}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}}) \leq H(\mathbf{X}_{1:N'}|\mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}}) \tag{4.38}$$
$$= 0 \tag{4.39}$$

2. If $S_{1:M'} \cap \mathcal{K} = \tau \neq \emptyset$, then there must exist $(S_{1:M'} \setminus \tau) \subseteq \{1,\dots,K\} \setminus (W_{1:N'} \cup \mathcal{K})$ such that

$$H(\mathbf{X}_{1:N'}|\mathbf{A}\|\mathbf{X}_{\mathcal{K}}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}\setminus\tau}) = H(\mathbf{X}_{1:N'}|\mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}\setminus\tau}, \mathbf{X}_{\mathcal{K}}) \tag{4.40}$$
$$\leq H(\mathbf{X}_{1:N'}|\mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_{S_{1:M'}}) \tag{4.41}$$
$$= 0 \tag{4.42}$$

In both case, for any $W_{1:N'}$, there always exists a subset of indices $S^* \subseteq \{1,\dots,K\} \setminus (W_{1:N'} \cup \mathcal{K})$, which can be $S_{1:M'}$ or $S_{1:M'} \setminus \tau$, such that given $\mathbf{X}_{S^*}$, $\mathbf{X}_{W_{1:N'}}$ can be decoded from $\mathbf{A}\|\mathbf{X}_{\mathcal{K}}$, which satisfies the necessary condition in Lemma 4.1. □

Corollary 4.1. *For any valid answer string* $\mathbf{A}$ *for messages* $\mathbf{X}_{1:K}$ *and* $\forall \mathcal{K} \subset \{1,\dots,K\}$*, the conditional answer string* $\mathbf{A}\|\mathbf{X}_{\mathcal{K}}$ *is also a valid answer string for messages* $\mathbf{X}_{1:K} \setminus \mathbf{X}_{\mathcal{K}}$*.*

Proof. According to Lemma 4.3, the conditional answer string $\mathbf{A}\|\mathbf{X}_{\mathcal{K}}$ satisfies the necessary condition for messages $\mathbf{X}_{1:K} \setminus \mathbf{X}_{\mathcal{K}}$. Hence it is a valid answer string by Definition 4.2. □

Example 4.3. *For answer string* $\mathbf{A}_1$ *defined in Equation* (4.20)*. The conditional answer string* $\mathbf{A}\|\mathbf{X}_{1,2,3}$ *can be expressed as follows:*

$$\mathbf{A}\|\mathbf{X}_{1,2,3} = \left\{ \begin{array}{l} c + c \\ c + c \\ \mathbf{X}_4 + \mathbf{X}_5 \\ \mathbf{X}_5 + \mathbf{X}_6 \\ \mathbf{X}_7 + \mathbf{X}_8 \\ \mathbf{X}_8 + \mathbf{X}_9 \end{array} \right\}, \tag{4.43}$$

where c is any constant value that the messages may take. We note that we can also assign different constant values to $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$*, respectively. It can be verified that* $\mathbf{A}\|\mathbf{X}_{1,2,3}$ *satisfies the necessary condition for* $\mathbf{X}_4, \dots, \mathbf{X}_9$ *and hence it is a valid answer string for* $\mathbf{X}_4, \dots, \mathbf{X}_9$*.*

4.2 The Capacity

Theorem 4.1. *For the single-server multi-message private information retrieval with side information problem with K messages, M side information messages and N demand messages,*

the capacity is

$$C(K,M,N) = \left(K - M - (T^* - N - 1)^+ \lfloor \frac{M}{N} \rfloor - \Theta\right)^{-1}, \tag{4.44}$$

where

$$T^* = \left\lfloor \frac{K - (N^2 + M + 1)}{N + \lfloor \frac{M}{N} \rfloor} \right\rfloor + N + 1, \tag{4.45}$$

$$\Theta = \left(K - N^2 - N - M - (T^* - N - 1)^+ (N + \lfloor \frac{M}{N} \rfloor)\right)^+. \tag{4.46}$$

Proof. The proofs for the converse and achievability are presented in Section 4.2.1 and Section 4.2.2, respectively. □

We have the following remarks regarding the formula (4.44).

- It can be verified that the capacity is lower-bounded by $(K-M)^{-1}$. This is not surprising since the MDS coding scheme which downloads $(K-M)L$ bits from the server always satisfies both *retrieval condition* and *privacy condition*.
- When $N > M$, which can be interpreted as the number of demand messages is larger than the number of side information messages, the capacity is $(K-M)^{-1}$.
- When $N^2 + N \geq K - M$, which can be interpreted as the number of demand messages is larger than the square root of the number of total messages (except the side information messages), the capacity is $(K-M)^{-1}$.
- When $N = 1$, the multi-message problem degrades into the single-message problem. In such cases, $T^* = \left\lfloor \frac{K-(1+M+1)}{1+M} \right\rfloor + 1 + 1 = \lfloor \frac{K-1}{1+M} \rfloor + 1$. Since $M \geq 0$, it is easy to see that T^* can also be expressed as $T^* = \lceil \frac{K}{1+M} \rceil$. Since $N = 1$, it can be verified that θ is always positive. The capacity for special cases when $N = 1$ can be shown to be $C(K,M,1) = \lceil \frac{K}{1+M} \rceil^{-1}$, which matches the capacity for single-server single-message PIR in [49].

4.2.1 Converse

In this section, we present the proof for the converse of Theorem 4.1. We need to show that the rate of any coding scheme which satisfies both *retrieval condition* and *privacy condition* is lower-bounded by

$$R \geq K - M - (T^* - N - 1)^+ \lfloor \frac{M}{N} \rfloor - \Theta. \tag{4.47}$$

For any query realization Q and corresponding answer string $\mathbf{A}$, the total number of download

bits can be expressed as

$$D = H(\mathbf{A}|\mathbf{Q} = Q) \tag{4.48}$$
$$= H(\mathbf{X}_{1:K}, \mathbf{A}|\mathbf{Q} = Q) - H(\mathbf{X}_{1:K}|\mathbf{A}, \mathbf{Q} = Q) \tag{4.49}$$
$$= H(\mathbf{X}_{1:K}|\mathbf{Q} = Q) + H(\mathbf{A}|\mathbf{Q} = Q, \mathbf{X}_{1:K}) - H(\mathbf{X}_{1:K}|\mathbf{A}, \mathbf{Q} = Q) \tag{4.50}$$
$$= KL - H(\mathbf{X}_{1:K}|\mathbf{A}, \mathbf{Q} = Q), \tag{4.51}$$

where Equation (4.51) is because that query $\mathbf{Q} = Q$ is assumed to be independent of the messages, i.e., $H(\mathbf{X}_{1:K}|\mathbf{Q} = Q) = H(\mathbf{X}_{1:K}) = KL$, and according to Equation (4.6), the answer string $\mathbf{A}$ is deterministic given query $\mathbf{Q} = Q$ and all messages $\mathbf{X}_{1:K}$.

According to Lemma 4.1, for any single index $W \in \{1, \ldots, K\}$, there exist at least one $S \subseteq \{1, \ldots, K\} \setminus \{W\}$ such that $H(\mathbf{X}_W|\mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_S) = 0$. Without loss of generality, we can make the following assumptions.

1. Decoding $\mathbf{X}_{W^{(j)}}$ with $W^j \in \{1, \ldots, K\}$ from $\mathbf{A}$ requires at least $|S^{(j)}|$ messages, denoted by $\mathbf{X}_{S^{(j)}}$, as side information.

2. $S^{(1)}$ has the smallest size over all $S^{(j)}$'s, i.e.,

$$S^{(1)} \in \arg\min_{S^{(j)}} |S^{(j)}| \tag{4.52}$$

We note that there may be multiple subsets that have the minimum size. Let $U^{(1)}$ denote the set of indices such that $\mathbf{X}_{U^{(1)}}$ can also be decoded given $\mathbf{X}_{S^{(1)}}$ as side information. Let $Z_1 = W^{(1)} \cup U^{(1)} \cup S^{(1)}$ denote the union set of indices[1]. Equation (4.51) can be further expanded as follows.

$$D = KL - H(\mathbf{X}_{Z_1}|\mathbf{A}, \mathbf{Q} = Q) - H(\mathbf{X}_{1:K}|\mathbf{A}, \mathbf{Q} = Q, \mathbf{X}_{Z_1}). \tag{4.53}$$

We note that according to the assumption for $U^{(1)}$, none of the messages $\mathbf{X}_{1:K} \setminus \mathbf{X}_{Z_1}$ can be decoded given $\mathbf{X}_{Z_1}$.

Similarly, we can construct $Z_i = W^{(i)} \cup U^{(i)} \cup S^{(i)}$ such that given $\mathbf{X}_{Z_1^{i-1}}$, decoding $X_{W^{(i)}}$ requires the smallest number of side information $\mathbf{X}_{S^{(i)}}$, and $\mathbf{X}_{U^{(i)}}$ can be decoded at the same time.

[1]We note that $W^{(1)}$ is a single index instead of a subset of indices. For the ease of notation we still use $W^{(1)} \cup U^{(1)} \cup S^{(1)}$ to denote the union set $\{W^{(1)}\} \cup U^{(1)} \cup S^{(1)}$.

Suppose after T iterations, we have $Z_1^T = \{1,\dots,K\}$. Then, we have

$$\begin{aligned} D =& KL - H(\mathbf{X}_{Z_1}|\mathbf{A},\mathbf{Q}=Q) - H(\mathbf{X}_{Z_2}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1}) - H(\mathbf{X}_{1:K}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^2}) && (4.54)\\ =& KL - H(\mathbf{X}_{Z_1}|\mathbf{A},\mathbf{Q}=Q) - H(\mathbf{X}_{Z_2}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1}) \\ & - \cdots - H(\mathbf{X}_{Z_T}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) - H(\mathbf{X}_{1:K}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^T}) && (4.55)\\ =& KL - \sum_{i=1}^{T} H(\mathbf{X}_{Z_i}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{i-1}}), && (4.56)\end{aligned}$$

where Equation (4.56) is because the last term $H(\mathbf{X}_{1:K}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^T})=0$ due to the assumption $Z_1^T=\{1,\dots,K\}$. To get the lower-bound for D, we need to maximize the summation in Equation (4.56).

For the last group of messages $\mathbf{X}_{Z_T}$, by assumption, given all messages of previous groups $\mathbf{X}_{Z_1^{T-1}}$, from the conditional answer string $\mathbf{A}\|\mathbf{X}_{Z_1^{T-1}}$, no message can be decoded given less than $|S^{(T)}|$ messages as side information messages and $\mathbf{X}_{W^{(T)}\cup U^{(T)}}$ can be decoded given messages $\mathbf{X}_{S^{(T)}}$, i.e.,

$$H(\mathbf{X}_{W^{(T)}\cup U^{(T)}}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}},\mathbf{X}_{S^{(T)}})=0. \tag{4.57}$$

Hence, we have

$$\begin{aligned} H(\mathbf{X}_{Z_T}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) =& H(\mathbf{X}_{S^{(T)}}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) + H(\mathbf{X}_{W^{(T)}\cup U^{(T)}}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}},\mathbf{X}_{S^{(T)}}) && (4.58)\\ \le& H(\mathbf{X}_{S^{(T)}}) && (4.59)\end{aligned}$$

and

$$\begin{aligned} H(\mathbf{A}\|\mathbf{X}_{Z_1^{T-1}}|\mathbf{Q}=Q) =& H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) && (4.60)\\ =& H(\mathbf{A},\mathbf{X}_{Z_T}|\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) - H(\mathbf{X}_{Z_T}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) && (4.61)\\ =& H(\mathbf{X}_{Z_T}|\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) + H(\mathbf{A}|\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}},\mathbf{X}_{Z_T}) \\ & - H(\mathbf{X}_{Z_T}|\mathbf{A},\mathbf{Q}=Q,\mathbf{X}_{Z_1^{T-1}}) && (4.62)\\ \ge& H(\mathbf{X}_{Z_T}) - H(\mathbf{X}_{S^{(T)}}) && (4.63)\\ =& H(\mathbf{X}_{W^{(T)}\cup U^{(T)}}) && (4.64)\end{aligned}$$

Hence, the conditional answer string $\mathbf{A}\|\mathbf{X}_{Z_1^{T-1}}$ has at least $|W^{(T)}\cup U^{(T)}|L$ bits, which is enough for messages $\mathbf{X}_{Z_T}$ such that given any $|S^{(T)}|$ messages, the other messages can be decoded from $\mathbf{A}\|\mathbf{X}_{Z_1^{T-1}}$.

By assumption, none of the messages in $\mathbf{X}_{Z_{T-1}}$ can be decoded from $\mathbf{A}\|\mathbf{X}_{Z_{T-2}}$, while $\mathbf{X}_{W^{(T-1)}\cup U^{(T-1)}}$ can be decoded from $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}}$ given $\mathbf{X}_{S^{(T-1)}}$. We note that no message from $\mathbf{X}_{Z_T}$ is required to decode $\mathbf{X}_{W^{(T-1)}\cup U^{(T-1)}}$ from $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}}$, even though $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}}$ is also a function of $\mathbf{X}_{Z_T}$. Thus it must be possible to divide $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}}$ into two parts. One part is a function of only $\mathbf{X}_{Z_{T-1}}$, denoted by $\mathbf{A}(\mathbf{X}_{Z_{T-1}})$, which permits the decoding of $\mathbf{X}_{W^{(T-1)}\cup U^{(T-1)}}$ with $\mathbf{X}_{S^{(T-1)}}$ as side information.

The other part is a function of both $\mathbf{X}_{Z_{T-1}}$ and $\mathbf{X}_{Z_T}$, denoted by $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})$, which satisfies $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T}) \| \mathbf{X}_{Z_{T-1}} = \mathbf{A} \| \mathbf{X}_{Z_1^{T-1}}$.

Lemma 4.4. *Giving any $|S^{(T)}|$ messages from $\mathbf{X}_{Z_T}$, without loss of optimality, we can assume that no information about $\mathbf{X}_{Z_{T-1}}$ can be inferred from $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})$, i.e., $\forall S \subset Z_T$ with $|S| \leq |S^{(T)}|$:*

$$H(\mathbf{X}_{Z_{T-1}} | \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T}), \mathbf{X}_S) = H(\mathbf{X}_{Z_{T-1}}). \tag{4.65}$$

Proof. By assumption, $\mathbf{A} \| \mathbf{X}_{Z_1^{T-1}} = \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T}) \| \mathbf{X}_{Z_{T-1}}$ satisfies two properties: (i) no message of $\mathbf{X}_{Z_T}$ can be decoded given less than $|S^{(T)}|$ messages and (ii) all messages of $\mathbf{X}_{Z_T}$ can be decoded given any $|S^{(T)}|$ messages.

If any function of $\mathbf{X}_{Z_{T-1}}$ (e.g. $f(\mathbf{X}_{Z_{T-1}})$) can be decoded from $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})$ with $|S| \leq |S^{(T)}|$, then in $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})$, there must exist an answer string $\tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S)$ which is only a function of $f(\mathbf{X}_{Z_{T-1}})$ and $\mathbf{X}_S$. From $\tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S) \| \mathbf{X}_{Z_{T-1}}$, there must exist one message in $\mathbf{X}_S$ that can be fully decoded or partially decoded by given only $|S| - 1 \leq |S^{(T)}| - 1$ messages.

For the first case where one message can be fully decoded, it contradicts property (i) that no message can be fully decoded by given less than $|S^{(T)}|$ messages. For the second case where one message (e.g. $\mathbf{X}_l$) can be partially decoded, according to property (ii), given any $|S^{(T)}|$ messages including $\mathbf{X}_S$ as side information messages, all other messages in $\mathbf{X}_{Z_T}$ can be decoded from $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T}) \| \mathbf{X}_{Z_{T-1}}$. Note that given $\mathbf{X}_S$ and $\mathbf{X}_{Z_{T-1}}$, $\tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S)$ is a constant. Hence, decoding the other $|W^{(T)} \cup U^{(T)}|$ messages cannot use $\tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S)$, which implies $(\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T}) \setminus \tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S)) \| \mathbf{X}_{Z_{T-1}}$ has $(|W^{(T)} \cup U^{(T)}|)L$ bits. By assumption $Z_T = W^{(T)} \cup U^{(T)} \cup S^{(T)}$, it is easy to see that an answer string with $(|W^{(T)} \cup U^{(T)}|)L$ bits is enough to satisfy property (ii) (e.g. MDS style coded answer string). Hence, $\tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S)$ can be replaced by $\tilde{\mathbf{A}}(f(\mathbf{X}_{Z_{T-1}}), \mathbf{X}_S) \| \mathbf{X}_S$ which satisfies (4.65), while the privacy and retrieval conditions are preserved and the number of download bits does not increase. □

Lemma 4.5. *For decoding any two messages, one in $\mathbf{X}_{Z_{T-1}}$ and another in $\mathbf{X}_{Z_T}$, from $\mathbf{A} \| \mathbf{X}_{Z_1^{T-2}}$, the number of required side information messages is $|S^{(T-1)}| + |S^{(T)}|$.*

Proof. By assumption, decoding any one message in $\mathbf{X}_{Z_T}$ from $\mathbf{A} \| \mathbf{X}_{Z_1^{T-1}}$ requires $|S^{(T)}|$ side information messages from $\mathbf{X}_{Z_T}$. According to Lemma 4.4, no information of $\mathbf{X}_{Z_{T-1}}$ can be inferred from $\mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})$ given any $|S^{(T)}|$ messages from $\mathbf{X}_{Z_T}$. Hence, the demand message in $\mathbf{X}_{Z_{T-1}}$ can only be decoded from $\mathbf{A}(\mathbf{X}_{Z_{T-1}})$, which requires $|S^{(T-1)}|$ side information messages. Therefore, decoding any two messages, one in $\mathbf{X}_{Z_{T-1}}$ and another in $\mathbf{X}_{Z_T}$, from $\mathbf{A} \| \mathbf{X}_{Z_1^{T-2}}$ requires $|S^{(T-1)}| + |S^{(T)}|$ side information messages. □

According to the assumption, given $\mathbf{X}_{S^{(T)}}$ and $\mathbf{X}_{S^{(T-1)}}$, messages $\mathbf{X}_{W^{(T)} \cup U^{(T)}}$ and $\mathbf{X}_{W^{(T-1)} \cup U^{(T-1)}}$ can be decoded from $\mathbf{A} \| \mathbf{X}_{Z_1^{T-2}}$, respectively. Thus, even we only want to decode 2 messages, the other messages will be decoded as a by-product.

Lemma 4.6. *Without loss of optimality, the conditional answer string* $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}} = \{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\}$ *can be replaced by* $\{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\|\mathbf{X}_{Z_{T-1}}\}$.

Proof. For any message $\mathbf{X}_W \in \mathbf{X}_{Z_{T-1}^T}$, if it can be decoded from $\{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\}$ with side information set S, it can also be decoded from $\{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\|\mathbf{X}_{Z_{T-1}}\}$ with the same S, since

$$H(\mathbf{X}_W|\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\|\mathbf{X}_{Z_{T-1}}, \mathbf{X}_S) \leq H(\mathbf{X}_W|\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T}), \mathbf{X}_S) = 0. \tag{4.66}$$

For any two messages in either $\mathbf{X}_{Z_{T-1}}$ or $\mathbf{X}_{Z_T}$, the minimum number of required side information for decoding them is $|S^{(T-1)}|$ or $|S^{(T)}|$, respectively, for both answer strings.

For any two messages, one in $\mathbf{X}_{Z_{T-1}}$ and another in $\mathbf{X}_{Z_T}$, according to Lemma 4.5, the minimum number of required side information is $|S^{(T-1)}| + |S^{(T)}|$ for $\{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\}$. And it can be easily verified that the minimum number of required side information for $\{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\}$ is also $|S^{(T-1)}| + |S^{(T)}|$.

Further, when decoding any two messages from both groups, all other messages from both groups can also be decoded. Hence, if any subset of messages from $\mathbf{X}_{Z_{T-1}^T}$ can be decoded from the original conditional answer string $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}} = \{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\}$ with side information $\mathbf{X}_S$, they can also be decoded from the modified conditional answer string $\{\mathbf{A}(\mathbf{X}_{Z_{T-1}}), \mathbf{A}(\mathbf{X}_{Z_{T-1}}, \mathbf{X}_{Z_T})\|\mathbf{X}_{Z_{T-1}}\}$ with the same side information. Therefore, the new answer string also satisfies the privacy condition and requires no more download bits than the old answer string. □

According to Lemma 4.6, the answer string $\mathbf{A}\|\mathbf{X}_{Z_1^{T-2}}$ can be replaced by an answer string which can be completely separated into two functions, each of which is only the function of either $\mathbf{X}_{Z_{T-1}}$ or $\mathbf{X}_{Z_T}$. Similarly, we can apply the same processes for all previous groups of messages iteratively and show that $\mathbf{A}\|\mathbf{X}_1^{T-i}$ can be replaced by answer string which can be completely separated into i functions and each function only depends on messages from one group. Finally, the answer string $\mathbf{A}$ can be replaced by a new answer string $\hat{\mathbf{A}}$ such that

(1) $\hat{\mathbf{A}}$ is a valid answer string from coding schemes that satisfy the privacy condition and downloads the same number of bits as $\mathbf{A}$.

(2) $\hat{\mathbf{A}}$ can be fully separated into T functions $\hat{\mathbf{A}}(\mathbf{X}_{Z_1}), \ldots, \hat{\mathbf{A}}(\mathbf{X}_{Z_T})$, where $\mathbf{A}(\mathbf{X}_{Z_t})$ for $t \in \{1, \ldots, T\}$ only depends on messages in $\mathbf{X}_{Z_t}$. $Z_i \cap Z_j = \emptyset$ for any $i \neq j \in \{1, \ldots, T\}$.

(3) From each $\hat{\mathbf{A}}(\mathbf{X}_{Z_t})$ for $t \in \{1, \ldots, T\}$, $\mathbf{X}_{W^{(t)} \cup U^{(t)}}$ can be decoded given $\mathbf{X}_{S^{(t)}}$ as side information and no message in $\mathbf{X}_{Z_t}$ can be decoded given less than $|S^{(t)}|$ side information messages ($|S^{(t)}| \geq 1$).

Lemma 4.7. *For any valid answer string* $\hat{\mathbf{A}}$ *which satisfies the necessary condition, for any* $t \in \{1, \ldots, T\}$, *we have* $|W^{(t)} \cup U^{(t)}| \geq N$.

Proof. For any $j \in S^{(t)}$, given $\mathbf{X}_{S^{(t)} \setminus j}$, none of $\mathbf{X}_{W^{(t)} \cup U^{(t)}}$ can be decoded. Hence, if there exists any t such that $|W^{(t)} \cup U^{(t)}| < N$, then $\mathbf{W} = W^{(t)} \cup U^{(t)} \cup j$ cannot be decoded from $\hat{\mathbf{A}}$, which violates the *necessary condition*. Therefore, for any valid $\hat{\mathbf{A}}$ satisfying the *necessary condition*, we have $|W^{(t)} \cup U^{(t)}| \geq N$. □

For any query realization Q and corresponding answer string $\mathbf{A}$, the number of download bits can be expressed as

$$
\begin{aligned}
D =& H(\mathbf{A}|\mathbf{Q} = Q) && (4.67)\\
=& H(\hat{\mathbf{A}}|\hat{\mathbf{Q}} = \hat{Q}) && (4.68)\\
=& KL - \sum_{i=1}^{T} H(\mathbf{X}_{Z_i}|\hat{\mathbf{A}}, \hat{\mathbf{Q}} = \hat{Q}, \mathbf{X}_{Z_1^{i-1}}) && (4.69)\\
=& KL - \sum_{i=1}^{T} H(\mathbf{X}_{Z_i}|\hat{\mathbf{A}}(\mathbf{X}_{Z_1}), \ldots, \hat{\mathbf{A}}(\mathbf{X}_{Z_T}), \hat{\mathbf{Q}} = \hat{Q}, \mathbf{X}_{Z_1^{i-1}}) && (4.70)\\
=& KL - \sum_{i=1}^{T} H(\mathbf{X}_{Z_i}|\hat{\mathbf{A}}(\mathbf{X}_{Z_i}), \hat{\mathbf{Q}} = \hat{Q}) && (4.71)\\
=& KL - \sum_{i=1}^{T} \left[H(\mathbf{X}_{S^{(i)}}|\hat{\mathbf{A}}(\mathbf{X}_{Z_i}), \hat{\mathbf{Q}} = \hat{Q}) + H(\mathbf{X}_{W^{(i)} \cup U^{(i)}}|\hat{\mathbf{A}}(\mathbf{X}_{Z_i}), \hat{\mathbf{Q}} = \hat{Q}, \mathbf{X}_{S^{(i)}}) \right] && (4.72)\\
\geq& KL - \sum_{i=1}^{T} H(\mathbf{X}_{S^{(i)}}) && (4.73)\\
=& KL - L \sum_{i=1}^{T} |S^{(i)}| && (4.74)
\end{aligned}
$$

Equation (4.71) is because only $\hat{\mathbf{A}}(\mathbf{X}_{Z_i})$ depends on $\mathbf{X}_{Z_i}$ and for any $j \neq i$, $\hat{\mathbf{A}}(\mathbf{X}_{Z_j})$ is independent of $\mathbf{X}_{Z_i}$. Equation (4.73) is because given $\mathbf{X}_{S^{(i)}}$ as side information, messages $\mathbf{X}_{W^{(i)} \cup U^{(i)}}$ can be decoded from $\hat{\mathbf{A}}(\mathbf{X}_{Z_i})$, i.e., $H(\mathbf{X}_{W^{(i)} \cup U^{(i)}}|\hat{\mathbf{A}}(\mathbf{X}_{Z_i}), \hat{\mathbf{Q}} = \hat{Q}, \mathbf{X}_{S^{(i)}}) = 0$.

According to the necessary condition of privacy condition shown in Lemma 4.1, for any $W_{1:N'} \subseteq \{1, \ldots, K\}$ with $N' \leq N$, there must exist $S_{1:M'} \subseteq \{1, \ldots, K\} \setminus W_{1:N'}$ with $M' \leq M$ such that $H(\mathbf{X}_{W_{1:N'}}|\hat{\mathbf{A}}, \hat{\mathbf{Q}} = \hat{Q}, \mathbf{X}_{S_{1:M'}}) = 0$. Consider $W_{1:N'} = \{W^{(1)}, \ldots, W^{(N')}\}$ with $N' \leq N$ and $N' \leq T$. To successfully decode $\mathbf{X}_{W_{1:N'}}$ from $\hat{\mathbf{A}}$, the following condition must be satisfied

$$
\sum_{t=1}^{N'} |S^{(t)}| \leq M. \tag{4.75}
$$

By assumption, we have $Z_1^T = \{1, \ldots, K\}$. The number of groups, T, must satisfy

$$K = \sum_{i=1}^{T} |Z_i| = \sum_{i=1}^{T} |W^{(i)}| + |U^{(i)}| + |S^{(i)}| \tag{4.76}$$

$$\geq \sum_{i=1}^{T} (N + |S^{(i)}|) \tag{4.77}$$

$$\geq TN + \sum_{i=1}^{T} |S^{(i)}|. \tag{4.78}$$

Equation (4.77) is because $|W^{(i)}| + |U^{(i)}| = |W^{(i)} \cup U^{(i)}| \geq N$ according to Lemma 4.7.

Depending on the relationship between the three parameters of the problem, which are K, M and N, we have the following three cases.

1. If $K \leq N^2 + M$, equivalently $T \leq N$ and $\sum_{i=1}^{T} |S^{(i)}| \leq M$. From Equation (4.74), we have

$$D \geq KL - L \sum_{i=1}^{T} |S^{(i)}| \geq (K - M)L. \tag{4.79}$$

2. If $N^2 + M < K \leq N^2 + M + N$, the number of groups, T, can be $N + 1$. The number of messages in the first N groups can be computed as follows.

$$|Z_i^N| = \sum_{i=1}^{N} |W^{(i)}| + |U^{(i)}| + |S^{(i)}| \tag{4.80}$$

$$\geq N^2 + \sum_{i=1}^{N} |S^{(i)}| \tag{4.81}$$

Additionally, for the $(N+1)$-th group, according to Lemma 4.7, we have $|W^{(N+1)}| + |U^{(N+1)}| \geq N$. Thus the number of messages in $S^{(N+1)}$ can be computed as follows

$$|S^{(N+1)}| = K - |Z_1^N| - |W^{(N+1)}| - |U^{(N+1)}| \tag{4.82}$$

$$\leq N^2 + M + N - (N^2 + \sum_{i=1}^{N} |S^{(i)}|) - N \tag{4.83}$$

$$\leq M - \sum_{i=1}^{N} |S^{(i)}| \tag{4.84}$$

Hence, for all $T = N + 1$ groups, we have

$$\sum_{i=1}^{N+1} |S^{(i)}| \leq M. \tag{4.85}$$

The total number of download bits can be lower-bounded by

$$D \geq KL - L\sum_{i=1}^{T} |S^{(i)}| \tag{4.86}$$

$$\geq (K-M)L. \tag{4.87}$$

3. If $K \geq N^2+M+N+1$, the number of groups, T, can be more than N. Since the sum of any N of $|S^{(i)}|$'s is upper bounded by M, to maximize the sum for all $|S^{(i)}|$'s, for $i \in \{1,\ldots,T\}$, the optimal choice for $S^{(i)}$'s would be that $N - \lfloor \frac{M}{N} \rfloor N$ of $|S^{(i)}|$'s are upper bounded by $\lfloor \frac{M}{N} \rfloor + 1$ and others are upper bounded by $\lfloor \frac{M}{N} \rfloor$. Without loss of generality, we can assume that

$$S^{(i)} \leq \begin{cases} \lfloor \frac{M}{N} \rfloor + 1, & i \in \{1,\ldots,N - \lfloor \frac{M}{N} \rfloor N\} \\ \lfloor \frac{M}{N} \rfloor, & i \in \{N - \lfloor \frac{M}{N} \rfloor N + 1,\ldots,T\}. \end{cases} \tag{4.88}$$

If $N > M$, then $\lfloor \frac{M}{N} \rfloor = 0$, which gives

$$|S^{(i)}| \leq \begin{cases} 1, & i \in \{1,\ldots,M\} \\ 0, & i \in \{M+1,\ldots,T\}. \end{cases} \tag{4.89}$$

Note that if we plug-in $\lfloor \frac{M}{N} \rfloor = 0$ into Equation (4.88), we can get $|S^{(i)}| \leq 1$ for all $i \in \{1,\ldots,N\}$. But since the total number of side information is M, the number of non-zero $|S^{(i)}|$'s can only be at most M. In such cases, the total number of download bits satisfies $D \geq (K-M)L$.

If $N \leq M$, equivalently $\lfloor \frac{M}{N} \rfloor \geq 1$, then the number of download bits satisfies

$$D \geq KL - L\sum_{i=1}^{T} |S^{(i)}| \tag{4.90}$$

$$= KL - L\sum_{i=1}^{N} |S^{(i)}| - L\sum_{i=N+1}^{T-1} |S^{(i)}| - L|S^{(T)}| \tag{4.91}$$

$$\geq KL - L(M - (T-N-1)\lfloor \frac{M}{N} \rfloor - |S^{(T)}|). \tag{4.92}$$

In such cases, the numbers of messages in Z_1^{T-1} satisfy

$$|Z_1^N| \geq N^2 + M, \tag{4.93}$$

$$|Z_{N+1}^{T-1}| \geq (T-N-1)(N + \lfloor \frac{M}{N} \rfloor). \tag{4.94}$$

For the last group, if $|Z_T| \leq N$, then we have $|S^{(T)}| = 0$. If $|Z_T| > N$, then $|S^{(T)}| = |Z_T| -$

$|W^{(T)} \cup U^T|$. Hence, we have

$$K = |Z_1^N| + |Z_{N+1}^{T-1}| + |Z_T| \tag{4.95}$$

$$\geq N^2 + M + (T - N - 1)(N + \lfloor \frac{M}{N} \rfloor) + |Z_T|. \tag{4.96}$$

And T can be upper bounded as follows.

$$T \leq \frac{K - (N^2 + M + |Z_T|)}{N + \lfloor \frac{M}{N} \rfloor} + N + 1 \tag{4.97}$$

$$\leq \frac{K - (N^2 + M + 1)}{N + \lfloor \frac{M}{N} \rfloor} + N + 1. \tag{4.98}$$

As T can only be an integer, the maximal value of T can be expressed as

$$T^* = \left\lfloor \frac{K - (N^2 + M + 1)}{N + \lfloor \frac{M}{N} \rfloor} \right\rfloor + N + 1, \tag{4.99}$$

which is the same as Equation (4.45). When $T = T^*$, the number of the message in the last group satisfies

$$|Z_T| \leq K - N^2 - M - (T^* - N - 1)(N + \lfloor \frac{M}{N} \rfloor). \tag{4.100}$$

Then the number of side information in the last group ($S^{(T)} = S^{(T^*)}$) can be upper bounded by

$$|S^{(T^*)}| = \left(K - N^2 - M - (T^* - N - 1)(N + \lfloor \frac{M}{N} \rfloor) - N \right)^+, \tag{4.101}$$

which is the same as Θ defined in Equation (4.46). And the total number of download bits can be lower bounded by

$$D \geq (K - M - (T^* - N - 1)\lfloor \frac{M}{N} \rfloor - |S^{(T^*)}|)L. \tag{4.102}$$

For all three cases, the normalized number of download bits can be lower bounded the same expression

$$R = \frac{D}{L} \geq K - M - (T^* - N - 1)^+ \lfloor \frac{M}{N} \rfloor - \Theta. \tag{4.103}$$

where T^* and Θ are defined by Equation (4.45) and (4.46), respectively.

4.2.2 Achievability

In this section, we prove the achievability of Theorem 4.1 by presenting a coding scheme for the single-server multi-message private information retrieval with side information problem which downloads the minimum normalized number of bits.

For any single-server multi-message PIR with side information problem with K total messages, M side information messages and N demand messages, we first compute T^* defined by Equation (4.45) and Θ defined by Eqn. (4.46). If $T^* = N+1$ and $\Theta = 0$, then $R^* = K - M$. It is trivial that the optimal coding scheme is the MDS coding scheme which satisfies given any M side information all of the other $K-M$ messages can be decoded. If $T^* > N+1$ or $T^* = N+1$, $\Theta > 0$, the coding scheme can be constructed as follows:

Step 1: The user creates T^* subsets, denoted by $\{\wp_1, \ldots, \wp_{T^*}\}$, which will be populated by messages. For all $i \in \{1, \ldots, T^*\}$, the size of $\wp_i$ satisfies:

$$|\wp_i| = \begin{cases} \lfloor \frac{M}{N} \rfloor + N + 1, & 1 \le i \le t \\ \lfloor \frac{M}{N} \rfloor + N, & t+1 \le i \le T^* - 1 \\ \Theta + N, & i = T^* \end{cases} \tag{4.104}$$

where $t = M - N\lfloor \frac{M}{N} \rfloor$. Let $c_i \in \mathbb{N}$ for $i \in \{1, \ldots, T^*\}$ denote the number of demand messages in subset $\wp_i$ and is initialized to be 0.

Step 2: For the first demand message $\mathbf{X}_{W_1}$, the user randomly selects one subset $\wp_i$ ($i \in \{1, \ldots, T^*\}$) to contain it with probability $\frac{|\wp_i|}{K}$, i.e.,

$$\Pr(\mathbf{X}_{W_1} \in \wp_i) = \frac{|\wp_i|}{K}. \tag{4.105}$$

The user updates $c_i = c_i + 1$. Then for the j-th demand message $\mathbf{X}_{W_j}$ ($j \in \{1, \ldots, N\}$), the user randomly selects one subset $\wp_u$ ($u \in \{1, \ldots, T^*\}$) to contain it with probability $\frac{|\wp_u| - c_u}{K - j + 1}$.

$$\Pr(\mathbf{X}_{W_j} \in \wp_u) = \frac{|\wp_u| - c_u}{K - j + 1}. \tag{4.106}$$

Iteratively, the user places all demand messages into the subsets.

Step 3: For each subset $\wp_i$ with $c_i > 0$, the user randomly selects m_i side information messages to put into $\wp_i$, where m_i satisfies:

$$m_i = \begin{cases} \lfloor \frac{M}{N} \rfloor + 1, & 1 \le i \le t \\ \lfloor \frac{M}{N} \rfloor, & t+1 \le i \le T^* - 1 \\ \Theta, & i = T^* \end{cases} \tag{4.107}$$

Step 4: The user randomly distributes the other messages to fill up the remaining empty spaces in each subset.

Step 5: The user sends queries to the server according to the coding scheme which satisfies the MDS-Condition in each subset of messages $\mathbf{X}_{\wp_i}$ ($\forall i \in \{1,\ldots,T^*\}$) with $R(|\wp_i|, m_i) = (|\wp_i| - m_i)L$ bits.

We name the coding scheme constructed by this method as Partition-and-MDS-Coding scheme, which is a modification of an optimal coding scheme for single demand message proposed in [42]. The way we select subsets for demand messages is related to the urn problem. The probability of any N messages to be the demand messages follows the binomial distribution.

Theorem 4.2. *The Partition-and-MDS-Coding scheme satisfies the Retrieval Condition and the Privacy Condition.*

Proof. For each subset of message $X_{\wp_i}$, if it contains demand messages, the number of download bits $R(|\wp_i|, m_i)$ and the number of side information messages m_i in such subset of messages satisfy $R(|\wp_i|, m_i) + m_i L = |\wp_i|L$. Additionally, the Partition-and-MDS-Coding scheme satisfies MDS-Condition in every subset of messages. Thus, all missing messages in $\wp_i$ can be successfully decoded, including the demand messages. Therefore, the Retrieval Condition is satisfied.

The probability that any N messages ($\{\mathbf{X}_{W_1},\ldots,\mathbf{X}_{W_N}\}$) are the demand messages can be computed as

$$\begin{aligned}
\Pr(\mathbf{W}_{1:N} = \{W_1,\ldots,W_N\}) =& N!\Pr(\mathbf{W}_1^N = W_1^N) && (4.108)\\
=& N!\Pr(\mathbf{W}_1 = W_1)\Pr(\mathbf{W}_2^N = W_2^N|\mathbf{W}_1 = W_1) && (4.109)\\
=& N!\prod_{i=1}^{N}\Pr(\mathbf{W}_i = W_i|\mathbf{W}_1^{i-1} = W_1^{i-1}) && (4.110)
\end{aligned}$$

According to the construction of the Partition-and-MDS-Coding scheme and assume that $W_i \in \wp_j$, we have

$$\begin{aligned}
&\Pr(\mathbf{W}_i = W_i|\mathbf{W}_1^{i-1} = W_1^{i-1})\\
=&\Pr(\mathbf{W}_i \in \wp_j|\mathbf{W}_1^{i-1} = W_1^{i-1}) \times \Pr(\mathbf{W}_i = W_i|\mathbf{W}_i \in \wp_j, \mathbf{W}_1^{i-1} = W_1^{i-1}) && (4.111)\\
=&\frac{|\wp_j| - |\wp_j \setminus (\wp_j \cap \{W_1,\ldots,W_{i-1}\})|}{K-i+1} \times \frac{1}{|\wp_j| - |\wp_j \setminus (\wp_j \cap \{W_1,\ldots,W_{i-1}\})|} && (4.112)\\
=&\frac{1}{K-i+1} && (4.113)
\end{aligned}$$

Hence, we have

$$\Pr(\mathbf{W}_{1:N} = \{W_1, \ldots, W_N\}) = N! \prod_{i=1}^{N} \frac{1}{K-i+1} \tag{4.114}$$

$$= \frac{N!}{K(K-1)\cdots(K-N+1)} \tag{4.115}$$

$$= \frac{1}{\binom{K}{N}} \tag{4.116}$$

Since there are $\binom{K}{N}$ possible demand message pairs with size N, every N-message pair is equally likely to be the demand messages, which satisfies the Privacy Condition of multi-message PIR. □

Example 4.4. *Consider a single-server multi-message private information retrieval with side information problem with the following setup:* $K = 13$, $M = 5$, $N = 2$, $W_{1,2} = \{2,5\}$ *and* $S_{1:5} = \{1,4,6,7,9\}$. *The coding scheme can be constructed by using the Partition-and-MDS-Coding method. In this example, we have* $T^* = 3 > N$ *and* $\Theta = 2 > 0$.

- *Step 1: The user first creates three subsets* ($\wp_1$, $\wp_2$, $\wp_3$) *with size* $|\wp_1| = 5$, $|\wp_2| = 4$ *and* $|\wp_3| = 4$.
- *Step 2: The user randomly selects one subset from* $\{\wp_1, \wp_2, \wp_3\}$ *to contain the first demand message* $\mathbf{X}_2$ *with probability* $\frac{5}{13}$, $\frac{4}{13}$ *and* $\frac{4}{13}$, *respectively. Suppose* $\wp_1$ *is chosen. Then for the second demand message* $\mathbf{X}_5$, *the user randomly selects one subset from* $\{\wp_1, \wp_2, \wp_3\}$ *with probability* $\frac{4}{12}$, $\frac{4}{12}$ *and* $\frac{4}{12}$, *respectively. Suppose* $\wp_3$ *is chosen.*
- *Step 3: For* $\wp_1$ *and* $\wp_3$, *the subsets which are chosen to contain demand messages, the user randomly distributes* 3 *and* 2 *side information messages into them, respectively. Suppose* $\mathbf{X}_1, \mathbf{X}_4, \mathbf{X}_6$ *are placed in* $\wp_1$ *and* $\mathbf{X}_7, \mathbf{X}_9$ *are placed in* $\wp_2$.
- *Step 4: The user randomly distributes the remaining messages into the subsets. Suppose we get* $\wp_1 = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4, \mathbf{X}_6, \mathbf{X}_8\}$, $\wp_2 = \{\mathbf{X}_3, \mathbf{X}_{10}, \mathbf{X}_{11}, \mathbf{X}_{13}\}$ *and* $\wp_3 = \{\mathbf{X}_5, \mathbf{X}_7, \mathbf{X}_9, \mathbf{X}_{12}\}$.
- *Step 5: The user generates and sends query* Q *to request the following answer strings* $\mathbf{A}$, *which consists of the MDS-coded messages from each subsets. For example, the first two coded messages are linear combinations of messages in subset* $\wp_1$ *and given any three messages, the other two messages can be decoded.*

$$\mathbf{A} = \begin{Bmatrix} \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_4 + \mathbf{X}_6 + \mathbf{X}_8 \\ \mathbf{X}_1 + 2\mathbf{X}_2 + 3\mathbf{X}_4 + 4\mathbf{X}_6 + 5\mathbf{X}_8 \\ \mathbf{X}_3 + \mathbf{X}_{10} + \mathbf{X}_{11} + \mathbf{X}_{13} \\ \mathbf{X}_3 + 2\mathbf{X}_{10} + 3\mathbf{X}_{11} + 4\mathbf{X}_{13} \\ \mathbf{X}_5 + \mathbf{X}_7 + \mathbf{X}_9 + \mathbf{X}_{12} \\ \mathbf{X}_5 + 2\mathbf{X}_7 + 3\mathbf{X}_9 + 4\mathbf{X}_{12} \end{Bmatrix} \tag{4.117}$$

From the user's perspective: X_2 can be decoded from T_1 and T_2 given that X_1, X_4 and X_6 are side information. X_5 can be decoded from T_5 and T_6 given that X_7 and X_9 are side information. Hence, the Retrieval Condition is satisfied.

From the server's perspective, the probability for any two messages to be the demand message is the same, which is $\frac{1}{\binom{13}{2}} = \frac{1}{78}$. Thus, the server cannot infer any information about the indices of the demand messages.

4.3 Numerical Examples

In this section, we present the numerical examples for single-server multi-message private information retrieval with side information.

Example 4.5. *We give the numerical simulation for single-server multi-message private information retrieval with side information for fixed $K = 100$ and $M \in \{1,\dots,10\}$ and $N \in \{1,\dots,9\}$. The black line in the plot is the normalized download bits for a trivial MDS coding scheme*

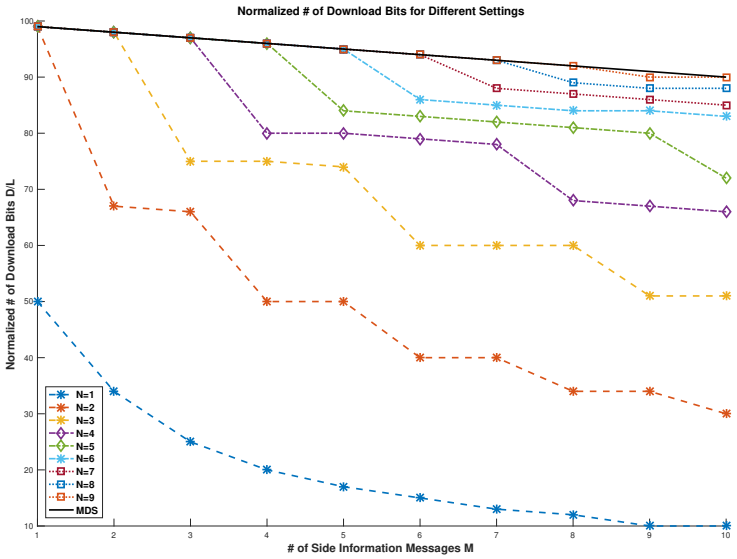

Figure 4.1 – Normalized number of download bits for fixed $K = 100$, $N \in \{1,\dots,9\}$, and $M \in \{1,\dots,10\}$.

which requires $K - M$ download bits per demand bit and be treated as the known upper bound for normalized download bits. It can be seen from the plot that when $N > M$, the optimal coding scheme (our partition-and-MDS coding scheme) has the same performance as the trivial MDS coding scheme. However, when $N \leq M$, our partition-and-MDS-coding scheme download much fewer bits than the trivial MDS coding scheme. Some numerical effect happens for $N \geq 2$ due to $\lfloor \frac{M}{N} \rfloor$. For example, when $N = 2$, from $M = 4$ to $M = 5$, the normalized number of download bits keeps the same, while from $M = 5$ to $M = 6$, the normalized number of download bits decreases

about 10.

Example 4.6. *We give another numerical simulation for single-server multi-message private information retrieval with side information for fixed* $N = 4$*, different* $M \in \{1,\dots,12\}$ *and different* $K \in \{N+M,\dots,100\}$*. As we can see from the plot, when* $M = 1,2,3$*, which are smaller than*

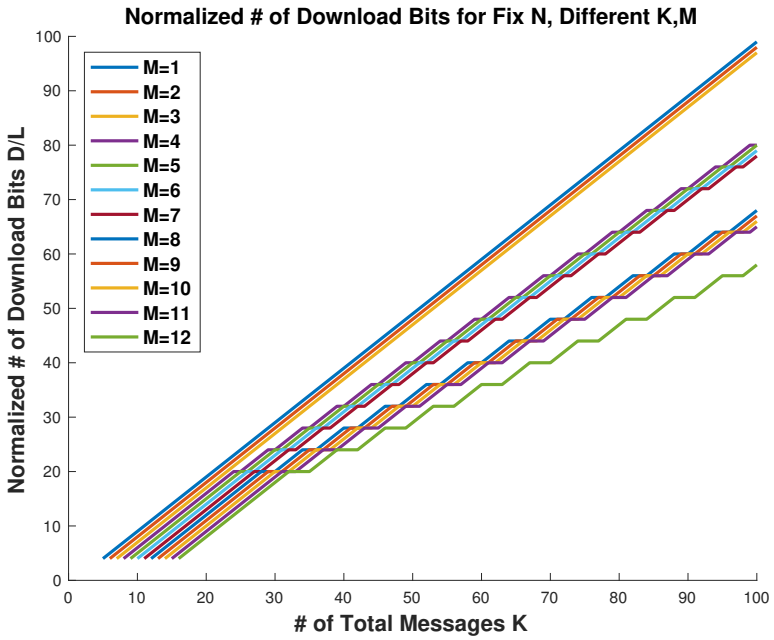

Figure 4.2 – Normalized number of download bits for fixed $N = 4$, $M \in \{1,\dots,12\}$, and $K \in \{N+M,\dots,100\}$.

$N = 4$, the normalized number of download bits per demand bit ($\frac{D}{L}$) grows linearly as K goes large, which is consistent with the theoretical result since in such cases $D = (K-M)L$. And when $M > N$, $\frac{D}{L}$ increases as K goes large, but with a smaller ratio and the staircase effect can be seen. Sometimes, when K increases by a small amount, the normalized number of download bits $\frac{D}{L}$ remains the same. We also plot the ration between D/L and $K-M$ in the following figure. It can be seen that for fixed N as K becomes large, to achieve the privacy, the percentage of missing bits we need to download decreases. And for larger M, the percentage of missing bits we need to download is even smaller. The reason that each curve becomes up and down as K goes large is that the normalized number of download bits may not increase when K increases. When D increases as K becomes large, the ratio $\frac{D}{L(K-M)}$ also increases. When D remains the same as K becomes large, the ration $\frac{D}{L(K-M)}$ decreases.

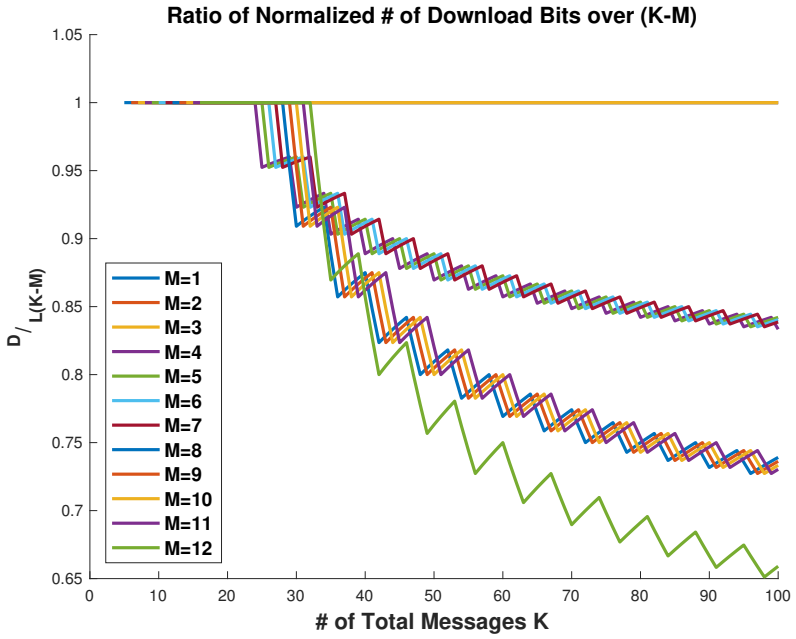

Figure 4.3 – Percentage of normalized number of download bits ($\frac{D}{(K-M)L}$) for fixed $N = 4$, $M \in \{1,\ldots,12\}$, and $K \in \{M+N,\ldots,100\}$.

4.4 Discussion and Conclusion

4.4.1 Privacy Condition in Single-Server PIR with Side Information

For the single-server private information retrieval with side information problem, the *privacy condition* can be used to derive the *necessary condition* defined in Definition 4.1. The *privacy condition* for single-server PIR is a statistic property for all the queries and answer strings, while the *necessary condition* is a property for each query and answer string. The *necessary condition* is much easier to be verified and can be verified for any single query and answer string. In the single-server PIR problem, it is sufficient to use one answer string which satisfies the *necessary condition* to represent the PIR coding scheme. The other answer strings in such coding scheme can be generated by randomly permuting the messages. Additionally, based on this *necessary condition*, the converse for the capacity can be derived.

4.4.2 Conclusion

In this chapter, we studied the single-server multi-message private information retrieval with side information problem. We established the capacity of this problem and gave a closed-form expression for the capacity. We presented the proof of the converse bound for the normalized total number of download bits and proposed an achievability scheme, Partition-and-MDS coding scheme, to construct optimal codes that satisfy both the *retrieval condition* and the *privacy condition*. The proposed achievability scheme is a linear coding scheme, which implies

that linear coding schemes are sufficient to optimally solve the multi-server single-message private information retrieval with side information problem.

In this problem, the *privacy condition* requires that each subset of messages with size N must have equal probability to be the demand messages from the server's perspective. Hence, for each subset of messages with size N, they must be decodable from the answer string given no more than M side information messages, which is referred to as the *necessary condition* in this chapter.

5 Multi-Server Single-Message PIR with Side Information

In Chapter 4, we study the multi-message single-server private information retrieval with side information, which is an extension of the single-server single-message private information retrieval with side information [49]. In this chapter, we investigate the multi-server single-message private information retrieval with side information, which can be interpreted as the extension of either multi-server single-message private information retrieval without side information [22] or single-server single-message private information retrieval with side information [49]. We prove the capacity for this problem by presenting the proof for the converse and proposing an achievability coding scheme.

5.1 Problem Statement

In the multi-server single-message PIR with Side Information problem, there is a database that consists of K messages, denoted by $X_{1:K} = \{X_1, \ldots, X_K\}$. The database is repeatively stored at N non-colluding servers without any coding. The random variables of the messages, $\mathbf{X}_i$'s for $i \in \{1, \ldots, K\}$, are assumed to be independent from each other and consists of L bits, *i.e.,*

$$H(\mathbf{X}_1) = \cdots = H(\mathbf{X}_K) = L, \tag{5.1}$$

$$H(\mathbf{X}_1, \ldots, \mathbf{X}_K) = H(\mathbf{X}_1) + \cdots + H(\mathbf{X}_K) = KL. \tag{5.2}$$

Let $W \in \{1, \ldots, K\}$ denote the demand index and $S \subseteq \{1, \ldots, K\} \setminus \{W\}$ denote the set of side information indices. The user wants to download message X_W from the servers, which is referred to as the demand message, and initially has M messages X_S with $|S| = M$, which is referred to as the side information messages. We assume that the servers only know the number (M) of side information messages that the user has but do not know the set of indices (S) of those side information messages. Suppose the demand message is uniformly chosen from all K messages and the M side information messages are uniformly chosen from the other $K-1$ messages. Let $\mathbf{W}$ denote the random variable for W, which is uniformly distributed

over $\{1,\dots,K\}$,

$$\Pr(\mathbf{W}=W)=\frac{1}{K},\qquad\qquad \forall W\in\{1,\dots,K\}, \tag{5.3}$$

Let $\mathbf{S}$ denote the random variable for S with $|S|=M$, which is uniformly distributed over $\{1,\dots,K\}\setminus\{W\}$ given $\mathbf{W}=W$,

$$\Pr(\mathbf{S}=S|\mathbf{W}=W)=\begin{cases}\frac{1}{\binom{K-1}{M}}, & S\subseteq\{1,\dots,K\}\setminus\{W\},\\ 0, & \text{otherwise.}\end{cases} \tag{5.4}$$

The discussion about the distribution models for $\mathbf{W}$ and $\mathbf{S}$ is given in Section 5.4.1.

The goal of the user is to download the demand message X_W from the servers and reveal no information about W to any of the servers. To achieve this goal, the user generates and sends queries to the servers. Let $\mathbf{Q}_j^{[W,S]}$ ($j\in[N]$) denote the random variable for the query which is generated for downloading message X_W while having side information X_S and sent to the j-th server. Following the literature, we assume that $\mathbf{Q}_j^{[W,S]}$ is a (stochastic) function of the indices W and S, but does not depend on contents of any of the messages, i.e.,

$$H(\mathbf{Q}_j^{[W,S]}|\mathbf{X}_{1:K})=H(\mathbf{Q}_j^{[W,S]}). \tag{5.5}$$

Let $Q_j^{[W,S]}$ denote the realization of $\mathbf{Q}_j^{[W,S]}$. Note that the notation $Q_j^{[W,S]}$ should not be interpreted as a function of W and S. The superscript is added for giving the additionally information that this query realization is generated for demand index W and side information indices S. We use Q_j without superscript to denote the query realization for server j when the demand index and side information indices are not specified. Once the j-th server receives the query $Q_j^{[W,S]}$, it returns corresponding answer string $A_j^{[W,S]}$ to the user. Let $\mathbf{A}_j^{[W,S]}$ denote the random variable for $A_j^{[W,S]}$. The answer string $\mathbf{A}_j^{[W,S]}$ is results of a deterministic function of the query $\mathbf{Q}_j^{[W,S]}$ and messages $\mathbf{X}_{1:K}$, i.e.,

$$H(\mathbf{A}_j^{[W,S]}|\mathbf{Q}_j^{[W,S]},\mathbf{X}_{1:K})=0,\forall\, j\in\{1,\dots,N\}. \tag{5.6}$$

The query $\mathbf{Q}_j^{[W,S]}$ is from an alphabet $\mathscr{Q}$ and the answer string $\mathbf{A}_j^{[W,S]}$ is from a corresponding alphabet $\mathscr{A}$. The PIR scheme is the set of queries and answer strings.

Let D denote the total number of bits downloaded from the servers for any coding scheme satisfying the above requirements

$$\begin{aligned}D&=H(\mathbf{A}_{1:N}^{[\mathbf{W},\mathbf{S}]}|\mathbf{W},\mathbf{S}) && (5.7)\\ &=\sum_{W\in\{1,\dots,K\}}\Pr(\mathbf{W}=W)\sum_{S\subseteq\{1,\dots,K\}\setminus\{W\},|S|=M}\Pr(\mathbf{S}=S|\mathbf{W}=W)H(\mathbf{A}_{1:N}^{[\mathbf{W},\mathbf{S}]}|\mathbf{W}=W,\mathbf{S}=S) && (5.8)\\ &=H(\mathbf{A}_{1:N}^{[\mathbf{W},\mathbf{S}]}|\mathbf{W}=W,\mathbf{S}=S). && (5.9)\end{aligned}$$

where Equation (5.9) is because the number of download bits is independent of the demand index and side information indices, i.e., $H(\mathbf{A}_{1:N}^{[\mathbf{W},\mathbf{S}]}|\mathbf{W}=W,\mathbf{S}=S)=H(\mathbf{A}_{1:N}^{[\mathbf{W},\mathbf{S}]}|\mathbf{W}=W',\mathbf{S}=S')$ for any $|S|=|S'|=M$. The rate of such coding scheme is defined as follows.

$$R=\lim_{L\to\infty}\frac{L}{D}. \tag{5.10}$$

The capacity is the supremum of all achievable rates. We use $C(K,M,N)$ to denote the capacity for MSPIR-SI problem with K messages, M side information messages and N servers.

$$C(K,M,N)=\sup\lim_{L\to\infty}\frac{L}{D} \tag{5.11}$$

5.1.1 Retrieval and Privacy Conditions

For any given W,S, let $Q_{1:N}^{[W,S]}=Q_1^{[W,S]},\ldots,Q_N^{[W,S]}$ and $A_{1:N}^{[W,S]}=A_1^{[W,S]},\ldots,A_N^{[W,S]}$ denote the queries and answer strings generated by some coding scheme for retrieving X_W with X_S as side information. To successfully decode the demand message X_W from the answer strings, the answer strings and queries must satisfy:

$$H(\mathbf{X}_W|\mathbf{A}_{1:N}^{[W,S]},\mathbf{Q}_{1:N}^{[W,S]},\mathbf{X}_S)=0,\forall W\in\{1,\ldots,K\},\forall S\subseteq\{1,\ldots,K\}\setminus\{W\}. \tag{5.12}$$

We refer to Condition (5.12) as the *retrieval condition* for multi-server single-message PIR with side information.

Additionally, private information retrieval requires that none of the servers individually should be able to infer any information about the index of the demand message. Hence, the queries must satisfy:

$$I(\mathbf{W};\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]})=0,\forall j\in\{1,\ldots,N\}. \tag{5.13}$$

According to the chain rule of mutual information, we can obtain

$$\begin{aligned}
I(\mathbf{W};\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{X}_{1:K}) &= I(\mathbf{W};\mathbf{X}_{1:K})+I(\mathbf{W};\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]}|\mathbf{X}_{1:K})+I(\mathbf{W};\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{X}_{1:K}) && (5.14)\\
&= I(\mathbf{W};\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]}|\mathbf{X}_{1:K})+I(\mathbf{W};\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{X}_{1:K}) && (5.15)\\
&= I(\mathbf{W};\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]}|\mathbf{X}_{1:K}) && (5.16)\\
&= I(\mathbf{W};\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]}). && (5.17)
\end{aligned}$$

Equation (5.15) is because $\mathbf{W}$ is independent of the messages, i.e., $I(\mathbf{W};\mathbf{X}_{1:K})=0$. Equation (5.16) is because $\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]}$ is deterministic given $\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]}$ and $\mathbf{X}_{1:K}$. i.e., $I(\mathbf{W};\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{X}_{1:K})=0$. Equation (5.17) is because $\mathbf{W}$ and $\mathbf{Q}_j^{[W,S]}$ are both independent of the messages. Thus, the answer strings, queries and messages must satisfies:

$$I(\mathbf{W};\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]},\mathbf{X}_{1:K})=0,\forall j\in\{1,\ldots,N\}. \tag{5.18}$$

We refer to Condition (5.18) as the *privacy condition* for multi-server single-message PIR with side information.

It can be shown that the *privacy condition* (5.18) is equivalent to the requirement that $\forall\, j \in \{1,\dots,N\}$, we have

$$H(\mathbf{W}) = H(\mathbf{W}|\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]}, \mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]}, \mathbf{X}_{1:K}). \tag{5.19}$$

Since the random variable of the demand index, $\mathbf{W}$, is assumed to be uniformly distributed over all indices $\{1,\dots,K\}$, $H(\mathbf{W})$ has the maximum entropy. For any specific query realization Q_j received by server j, messages $X_{1:K}$ and corresponding answer string A_j, we have

$$H(\mathbf{W}|\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]} = A_j, \mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]} = Q_j, \mathbf{X}_{1:K} = X_{1:K}) \leq H(\mathbf{W}). \tag{5.20}$$

Thus, to satisfy Equation (5.19), we must have $\forall\, j \in \{1,\dots,N\}$ and $\forall\, W \in \{1,\dots,K\}$:

$$\Pr(\mathbf{W} = W|\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]} = A_j, \mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]} = Q_j, \mathbf{X}_{1:K} = X_{1:K}) = \Pr(\mathbf{W} = W) = \frac{1}{K}. \tag{5.21}$$

Note that given $\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]} = Q_j$ and all messages $\mathbf{X}_{1:K} = X_{1:K}$, the answer string $\mathbf{A}_j^{[\mathbf{W},\mathbf{S}]}$ is deterministic. Thus, for any specific valid query $Q_j \in \mathcal{Q}$ and its corresponding answer strings $A_j \in \mathcal{A}$, server j should not be able to infer any information about the realization of the demand index $\mathbf{W}$. In particular, this observation implies the following lemma.

Lemma 5.1. *For any query realization $Q_j \in \mathcal{Q}$ and any two indices $W, W' \in \{1,\dots,K\}$, there exist $S_j \subseteq \{1,\dots,K\} \setminus \{W\}$ with $|S_j| \leq M$ and $S'_j \subseteq \{1,\dots,K\} \setminus \{W'\}$ with $|S'_j| \leq M$ such that $\forall\, \mathcal{K} \subseteq \{1,\dots,K\}$:*

$$H(\mathbf{A}_j^{[W,S_j]}|\mathbf{Q}_j^{[W,S_j]} = Q_j, \mathbf{X}_{\mathcal{K}}) = H(\mathbf{A}_j^{[W',S'_j]}|\mathbf{Q}_j^{[W',S'_j]} = Q_j, \mathbf{X}_{\mathcal{K}}). \tag{5.22}$$

Proof. We prove this lemma by contradiction. For any query realization $Q_j \in \mathcal{Q}$, server j generates the corresponding answer string A_j, which is the result of a deterministic function of Q_j and all messages. For the remainder of the proof, we denote it as

$$A_j = f_{Q_j}(X_1,\dots,X_K). \tag{5.23}$$

According to the *privacy condition* and Equation (5.21), for any index $W \in \{1,\dots,K\}$, we have

$$\Pr(\mathbf{W} = W|\mathbf{A}_j = A_j, \mathbf{Q}_j = Q_j, \mathbf{X}_{1:K} = X_{1:K}) = \frac{1}{K}. \tag{5.24}$$

Now, suppose that for some message W', there *do not* exist side information indices $S'_j \subseteq \{1,\dots,K\} \setminus \{W'\}$ for which the resulting query could be $\mathbf{Q}_j^{[W',S'_j]} = Q_j$ (with positive probability). Clearly, then, W' cannot satisfy Eqn. (5.24). Hence, for any two indices W and W', there must

exist $S_j \subseteq \{1,\dots,K\}\setminus\{W\}$ and $S'_j \subseteq [K]\setminus\{W'\}$ such that with positive probability

$$\mathbf{Q}_j^{[W,S_j]} = Q_j, \qquad \mathbf{A}_j^{[W,S_j]} = A_j = f_{Q_j}(X_1,\dots,X_K), \tag{5.25}$$

$$\mathbf{Q}_j^{[W',S'_j]} = Q_j, \qquad \mathbf{A}_j^{[W',S'_j]} = A_j = f_{Q_j}(X_1,\dots,X_K). \tag{5.26}$$

Since $\mathbf{A}_j^{[W,S_j]}$ and $\mathbf{A}_j^{[W',S'_j]}$ are the same function f_{Q_j} of the same random variables $\mathbf{X}_{1:K} = X_{1:K}$, we have

$$H(\mathbf{A}_j^{[W,S_j]}|\mathbf{Q}_j^{[W,S_j]} = Q_j, \mathbf{X}_{\mathcal{K}}) = H(\mathbf{A}_j^{[W',S'_j]}|\mathbf{Q}_j^{[W',S'_j]} = Q_j, \mathbf{X}_{\mathcal{K}}). \tag{5.27}$$

□

For any query realization Q_j, from the user's perspective, the query Q_j is generated for index pair (W,S). However, from the sever j's perspective, the query Q_j can be possibly generated for (W,S) or (W',S'). The server j should not be able to distinguish them. For any index $W' \in \{1,\dots,K\}$, it is always possible to find at least one corresponding $S' \subseteq \{1,\dots,K\}\setminus\{W'\}$ such that Q_j is generated for downloading $X_{W'}$ with side information messages $X_{S'}$, i.e.,

$$\Pr(\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]} = Q_j|\mathbf{W} = W', \mathbf{S} = S') > 0. \tag{5.28}$$

We refer to each S' as the virtual side information for W' in query Q_j, which is formally defined in Definition 5.1.

Definition 5.1 (Virtual Side Information)**.** *For any query realization Q_j from any valid PIR with side information coding scheme and any index $i \in \{1,\dots,K\}$, define*

$$\mathscr{E}_{Q_j}(i) = \{v : \Pr(\mathbf{Q}_j^{[\mathbf{W},\mathbf{S}]} = Q_j|\mathbf{W} = i, \mathbf{S} = v) > 0\}. \tag{5.29}$$

We refer to each $v \in \mathscr{E}_{Q_j}(i)$ as one virtual side information for index i in query Q_j. Let $V_{Q_j} = [v_1,\dots,v_K]$ denote a virtual side information vector for query Q_j, where $v_i \in \mathscr{E}_{Q_j}(i)$.

Regarding the virtual side information, we have the following remarks.

- The set $\mathscr{E}_{Q_j}(i)$ may contain more than one element. Hence, the virtual side information is not necessarily unique for fixed query Q_j and index i.
- Each virtual side information $v \in \mathscr{E}_{Q_j}(i)$ is a subset of indices. The size of each v is no larger than M, i.e., $|v| \le M$.
- The server j cannot infer which index pair (i,v_i) the query Q_j is generated for. Thus, any (i,v_i) $(\forall\, i \in \{1,\dots,K\})$ can be the demand and side information index pair (W,S).
- The virtual side information is defined for each query Q_j. For queries generated for different server Q_j and $Q_{j'}$ with $j \neq j'$, the virtual side information for the same index may be different, i.e., $\mathscr{E}_{Q_j}(i) \not\equiv \mathscr{E}_{Q_{j'}}(i)$.

This definition is the key to our development here. The important observation is that for any PIR with side information coding scheme, if we consider any possible query received by a server, then this query must be consistent with *every* message in the database. If this was not the case, then upon receiving that particular query, the server would be able to rule out a certain message as being the demand message, which in turn would contradict the privacy condition. Hence, the considered scheme could not be valid. According to Lemma **??**, there always exists at least one set of virtual side information messages for every message in any query realization from valid PIR with side information coding scheme. The virtual side information for multi-server PIR with side information is different from a similar concept defined in [49] for single-server PIR with side information and the detailed discussion is given in Section 5.4.2.

Example 5.1 (Virtual Side Information)**.** *Consider a MSPIR-SI problem with two servers ($N = 2$), five messages ($K = 5$) and two side information messages ($M = 2$). Suppose the user wants to download message $\mathbf{X}_1$ ($W = 1$) and has $\mathbf{X}_2$ and $\mathbf{X}_3$ as side information ($S = \{2,3\}$). Additionally, suppose each message can be divided into 4 chunks, i.e,. $\mathbf{X}_i = \{\mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{X}_{i3}, \mathbf{X}_{i4}\}$, $\forall i \in \{1,2,3,4\}$. Suppose the user generates queries Q_1 and Q_2 from a valid PIR coding scheme [1] for asking the following answer strings from Server 1 and Server 2. It can be verified that from $\mathbf{A}_1^{[W,S]}$ and*

$\mathbf{A}_1^{[W,S]}$ for Server 1	$\mathbf{A}_2^{[W,S]}$ for Server 2
$\mathbf{X}_{11}+\mathbf{X}_{21}+\mathbf{X}_{31}$	$\mathbf{X}_{12}+\mathbf{X}_{22}+\mathbf{X}_{32}$
$\mathbf{X}_{41}+\mathbf{X}_{51}$	$\mathbf{X}_{42}+\mathbf{X}_{52}$
$\mathbf{X}_{13}+\mathbf{X}_{23}+\mathbf{X}_{33}+\mathbf{X}_{42}+\mathbf{X}_{52}$	$\mathbf{X}_{14}+\mathbf{X}_{24}+\mathbf{X}_{34}+\mathbf{X}_{41}+\mathbf{X}_{51}$

$\mathbf{A}_2^{[W,S]}$, the sum $\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3$ can be decoded and hence $\mathbf{X}_1$ can be decoded given $\mathbf{X}_2$ and $\mathbf{X}_3$ as side information. From the Server 1's perspective, when Q_1 is received to request $\mathbf{A}_1^{[W,S]}$, without the knowledge of Q_2, the only information it can infer from Q_1 is:

(1) If $\mathbf{X}_1$ is the demand message, then $\mathbf{X}_2$ and $\mathbf{X}_3$ must the be side information messages.

(2) If $\mathbf{X}_4$ is the demand message, then $\mathbf{X}_5$ must the be one of the side information messages. The other side information message is not required for decoding $\mathbf{X}_4$.

Thus, the query Q_1 may possibly generated for $(W = 1, S = \{2,3\})$ or $(W = 4, S = \{5,*\})$, where $*$ denotes any other index. For Q_1, the virtual side information for index 1 is $\{2,3\}$ and the virtual side information for index 4 could be any one of $\{\{1,5\},\{2,5\},\{3,5\},\{5\}\}$. And the virtual side information vectors is not unique which could be

$$V_{Q_1} = [v_1, v_2, v_3, v_4, v_5] \tag{5.30}$$

$$v_1 = \{2,3\}, v_2 = \{1,3\}, v_3 = \{1,2\}, v_4 = \{1,5\}, v_5 = \{4\} \tag{5.31}$$

[1] The construction of such coding scheme is shown in Section 5.3.2.

5.2 Useful Techniques and Insights

In this section, we introduce two useful tools which are later used in the proof for the converse of the capacity of multi-server single-message PIR with side information.

5.2.1 PIR Scheme for More Messages

Consider one multi-server single-message PIR with side information coding scheme $\mathscr{C}$ for K message, M side information messages and N servers. The coding scheme $\mathscr{C}$ satisfies both *privacy condition* and *retrieval condition* and is referred to as (K,M,N)-PIR coding scheme. Suppose we only know the set of all answer strings $\mathscr{A}$ and set of all queries $\mathscr{Q}$ of $\mathscr{C}$. Based on the answer strings $\mathscr{A}$ and $\mathscr{Q}$, we present a way to generate another coding scheme $\hat{\mathscr{C}}$ with $\hat{\mathscr{A}}$ and $\hat{\mathscr{Q}}$ for parameter triplet $(K+t,M,N)$ for any $0<t\leq M+1$ as follows:

1. The user generates two empty subsets $\wp_1$ and $\wp_2$ with size $|\wp_1|=K$ and $|\wp_2|=t$.
2. The user randomly selects one subset (either $\wp_1$ or $\wp_2$) to place the demand message $\mathbf{X}_W$ with probability proportional to the size of the subset.

$$\Pr(\mathbf{X}_W\in\wp_1)=\frac{|\wp_1|}{|\wp_1|+|\wp_2|}=\frac{K}{K+t} \tag{5.32}$$

$$\Pr(\mathbf{X}_W\in\wp_2)=\frac{|\wp_2|}{|\wp_1|+|\wp_2|}=\frac{t}{K+t} \tag{5.33}$$

3. If $\mathbf{X}_W\in\wp_1$, the user places all side information message in $\wp_1$. If $\mathbf{X}_W\in\wp_2$, the user randomly selects $t-1$ side information messages to place in $\wp_2$[2].
4. The user randomly distributes the other messages to fill up $\wp_1$ and $\wp_2$.
5. If $\mathbf{X}_W\in\wp_1$, the user generates queries $Q_1^{[W,S]},\ldots,Q_N^{[W,S]}$ from the (K,M,N)-PIR coding scheme for the messages in $\wp_1$ and send them to the servers, respectively. If $\mathbf{X}_W\in\wp_2$, the user randomly select $\mathbf{X}_{W'}\in\wp_1$ and $\mathbf{X}_{S'}\subseteq(\wp_1\setminus\{X_{W'}\})$ to generate queries $Q_1^{[W',S']},\ldots,Q_N^{[W',S']}$ from the (K,M,N)-PIR coding scheme for the messages in $\wp_1$ and send them to the servers, respectively.
6. The user additionally query Q_b which requests the sum of all messages in $\wp_2$ and sends Q_b to Server 1.
7. The server which receives query $Q_j^{[W,S]}$ or $Q_j^{[W',S']}$ replies with the corresponding answer string $\mathbf{A}_j=f_{Q_j^{[W,S]}}(\mathbf{X}_{\wp_1})$ or $\mathbf{A}_j=f_{Q_j^{[W',S']}}(\mathbf{X}_{\wp_1})$, respectively. Server 1 additionally replies another answer string $\mathbf{B}=\sum_{\mathbf{X}\in\wp_2}\mathbf{X}$.

If $\mathbf{X}_W\in\wp_1$, the generated queries $\hat{Q}_1^{[W,S]},\ldots,\hat{Q}_N^{[W,S]}$ and answer strings $\hat{\mathbf{A}}_1^{[W,S]},\ldots,\hat{\mathbf{A}}_N^{[W,S]}$ are shown as follows:

[2] When $\mathbf{X}_W\in\wp_2$ and $t<M-1$, not all side information messages can be placed in $\wp_2$.

$\hat{\mathbf{A}}_1^{[W,S]}$ for Server 1	$\hat{\mathbf{A}}_2^{[W,S]}$ for Server 2	...	$\hat{\mathbf{A}}_N^{[W,S]}$ for Server N
$\mathbf{A}_1^{[W,S]}, B$	$\mathbf{A}_2^{[W,S]}$	...	$\mathbf{A}_N^{[W,S]}$

$$\hat{Q}_1^{[W,S]} = \{Q_1^{[W,S]}, Q_b\} \tag{5.34}$$
$$\hat{Q}_{2:N}^{[W,S]} = Q_{2:N}^{[W,S]} \tag{5.35}$$

$$\hat{\mathbf{A}}_1^{[W,S]} = \{\mathbf{A}_1^{[W,S]}, \mathbf{B}\} \tag{5.36}$$
$$\hat{\mathbf{A}}_{2:N}^{[W,S]} = \mathbf{A}_{2:N}^{[W,S]} \tag{5.37}$$

If $\mathbf{X}_W \in \wp_2$, the generated queries $\hat{Q}_1^{[W,S]}, \ldots, \hat{Q}_N^{[W,S]}$ and answer strings $\hat{\mathbf{A}}_1^{[W,S]}, \ldots, \hat{\mathbf{A}}_N^{[W,S]}$ are shown as follows:

$\hat{\mathbf{A}}_1^{[W,S]}$ for Server 1	$\hat{\mathbf{A}}_2^{[W,S]}$ for Server 2	...	$\hat{\mathbf{A}}_N^{[W,S]}$ for Server N
$\mathbf{A}_1^{[W',S']}, \mathbf{B}$	$\mathbf{A}_2^{[W',S']}$	...	$\mathbf{A}_N^{[W',S']}$

$$\hat{Q}_1^{[W,S]} = \{Q_1^{[W',S']}, Q_b\} \tag{5.38}$$
$$\hat{Q}_{2:N}^{[W,S]} = Q_{2:N}^{[W',S']} \tag{5.39}$$

$$\hat{\mathbf{A}}_1^{[W,S]} = \{\mathbf{A}_1^{[W',S']}, \mathbf{B}\} \tag{5.40}$$
$$\hat{\mathbf{A}}_{2:N}^{[W,S]} = \mathbf{A}_{2:N}^{[W',S']} \tag{5.41}$$

Lemma 5.2. *The constructed $(K+t, M, N)$-PIR coding scheme satisfies both retrieval condition and privacy condition.*

Proof. If $\mathbf{X}_W \in \wp_1$, according to Equation (5.36) and (5.37), the answer strings $\hat{\mathbf{A}}_1^{[W,S]}, \ldots, \hat{\mathbf{A}}_N^{[W,S]}$ can be used to recover the answer strings $\mathbf{A}_1^{[W,S]}, \ldots, \mathbf{A}_N^{[W,S]}$, which is generated by a (K, M, N)-PIR coding scheme for messages in $\wp_1$ and hence satisfy *retrieval condition*. Therefore, from $\mathbf{A}_1^{[W,S]}, \ldots, \mathbf{A}_N^{[W,S]}$, $\mathbf{X}_W$ can be decoded with $\mathbf{X}_S$ as side information. If $\mathbf{X}_W \in \wp_2$, according to Equation (5.40), the answer string $\hat{\mathbf{A}}_1^{[W,S]}$ can be use to recover answer string $\mathbf{B}$, which is the sum of $\mathbf{X}_W$ and messages in $\mathbf{X}_S$. The user can decode the demand message $\mathbf{X}_W$ from $\mathbf{B}$ with $\mathbf{X}_S$ as side information. Hence, the *retrieval condition* (5.12) is satisfied.

The probability for any index $W \in \{1, \ldots, K+t\}$ to be the demand index can be computed as

$$\Pr(\mathbf{W} = W | \hat{\mathbf{A}}_{1:N}, \mathbf{Q}_{1:N}) = \Pr(\mathbf{W} = W, X_W \in \wp_i | \hat{\mathbf{A}}_{1:N}, \mathbf{Q}_{1:N}) \tag{5.42}$$
$$= \Pr(\mathbf{W} = W | X_W \in \wp_i, \hat{\mathbf{A}}_{1:N}, \mathbf{Q}_{1:N}) \Pr(X_W \in \wp_i | \hat{\mathbf{A}}_{1:N}, \mathbf{Q}_{1:N}) \tag{5.43}$$
$$= \frac{1}{|\wp_i|} \frac{|\wp_i|}{K+t} \tag{5.44}$$
$$= \frac{1}{K+t}. \tag{5.45}$$

From each server's perspective, every message has the same probability ($\frac{1}{K+t}$) to be the demand message. Therefore, the *privacy condition* (5.18) is also satisfied. □

Lemma 5.3. *For any queries* $\mathbf{Q}_1^{[W,S]},\ldots,\mathbf{Q}_N^{[W,S]}$ *and answer strings* $\mathbf{A}_1^{[W,S]},\ldots,\mathbf{A}_N^{[W,S]}$ *from any* (K,M,N)*-PIR coding scheme and any* $t\in\{1,\ldots,M+1\}$*, there exist queries* $\hat{Q}_1^{[W,S]},\ldots,\hat{Q}_N^{[W,S]}$ *with answer strings* $\hat{\mathbf{A}}_1^{[W,S]},\ldots,\hat{\mathbf{A}}_N^{[W,S]}$ *from the constructed* $(K+t,M,N)$*-PIR coding scheme such that*

$$H(\hat{\mathbf{A}}_j^{[W,S]}|\hat{\mathbf{Q}}_j^{[W,S]}=\hat{Q}_j^{[W,S]},\mathbf{X}_\tau)=H(\mathbf{A}_j^{[W,S]}|\mathbf{Q}_j^{[W,S]}=Q_j^{[W,S]}),\forall\, j\in\{1,\ldots,N\},\tag{5.46}$$

for some $\tau\subseteq\{1,\ldots,K+t\}\setminus(W\cup S)$ *with* $|\tau|=t$.

Proof. Without loss of generality, we assume that the (K,M,N)-PIR coding scheme applies on messages $\{\mathbf{X}_1,\ldots,\mathbf{X}_K\}$. This assumption also implies $W\in\{1,\ldots,K\}$ and $S\subseteq\{1,\ldots,K\}\setminus\{W\}$. Then we can pick $\tau=\{K+1,\ldots,K+t\}$ and hence $\mathbf{X}_\tau=\{\mathbf{X}_{K+1},\ldots,\mathbf{X}_{K+t}\}$. For any $W\in\{1,\ldots,K\}$, $S\subseteq\{1,\ldots,K\}\setminus\{W\}$, suppose the (K,M,N)-PIR coding scheme generates $Q_1^{[W,S]},\ldots,Q_N^{[W,S]}$. It is always possible to find corresponding queries $\hat{Q}_1^{[W,S]},\ldots,\hat{Q}_N^{[W,S]}$ from the $(K+t,M,N)$-PIR coding scheme such that $\wp_1=\{\mathbf{X}_1,\ldots,\mathbf{X}_K\}$ and $\wp_2=\{\mathbf{X}_{K+1},\ldots,\mathbf{X}_{K+t}\}$, $\hat{Q}_1^{[W,S]}=\{Q_1^{[W,S]},Q_b\}$ and $\hat{Q}_{2:N}^{[W,S]}=Q_{2:N}^{[W,S]}$. We note that answer string $\mathbf{B}$ is deterministic given messages $\mathbf{X}_\tau$, i.e.,

$$H(\mathbf{B}|\hat{\mathbf{Q}}_1=\{Q_1^{[W,S]},Q_b\},\mathbf{X}_\tau)=0.\tag{5.47}$$

Thus, for Server 1, we have

$$\begin{aligned}H(\hat{\mathbf{A}}_1^{[W,S]}|\hat{\mathbf{Q}}_1^{[W,S]}=\hat{Q}_1^{[W,S]},\mathbf{X}_\tau)&=H(\mathbf{A}_1^{[W,S]},\mathbf{B}|\hat{\mathbf{Q}}_1^{[W,S]}=\{Q_1^{[W,S]},Q_b\},\mathbf{X}_\tau)&(5.48)\\&=H(\mathbf{A}_1^{[W,S]}|\mathbf{Q}_1^{[W,S]}=Q_1^{[W,S]}),&(5.49)\end{aligned}$$

and for Server $j\in\{2,\ldots,N\}$, we have

$$H(\hat{\mathbf{A}}_j^{[W,S]}|\hat{\mathbf{Q}}_j^{[W,S]}=\hat{Q}_j^{[W,S]},\mathbf{X}_\tau)=H(\mathbf{A}_j^{[W,S]}|\mathbf{Q}_j^{[W,S]}=Q_j^{[W,S]})\tag{5.50}$$

Therefore, for all $j\in\{1,\ldots,N\}$, Equation (5.46) is satisfied. □

Remark 5.1. *We note that the constructed coding scheme with answer strings* $\hat{\mathbf{A}}_1^{[W,S]},\ldots,\hat{\mathbf{A}}_N^{[W,S]}$ *is not necessarily optimal.*

5.2.2 PIR Scheme for Fewer Messages

In this section, we introduce a simple and novel approach, *setting constants*, to generate answer strings for (K',M,N)-PIR coding schemes from answer strings for $(K.M,N)$-PIR coding schemes with $K'<K$.

Lemma 5.4. *Given answer strings* $\mathscr{A}$ *for any* $(K.M,N)$*-PIR coding scheme for message* $\mathbf{X}_{1:K}$*, the answer strings* $\ddot{\mathscr{A}}$ *for* (K',M,N)*-PIR coding scheme for messages* $\mathbf{X}_{1:K}\setminus\mathbf{X}_{\mathscr{K}}$ *for any* $\mathscr{K}\subset\{1,\ldots,K\}$*can be generated by setting messages* $\mathbf{X}_{\mathscr{K}}$ *to constants.*

Proof. Let $\mathbf{A}_1,\ldots,\mathbf{A}_N$ and $Q_1,\ldots,Q_N$ denote one group of answer strings and queries generated by any $(K.M,N)$-PIR coding scheme for message $\mathbf{X}_{1:K}$. Each $\mathbf{A}_j$ $(\forall\, j \in \{1,N\})$ is a deterministic function of Q_j and $\mathbf{X}_{1:K}$ and can be written as follows.

$$\mathbf{A}_j = f_{Q_j}(\mathbf{X}_1,\ldots,\mathbf{X}_K) = f_{Q_j}(\mathbf{X}_{1:K}) \tag{5.51}$$

By setting the messages $\mathbf{X}_{\mathscr{K}}$ into constant c, where c is any value that could possibly taken by the messages, we get $\ddot{\mathbf{A}}$

$$\begin{aligned}\ddot{\mathbf{A}}_j =& f_{Q_j}(\mathbf{X}_1,\ldots,\mathbf{X}_K|\mathbf{X}_{\mathscr{K}} = c) &(5.52)\\ =& f_{\ddot{Q}_j}(\mathbf{X}_{1:K}\setminus\mathbf{X}_{\mathscr{K}}) &(5.53)\end{aligned}$$

Let the probability of choosing $\ddot{\mathbf{A}}_1,\ldots,\ddot{\mathbf{A}}_N$ to be the answer strings for given $\mathbf{W}=W$ be the same probability as $\mathbf{A}_1,\ldots,\mathbf{A}_N$, i.e.,

$$\Pr(\ddot{\mathbf{A}}_1,\ldots,\ddot{\mathbf{A}}_N|\mathbf{W}=W) = \Pr(\mathbf{A}_1,\ldots,\mathbf{A}_N|\mathbf{W}=W). \tag{5.54}$$

For Server j,

$$\Pr(\ddot{\mathbf{A}}_j|\mathbf{W}=W) = \Pr(\mathbf{A}_j|\mathbf{W}=W). \tag{5.55}$$

For any message $\mathbf{X}_W$ which can be decoded from $\mathbf{A}_1,\ldots,\mathbf{A}_N$ which can be decoded given $\mathbf{X}_S$ as side information and $W \notin \mathscr{K}$, i.e.,

$$H(\mathbf{X}_W|\mathbf{A}_1,\ldots,\mathbf{A}_N,\mathbf{X}_S) = 0, \tag{5.56}$$

we have

$$\begin{aligned}H(\mathbf{X}_W|\ddot{\mathbf{A}}_1,\ldots,\ddot{\mathbf{A}}_N,\mathbf{X}_S) =& H(\mathbf{X}_W|f_{\ddot{Q}_1}(\mathbf{X}_{1:K}\setminus\mathbf{X}_{\mathscr{K}}),\ldots,f_{\ddot{Q}_N}(\mathbf{X}_{1:K}\setminus\mathbf{X}_{\mathscr{K}}),\mathbf{X}_S) &(5.57)\\ =& H(\mathbf{X}_W|f_{Q_1}(\mathbf{X}_{1:K}),\ldots,f_{Q_N}(\mathbf{X}_{1:K}),\mathbf{X}_S,\mathbf{X}_{\mathscr{K}}) &(5.58)\\ \le& H(\mathbf{X}_W|\mathbf{A}_1,\ldots,\mathbf{A}_N,\mathbf{X}_S) &(5.59)\\ =& 0. &(5.60)\end{aligned}$$

Thus, message $\mathbf{X}_W$ can also be decoded from $\ddot{\mathbf{A}}_1,\ldots,\ddot{\mathbf{A}}_N$ given $\mathbf{X}_S$ as side information, which satisfies the *retrieval condition*.

The probability for any index $W \in \{1,\ldots,K\}\setminus\mathscr{K}$ to be the demand index given the answer

string $\ddot{\mathbf{A}}_j$ and queries $\ddot{\mathbf{Q}}_j$ for all $j \in \{1,\dots,N\}$ can be computed as

$$\Pr(\mathbf{W}=W|\ddot{\mathbf{A}}_j) = \frac{\Pr(\mathbf{W}=W,\ddot{\mathbf{A}}_j)}{\Pr(\ddot{\mathbf{A}}_j)} \tag{5.61}$$

$$= \frac{\Pr(\mathbf{W}=W)\Pr(\ddot{\mathbf{A}}_j|\mathbf{W}=W)}{\sum_{W'\in\{1,\dots,K\}\setminus\mathscr{K}}\Pr(\ddot{\mathbf{A}}_j|\mathbf{W}=W')} \tag{5.62}$$

$$= \frac{\Pr(\mathbf{W}=W)\Pr(\mathbf{A}_j|\mathbf{W}=W)}{\sum_{W'\in\{1,\dots,K\}\setminus\mathscr{K}}\Pr(\mathbf{A}_j|\mathbf{W}=W')} \tag{5.63}$$

$$= \Pr(\mathbf{W}=W|\mathbf{A}_j,\mathbf{W}\notin\mathscr{K}) \tag{5.64}$$

$$= \frac{1}{K-|\mathscr{K}|}. \tag{5.65}$$

From Server j's perspective, every index in $\{1,\dots,K\}\setminus\mathscr{K}$ has the same probability to be de demand index, which satisfies the *privacy condition*.

Therefore, the constructed (K',M,N)-PIR coding scheme with answer strings $\ddot{\mathscr{A}}$ is a valid PIR coding scheme which satisfies both *retrieval condition* and *privacy condition*. □

We have the following remarks for the operation, *Setting Constants*.

- The constant c can be any possible value which the message is allowed to take. And for different answer string $\mathbf{A}_j$, c can also be chosen differently. Since the user only knows the value of the side information messages and does not know the value of demand message and other unwanted messages, the answer strings should be compatible with all possible values. There is no need to worry about not well-defined functions like 0 in the denominator.
- Setting constant can be interpreted as telling the server that none of those messages is the demand message. From the original answer string $\mathbf{A}_j$, Server j can not infer which index in $\{1,\dots,K\}$ is the demand index. From the generated answer string $\ddot{\mathbf{A}}_j$, Server j can only know $\mathscr{K}$ contains no demand index, but still cannot infer which index in $\{1,\dots,K\}\setminus\mathscr{K}$ is the demand index.

5.2.3 No Need to Reuse Indices

In this section, we present an informative insight into the virtual side information. According to Lemma 5.1, for any query realization Q_j, it is always possible to find at least one virtual side information for every index. Suppose for query Q_j, the virtual side information for W_0 is denoted by S_0 and there exists another index $W_1 \in \{1,\dots,K\}\setminus(W_0\cup S_0)$, the virtual side information for W_1 is denoted by $S_{1,j}$. The query Q_j can be treated as $Q_j^{[W_0,S_0]}$ or $Q_j^{[W_1,S_{1,j}]}$ and satisfies

$$H(\mathbf{A}_j^{[W_0,S_0]}|\mathbf{Q}_j^{[W_0,S_0]}=Q_j^{[W_0,S_0]},\mathbf{X}_{W_0\cup S_0}) = H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]}=Q_j^{[W_1,S_{1,j}]},\mathbf{X}_{W_0\cup S_0}). \tag{5.66}$$

In general, $S_{1,j}$ is only restricted to satisfy $S_{1,j} \subseteq \{1,\dots,K\} \setminus \{W_1\}$ and $|S_{1,j}| \leq M$, which means $S_{1,j} \cap (W_0 \cup S_0)$ may not always be empty. One interesting question would be can we restrict the $S_{1,j} \cap (W_0 \cup S_0) = \emptyset$ and still can find $S_{1,j}$ which satisfies Equation (5.66)?

Lemma 5.5. *Consider one query realization $Q_j^{[W_0,S_0]}$ and corresponding answer string $\mathbf{A}_j^{[W_0,S_0]}$ which are generated for demand index W_0 and side information indices S_0 by a PIR with side information coding scheme. Suppose $Q_j^{[W_0,S_0]}$ has virtual side information $S_{1,j}$ for index W_1 such that $W_1 \notin W_0 \cup S_0$. If $S_{1,j} \cap (W_0 \cup S_0) = \tau \neq \emptyset$, it is possible to construct another PIR with side information coding scheme with query realization $\hat{Q}_j^{[W_0,S_0]}$ and answer string $\hat{\mathbf{A}}_j^{[W_0,S_0]}$ for demand index W_0 and side information indices S_0 and have virtual side information $\hat{S}_{1,j}$ for index W_1 which satisfy $\hat{S}_{1,j} = S_{1,j} \setminus \tau$ and*

$$H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]},\mathbf{X}_{W_0\cup S_0}) = H(\hat{\mathbf{A}}_j^{[W_1,\hat{S}_{1,j}]}|\hat{\mathbf{Q}}_j^{[W_1,\hat{S}_{1,j}]} = \hat{Q}_j^{[W_1,\hat{S}_{1,j}]},\mathbf{X}_{W_0\cup S_0}). \quad (5.67)$$

Proof. According to Lemma 5.1, we have

$$H(\mathbf{A}_j^{[W_0,S_0]}|\mathbf{Q}_j^{[W_0,S_0]} = Q_j^{[W_0,S_0]},\mathbf{X}_{W_0\cup S_0}) = H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]},\mathbf{X}_{W_0\cup S_0}), \quad (5.68)$$

where query $Q_j^{[W_0,S_0]} = Q_j^{[W_1,S_{1,j}]}$ are the same query and $\mathbf{A}_j^{[W_0,S_0]} = \mathbf{A}_j^{[W_1,S_{1,j}]}$ are the same answer string. We note that query $\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]}$ and the corresponding answer string $\mathbf{A}_j^{[W_1,S_{1,j}]}$ are also valid for demand index W_1 and side information indices $S_{1,j}$. There exists a group of answer strings, denoted by $\mathbf{A}_{1:N}^{[W_1,S_{1,j}]}$, from which $\mathbf{X}_{W_1}$ can be decoded given $\mathbf{X}_{S_{1,j}}$ as side information. Note that Server j only knows that its own answer string is $A_j^{[W_1,S_{1,j}]}$ and does not know answer strings from other servers.

Let $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}]}$ denote the answer string which are obtained by setting $\mathbf{X}_{W_0\cup S_0}$ to constants in $\mathbf{A}_j^{[W_1,S_{1,j}]}$ as described in Section 5.2.2, i.e.

$$\mathbf{A}_j^{[W_1,S_{1,j}]} = f_{Q_j^{[W_1,S_{1,j}]}}(\mathbf{X}_1,\mathbf{X}_2,\dots,\mathbf{X}_K), \quad (5.69)$$

$$\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}]} = f_{Q_j^{[W_1,S_{1,j}]}}(\mathbf{X}_1,\mathbf{X}_2,\dots,\mathbf{X}_K|\mathbf{X}_i = c, \forall i \in W_0 \cup S_0), \quad (5.70)$$

where c is any constant value the messages could possibly take. Since the original answer sting $\mathbf{A}_j^{[W_1,S_{1,j}]}$ is one answer string from (K,M,N)-PIR coding scheme, denoted by $\mathscr{C}_1$, according to Lemma 5.4, answer string $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}]}$ is from an $(K-|W_0\cup S_0|,M,N)$-PIR coding scheme. The reused virtual side information messages for W_1 indexed by τ are constants in answer string $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}]}$. Thus, one virtual side information for W_1 in the query corresponding to $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}]}$ is $S_{1,j} \setminus \tau$, which does not contain any index from $W_0 \cup S_0$. Hence, we can rewrite $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}]}$ as $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}\setminus\tau]}$.

According to Lemma 5.2, it is possible to construct a valid (K,M,N)-PIR coding scheme with answer string $\hat{\mathbf{A}}_j^{[W_1,S_{1,j}\setminus\tau]}$ based on the $(K-|W_0\cup S_0|,M,N)$-PIR coding scheme with answer

string $\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}\setminus\tau]}$ such that

$$\begin{aligned}
&H(\hat{\mathbf{A}}_j^{[W_1,S_{1,j}\setminus\tau]}|\hat{\mathbf{Q}}_j^{[W_1,S_{1,j}\setminus\tau]}=\hat{Q}_j^{[W_1,S_{1,j}\setminus\tau]},\mathbf{X}_{W_0\cup S_0})\\
=&H(\ddot{\mathbf{A}}_j^{[W_1,S_{1,j}\setminus\tau]}|\ddot{\mathbf{Q}}_j^{[W_1,S_{1,j}\setminus\tau]}=\ddot{Q}_j^{[W_1,S_{1,j}\setminus\tau]}) &(5.71)\\
=&H(f_{Q_j^{[W_1,S_{1,j}]}}(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_K|\mathbf{X}_i=c,\forall i\in W_0\cup S_0)|\ddot{\mathbf{Q}}_j^{[W_1,S_{1,j}\setminus\tau]}=\ddot{Q}_j^{[W_1,S_{1,j}\setminus\tau]}) &(5.72)\\
=&H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]}=Q_j^{[W_1,S_{1,j}]},\mathbf{X}_{W_0\cup S_0}), &(5.73)
\end{aligned}$$

where Equation (5.71) is from Lemma 5.3 Equation (5.46). □

Before moving on, we would like to mention that

- For any answer string $\mathbf{A}_j$ with $S_{1,j}\cap(W_0\cup S_0)=\tau\neq\emptyset$, with out loss of optimality, we can construct another answer string $\hat{\mathbf{A}}_j$ which has $\hat{S}_{1,j}$ as the virtual side information for W_1 and $\hat{S}_{1,j}$ does not reuse indices in $W_0\cup S_0$.
- It is good to note that we use the notation $S_{1,j}$ for the virtual side information for W_1 in query $Q_j^{[W_0,S_0]}$, which implies the fact that the virtual side information for the non-demand index can be different for queries for different servers. Basically, it is possible that the virtual side information for W_1 are $S_{1,1}$ for Server 1 and $S_{1,2}$ for Server 2 such that $S_{1,1}\neq S_{1,2}$.
- The virtual side information for W_0 is always S_0 for queries for all servers, since W_0 and S_0 are the real demand and side information indices.

Example 5.2. *Suppose a* $(5,2,2)$*-PIR with side information coding scheme generates answer strings* $\mathbf{A}_1^{[W,S]}$ *and* $\mathbf{A}_2^{[W,S]}$ *for* $W=1$ *and* $S=\{2,3\}$.

$\mathbf{A}_1^{[W,S]}$ for Server 1	$\mathbf{A}_2^{[W,S]}$ for Server 2
$\mathbf{X}_{11}+\mathbf{X}_{21}+\mathbf{X}_{31}$	$\mathbf{X}_{12}+\mathbf{X}_{22}+\mathbf{X}_{32}$
$\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{31}$	$\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{32}$
$\mathbf{X}_{13}+\mathbf{X}_{23}+\mathbf{X}_{33}+\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{32}$	$\mathbf{X}_{14}+\mathbf{X}_{24}+\mathbf{X}_{34}+\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{31}$

From Server 1*'s perspective, the virtual side information for* $W_0=1$ *is* $S_0=\{2,3\}$ *and virtual side information for* $W_1=4$ *is* $S_{1,1}=\{5,3\}$. *In such case,* $\tau=S_{1,1}\cap(W_0\cup S_0)=\{3\}$. *The corresponding* $\mathbf{A}_1^{[W_1,S_{1,1}]}$ *and* $\mathbf{A}_2^{[W_1,S_{1,1}]}$ *can be expressed as*

$\mathbf{A}_1^{[W_1,S_{1,1}]}$ for Server 1	$\mathbf{A}_2^{[W_1,S_{1,1}]}$ for Server 2
$\mathbf{X}_{11}+\mathbf{X}_{21}+\mathbf{X}_{31}$	$\mathbf{X}_{13}+\mathbf{X}_{23}+\mathbf{X}_{33}$
$\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{31}$	$\mathbf{X}_{43}+\mathbf{X}_{53}+\mathbf{X}_{33}$
$\mathbf{X}_{13}+\mathbf{X}_{23}+\mathbf{X}_{33}+\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{32}$	$\mathbf{X}_{11}+\mathbf{X}_{21}+\mathbf{X}_{31}+\mathbf{X}_{44}+\mathbf{X}_{54}+\mathbf{X}_{34}$

Note that $\mathbf{A}_{1:2}^{[W,S]}$ *cannot be used to decode* $\mathbf{X}_{W_1} = \mathbf{X}_4$ *even given* $\mathbf{X}_{S_{1,1}} = \{\mathbf{X}_5, \mathbf{X}_3\}$ *as side information but* $\mathbf{A}_{1:2}^{[W_1,S_{1,1}]}$ *can be used to decode* $\mathbf{X}_{W_1}$ *given* $\mathbf{X}_{S_1}$ *as side information. The modified answer string* $\ddot{\mathbf{A}}_1^{[W_1,S_{1,1}]}$ *can be obtained by setting all component of* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ *to be constant and can be expressed as*

$$\ddot{\mathbf{A}}_1^{[W_1,S_{1,1}]} = \left\{ \begin{array}{l} c_1 + c_1 + c_1 \\ \mathbf{X}_{41} + \mathbf{X}_{51} + c_1 \\ c_3 + c_3 + c_3 + \mathbf{X}_{42} + \mathbf{X}_{52} + c_2 \end{array} \right\} = \left\{ \begin{array}{l} \mathbf{X}_{41} + \mathbf{X}_{51} + c_1 \\ c_3 + c_3 + c_3 + \mathbf{X}_{42} + \mathbf{X}_{52} + c_2 \end{array} \right\} \tag{5.74}$$

where c_1, c_2, c_3, c_4 *are four constants for four components of* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$*. Since* $c_1 + c_1 + c_1$ *is a constant and the user knows it, Server* 1 *does not have to send it back to the user. Similarly, we have*

$$\ddot{\mathbf{A}}_2^{[W_1,S_{1,1}]} = \left\{ \begin{array}{l} c_3 + c_3 + c_3 \\ \mathbf{X}_{43} + \mathbf{X}_{53} + c_3 \\ c_1 + c_1 + c_1 + \mathbf{X}_{44} + \mathbf{X}_{54} + c_4 \end{array} \right\} = \left\{ \begin{array}{l} X_{43} + X_{53} + c_3 \\ c_1 + c_1 + c_1 + \mathbf{X}_{44} + \mathbf{X}_{54} + c_4 \end{array} \right\} \tag{5.75}$$

Then, the constructed coding scheme with answer string $\hat{\mathbf{A}}_1^{[W_1,S_{1,1}\setminus\tau]}$ *and* $\hat{\mathbf{A}}_2^{[W_1,S_{1,1}\setminus\tau]}$ *are*

$$\hat{\mathbf{A}}_1^{[W_1,S_{1,1}\setminus\tau]} = \left\{ \begin{array}{l} \ddot{\mathbf{A}}_1^{[W_1,S_{1,1}]} \\ \mathbf{B} \end{array} \right\} = \left\{ \begin{array}{l} \mathbf{X}_{41} + \mathbf{X}_{51} + c_1 \\ c_3 + c_3 + c_3 + \mathbf{X}_{42} + \mathbf{X}_{52} + c_2 \\ \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 \end{array} \right\} \tag{5.76}$$

$$\hat{\mathbf{A}}_2^{[W_1,S_{1,1}\setminus\tau]} = \{\ddot{\mathbf{A}}_2^{[W_1,S_{1,1}]}\} = \left\{ \begin{array}{l} \mathbf{X}_{43} + \mathbf{X}_{53} + c_3 \\ c_1 + c_1 + c_1 + \mathbf{X}_{44} + \mathbf{X}_{54} + c_4 \end{array} \right\} \tag{5.77}$$

In $\hat{\mathbf{A}}_1^{[W_1,S_{1,1}\setminus\tau]}$*, the virtual side information,* $S_{1,1}$*, for* $W_1 = 4$ *does not contain index* 3 *any more. It can be verified that*

$$H(\hat{\mathbf{A}}_1^{[W_1,S_{1,1}\setminus\tau]} | \hat{\mathbf{Q}}_1^{[W_1,S_{1,1}\setminus\tau]} = \hat{Q}_1^{[W_1,S_{1,1}\setminus\tau]}, \mathbf{X}_{W_0\cup S_0}) = H(\mathbf{A}_1^{[W_1,S_{1,1}]} | \mathbf{Q}_1^{[W_1,S_{1,1}]} = Q_1^{[W_1,S_{1,1}]}, \mathbf{X}_{W_0\cup S_0}). \tag{5.78}$$

5.2.4 The Symmetry in Unwanted Indices

In previous Section 5.2.3, we show that without loss of optimality, we can assume that the virtual side information $S_{1,j}$ for W_1 contains no index from $W_0 \cup S_0$. Thus, we have $S_{1,j} \subseteq \{1,\ldots,K\} \setminus (W_0 \cup S_0 \cup W_1)$. Another interesting question is does the actually indices in $S_{1,j}$ changes the value of the conditional entropy $H(\mathbf{A}_j^{[W_1,S_j]} | \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0})$? Intuitively, due to the symmetry in unwanted indices, the value of the conditional entropy term should not depend on the indices in $S_{1,j}$. We show that the size of $S_{1,j}$ is more important in the

following Lemma 5.6.

Lemma 5.6. *Let* $g(S_j) = H(\mathbf{A}_j^{[W_1,S_j]}|\mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0})$. *For any two subset of indices,* $S_{1,j}, S'_{1,j} \subseteq \{1,\dots,K\}\setminus\{W_1\}$, *if* $H(\mathbf{X}_{S_{1,j}}|\mathbf{X}_{W_0\cup S_0}) = H(\mathbf{X}_{S'_{1,j}}|\mathbf{X}_{W_0\cup S_0})$, *we have* $g(S_{1,j}) = g(S'_{1,j})$.

Proof. For any $S_j \subseteq \{1,\dots,K\}\setminus\{W_1\}$, we have:

$$
\begin{aligned}
& H(\mathbf{A}_j^{[W_1,S_j]}|\mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) \\
=& H(\mathbf{X}_{W_1\cup S_j}, \mathbf{A}_j^{[W_1,S_j]}|\mathbf{Q}_j^{[W_0,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) \\
& - H(\mathbf{X}_{W_1\cup S_j}|\mathbf{A}_j^{[W_1,S_j]}, \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) && (5.79) \\
=& H(\mathbf{X}_{W_1\cup S_j}|\mathbf{Q}_j^{[W_0,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) \\
& + H(\mathbf{A}_j^{[W_1,S_j]}|\mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S_j}) \\
& - H(\mathbf{X}_{W_1\cup S_j}|\mathbf{A}_j^{[W_1,S_j]}, \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) && (5.80) \\
=& H(\mathbf{X}_{W_1}|\mathbf{Q}_j^{[W_0,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) \\
& + H(\mathbf{X}_{S_j}|\mathbf{Q}_j^{[W_0,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0\cup W_1}) && (5.81) \\
& + H(\mathbf{A}_j^{[W_1,S_j]}|\mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S_j}) \\
& - H(\mathbf{X}_{S_j}|\mathbf{A}_j^{[W_1,S_j]}, \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0}) \\
& - H(\mathbf{X}_{W_1}|\mathbf{A}_j^{[W_1,S_j]}, \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0\cup S_j}) && (5.82) \\
=& \underbrace{H(\mathbf{X}_{W_1}|\mathbf{X}_{W_0\cup S_0})}_{=L} + \underbrace{H(\mathbf{X}_{S_j}|\mathbf{X}_{W_0\cup S_0})}_{(a)} \\
& + \underbrace{H(\mathbf{A}_j^{[W_1,S_j]}|\mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S_j})}_{(b)} \\
& - \underbrace{H(\mathbf{X}_{S_j}|\mathbf{A}_j^{[W_1,S_j]}, \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0})}_{(c)} \\
& - \underbrace{H(\mathbf{X}_{W_1}|\mathbf{A}_j^{[W_1,S_j]}, \mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_j]}, \mathbf{X}_{W_0\cup S_0\cup S_j})}_{(d)} && (5.83)
\end{aligned}
$$

The answer string $\mathbf{A}_j^{[W_1,S_j]}$ is generated by a deterministic function of the selected query realization $\mathbf{Q}_j^{[W_1,S_j]} = Q_j^{[W_1,S_1]}$ and all messages $\mathbf{X}_{1:K}$, which can be expressed as $\mathbf{A}_j^{[W_1,S_j]} = f_{Q_j^{[W_1,S_j]}}(\mathbf{X}_1,\dots,\mathbf{X}_K)$. Let us first consider $S_j = S_{1,j}$. For any query $\mathbf{Q}_j^{[W_1,S_{1,j}]} = \tilde{Q}_j^{[W_1,S_{1,j}]}$ generated for $(W_1, S_{1,j})$, we have the answer string $\mathbf{A}_j^{[W_1,S_{1,j}]} = f_{\tilde{Q}_j^{[W_1,S_{1,j}]}}(\mathbf{X}_1,\dots,\mathbf{X}_K)$. Consider another answer strings $\mathbf{A}_j'^{[W_1,S_{1,j}]} = f_{\tilde{Q}_j^{[W_1,S_{1,j}]}}(\mathbf{X}'_1,\dots,\mathbf{X}'_K)$, where $\{\mathbf{X}'_1,\dots,\mathbf{X}'_K\}$ are one particular permutation of $\{\mathbf{X}_1,\dots,\mathbf{X}_K\}$. Define the mapping from $\{\mathbf{X}_1,\dots,\mathbf{X}_K\}$ to $\{\mathbf{X}'_1,\dots,\mathbf{X}'_K\}$ for $S'_{1,j}$ as follows. $\forall i \in \{1,\dots,K\}$:

1. If $i \in W_0 \cup S_0 \cup W_1$ or $i \notin S_{1,j} \cup S'_{1,j}$ or $i \in S_{1,j} \cap S'_{1,j}$, $\mathbf{X}'_i = \mathbf{X}_i$
2. If $i \in S_{1,j}$ and $i \notin S'_{1,j}$, $\mathbf{X}'_i = \mathbf{X}_{\hat{i}}$ and $\mathbf{X}'_{\hat{i}} = \mathbf{X}_i$, where $\hat{i} \in S'_{1,j}$ and $\hat{i} \notin S_{1,j}$.

Specifically, we swap the messages with indices only in either $S_{1,j}$ or $S'_{1,j}$. Since $\mathbf{A}_j^{[W_1,S_{1,j}]}$ is a valid answer string from Server j for $(W_1, S_{1,j})$, it is easy to see that $\mathbf{A}_j'^{[W_1,S_{1,j}]}$ is a also valid answer string from Server j for $(W_1, S'_{1,j})$. Hence, we have

$$f_{\tilde{Q}_j^{[W_1,S_{1,j}]}}(\mathbf{X}'_1,\ldots,\mathbf{X}'_K) = \mathbf{A}_j'^{[W_1,S_{1,j}]} = \mathbf{A}_j^{[W_1,S'_{1,j}]} = f_{\tilde{Q}^{[W_1,S'_{1,j}]}}(\mathbf{X}_1,\ldots,\mathbf{X}_K). \tag{5.84}$$

For any two subsets of indices, $S_{1,j}$ and $S'_{1,j}$, such that $H(\mathbf{X}_{S_{1,j}}|\mathbf{X}_{W_0\cup S_0}) = H(\mathbf{X}_{S'_{1,j}}|\mathbf{X}_{W_0\cup S_0})$, terms (a) of Equation (5.83) are the same. For terms (b), we have

$$\begin{aligned}
&H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = \tilde{Q}_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S_{1,j}}) \\
=&H(f_{\tilde{Q}_j^{[W_1,S_{1,j}]}}(\mathbf{X}_1,\ldots,\mathbf{X}_K)|\mathbf{Q}_j^{[W_1,S_{1,j}]} = \tilde{Q}_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S_{1,j}}) &(5.85)\\
=&H(f_{\tilde{Q}_j^{[W_1,S_{1,j}]}}(\mathbf{X}'_1,\ldots,\mathbf{X}'_K)|\mathbf{Q}_j^{[W_1,S_{1,j}]} = \tilde{Q}_j^{[W_1,S_{1,j}]}, \mathbf{X}'_{W_0\cup S_0\cup W_1\cup S_{1,j}}) &(5.86)\\
=&H(f_{\tilde{Q}_j^{[W_1,S'_{1,j}]}}(\mathbf{X}_1,\ldots,\mathbf{X}_K)|\mathbf{Q}_j^{[W_1,S'_{1,j}]} = \tilde{Q}_j^{[W_1,S'_{1,j}]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S'_{1,j}}) &(5.87)\\
=&H(\mathbf{A}_j^{[W_1,S'_{1,j}]}|\mathbf{Q}_j^{[W_1,S'_{1,j}]} = \tilde{Q}_j^{[W_1,S'_{1,j}]}, \mathbf{X}_{W_0\cup S_0\cup W_1\cup S'_{1,j}}). &(5.88)
\end{aligned}$$

Hence, terms (b) are the same for $S_{1,j}$ and $S'_{1,j}$. Similarly, it can be shown that terms (c) and (d) are also the same, respectively. Thus, if $H(\mathbf{X}_{S_{1,j}}|\mathbf{X}_{W_0\cup S_0}) = H(\mathbf{X}_{S'_{1,j}}|\mathbf{X}_{W_0\cup S_0})$, for any query $\tilde{Q}_j^{[W_1,S_{1,j}]}$ generated for $(W_1, S_{1,j})$, it is always possible to generate another query $\tilde{Q}_j^{[W_1,S'_{1,j}]}$ for $(W_1, S'_{1,j})$ such that the value of two conditional entropy terms are the same. Therefore, we have $g(S_{1,j}) = g(S'_{1,j})$ if $H(\mathbf{X}_{S_{1,j}}|\mathbf{X}_{W_0\cup S_0}) = H(\mathbf{X}_{S'_{1,j}}|\mathbf{X}_{W_0\cup S_0})$. □

For Lemma 5.6, we have the following remarks.

- The value of $g(S)$ only depends on $H(\mathbf{X}_S|\mathbf{X}_{W_0\cup S_0})$, which is not surprising because the all messages except the demand and side information messages ($\mathbf{X}_{W_0\cup S_0}$) should be fully symmetric in the answer strings and can be re-indexed arbitrarily.
- We note that it is not straightforward to compute the value of $g(S)$ from $H(\mathbf{X}_S|\mathbf{X}_{W_0\cup S_0})$.

Example 5.3. *Consider an* $(8,2,2)$*-PIR with side information coding scheme which generates answer strings* $\mathbf{A}_1^{[W,S]}$ *and* $\mathbf{A}_2^{[W,S]}$ *for* $W = 1$ *and* $S = \{2,3\}$.

From Server 1*'s perspective, the virtual side information for* $W_1 = 4$ *is* $S_{1,1} = \{5,6\}$. *From Server* 2*'s perspective, the virtual side information for* $W_1 = 4$ *is* $S_{1,2} = \{5,7\}$. *Hence,* $\mathbf{A}_1^{[W,S]}$ *and* $\mathbf{A}_2^{[W,S]}$ *can also be interpreted as* $\mathbf{A}_1^{[W_1,S_{1,1}]}$ *and* $\mathbf{A}_2^{[W_1,S_{1,2}]}$*, respectively. We note that* $S_{1,1}$ *and* $S_{1,2}$ *are different.*

$\mathbf{A}_1^{[W,S]}$ for Server 1	$\mathbf{A}_2^{[W,S]}$ for Server 2
$\mathbf{X}_{11}+\mathbf{X}_{21}+\mathbf{X}_{31}$	$\mathbf{X}_{12}+\mathbf{X}_{22}+\mathbf{X}_{32}$
$\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{61}$	$\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{72}$
$\mathbf{X}_{71}+\mathbf{X}_{81}$	$\mathbf{X}_{62}+\mathbf{X}_{82}$
$\mathbf{X}_{13}+\mathbf{X}_{23}+\mathbf{X}_{33}+\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{62}+\mathbf{X}_{72}+\mathbf{X}_{82}$	$\mathbf{X}_{14}+\mathbf{X}_{24}+\mathbf{X}_{34}+\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{61}+\mathbf{X}_{71}+\mathbf{X}_{81}$

By swapping the different indices in $S_{1,1}$ and $S_{1,2}$, we can get the answer strings $\mathbf{A}_1'^{[W_1,S'_{1,1}]}$ and $\mathbf{A}_2'^{[W_1,S'_{1,2}]}$.

$\mathbf{A}_1'^{[W_1,S'_{1,1}]}$ for Server 1	$\mathbf{A}_2'^{[W_1,S'_{1,2}]}$ for Server 2
$\mathbf{X}_{11}+\mathbf{X}_{21}+\mathbf{X}_{31}$	$\mathbf{X}_{12}+\mathbf{X}_{22}+\mathbf{X}_{32}$
$\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{71}$	$\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{62}$
$\mathbf{X}_{61}+\mathbf{X}_{81}$	$\mathbf{X}_{72}+\mathbf{X}_{82}$
$\mathbf{X}_{13}+\mathbf{X}_{23}+\mathbf{X}_{33}+\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{62}+\mathbf{X}_{72}+\mathbf{X}_{82}$	$\mathbf{X}_{14}+\mathbf{X}_{24}+\mathbf{X}_{34}+\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{61}+\mathbf{X}_{71}+\mathbf{X}_{81}$

We note that $S'_{1,1}=S_{1,2}$ and $g(S_{1,1})$ and $g(S'_{1,1})$ can be computed as follows.

$$\begin{aligned} g(S_{1,1}) &= H(\mathbf{A}_1^{[W_1,S_{1,1}]}|\mathbf{Q}_1^{[W_1,S_{1,1}]}=Q_1^{[W_1,S_{1,1}]},\mathbf{X}_{\{1,2,3\}}) && (5.89)\\ &= H(\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{71},\mathbf{X}_{61}+\mathbf{X}_{81},\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{62}+\mathbf{X}_{72}+\mathbf{X}_{82}) && (5.90)\\ &= \frac{3}{4}L && (5.91)\\ g(S'_{1,1}) &= H(\mathbf{A}_1'^{[W_1,S'_{1,1}]}|\mathbf{Q}_1'^{[W_1,S'_{1,1}]}=Q_1'^{[W_1,S'_{1,1}]},\mathbf{X}_{\{1,2,3\}}) && (5.92)\\ &= H(\mathbf{X}_{42}+\mathbf{X}_{52}+\mathbf{X}_{62},\mathbf{X}_{72}+\mathbf{X}_{82},\mathbf{X}_{41}+\mathbf{X}_{51}+\mathbf{X}_{61}+\mathbf{X}_{71}+\mathbf{X}_{81}) && (5.93)\\ &= \frac{3}{4}L && (5.94) \end{aligned}$$

Hence, the actual indices in virtual side information $S_{1,j}$ are not important. It is always possible to generate answer string $\mathbf{A}'_j$ with virtual side information $S'_{1,j}$ such that $g(S_{1,j})=g(S'_{1,j})$.

5.3 The capacity

Theorem 5.1. *The capacity of multi-server single-message private information retrieval with side information for K messages, M side information messages, and N servers is*

$$C(K,M,N)=\left(1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^{\lceil\frac{K}{M+1}\rceil-1}}\right)^{-1}. \tag{5.95}$$

Proof. The proofs for the converse and achievability are presented in Section 5.3.1 and Section 5.3.2, respectively. □

We have the following remarks regarding formula (5.95).

- The capacity $C(K,M,N)$ is a non-increasing function of K and a non-decreasing function of N and M, which is intuitive since the larger number of messages, less number of servers and less number of side information messages are, the more download bits are required.

- By setting the number of side information message(s) $M=0$, we get the formula for multi-server single-message private information retrieval without side information given in [22].

- By setting the number of servers $N=1$, we recover the formula for single-server single-message private information retrieval with side information given in [49].

- The side information effectively reduces the total number of messages linearly from K to $\lceil \frac{K}{M+1} \rceil$. Specifically, the multi-server single-message PIR with side information problem for K messages and M side information messages has the same capacity as multi-server single-message PIR without side information for $\lceil \frac{K}{M+1} \rceil$ messages.

5.3.1 Converse

In this section, we present the proof for the converse of Theorem 5.1. We need to show that

$$D \geq L\left(1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^{\lceil \frac{K}{M+1} \rceil -1}}\right). \tag{5.96}$$

Let $W_0 = W$ denote the demand index and $S_0 = S$ denote the side information indices. The total number of download bits (D) for any specific query realizations $Q_{1:N}^{[W_0,S_0]}$ can be lower bounded as follows.

$$\begin{aligned}
D \geq & H(\mathbf{A}_{1:N}^{[W_0,S_0]} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{S_0}) && (5.97)\\
= & H(\mathbf{A}_{1:N}^{[W_0,S_0]}, \mathbf{X}_{W_0} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{S_0}) \\
& - H(\mathbf{X}_{W_0} | \mathbf{A}_{1:N}^{[W_0,S_0]}, \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{S_0}) && (5.98)\\
= & H(\mathbf{X}_{W_0} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{S_0}) \\
& + H(\mathbf{A}_{1:N}^{[W_0,S_0]} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{W_0 \cup S_0}) && (5.99)\\
= & L + H(\mathbf{A}_{1:N}^{[W_0,S_0]} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{W_0 \cup S_0}) && (5.100)\\
\geq & L + H(\mathbf{A}_{j}^{[W_0,S_0]} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{W_0 \cup S_0}) && (5.101)\\
= & L + H(\mathbf{A}_{j}^{[W_0,S_0]} | \mathbf{Q}_{j}^{[W_0,S_0]} = Q_{j}^{[W_0,S_0]}, \mathbf{X}_{W_0 \cup S_0}). && (5.102)
\end{aligned}$$

Equation (5.99) is because $H(\mathbf{X}_{W_0} | \mathbf{A}_{1:N}^{[W_0,S_0]}, \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{S_0}) = 0$ from the *retrieval condition* (5.12). Equation (5.100) is because $H(\mathbf{X}_{W_0} | \mathbf{Q}_{1:N}^{[W_0,S_0]} = Q_{1:N}^{[W_0,S_0]}, \mathbf{X}_{S_0}) = H(\mathbf{X}_{W_0}) = L$. Equation (5.102) is because $\mathbf{A}_j^{[W_0,S_0]}$ only depends on $Q_j^{[W_0,S_0]}$ and $\mathbf{X}_{1:K}$, and is independent of any $Q_{j'}^{[W_0,S_0]}$ ($j' \neq j$) given $Q_j^{[W_0,S_0]}$. For the ease of notation, we use $W_0 \cup S_0$ to denote $\{W_0\} \cup S_0$.

According to Lemma 5.1, for query realization $Q_j^{[W_0,S_0]}$ and answer strings $\mathbf{A}_j^{[W_0,S_0]}$, for any $W_1 \in \{1,\ldots,K\}\setminus\{W_0\}$, there must exist $S_{1,j} \subseteq \{1,\ldots,K\}\setminus\{W_1\}$ and $|S_{1,j}| \leq M$ such that $\forall\, j \in \{1,\ldots,N\}$

$$H(\mathbf{A}_j^{[W_0,S_0]}|\mathbf{Q}_j^{[W_0,S_0]} = Q_j^{[W_0,S_0]}, \mathbf{X}_{W_0\cup S_0}) = H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0\cup S_0}), \tag{5.103}$$

where $Q_j^{[W_0,S_0]}$ and $Q_j^{[W_1,S_{1,j}]}$ are actually the same query corresponding to the same answer strings which may be possibly generated for (W_0,S_0) or $(W_1,S_{1,j})$.

We note that $S_{1,j}$ $(\forall\, j \in \{1,\ldots,N\})$ is determined by the query $Q_j^{[W_1,S_{1,j}]}$. Since $Q_1^{[W_1,S_{1,1}]} = Q_1^{[W_0,S_0]},\ldots,Q_N^{[W_1,S_{1,N}]} = Q_N^{[W_0,S_0]}$ are generated for decoding X_{W_0} with X_{S_0} as side information messages, the virtual side information indices $S_{1,j}$ for index W_1 and query $Q_j^{[W_0,S_0]}$ may be chosen differently for each server j, i.e., the sizes of $S_{1,j}$'s may be different, or even if the sizes are the same, the indices may also be different. The difference among $S_{1,j}$'s for different j is the main difficulty of proving the converse for multi-server private information retrieval with side information. However, we prove that the number of download bits for answer string corresponding to query with different virtual side information can be lower bounded by another answer string corresponding to query with the same common virtual side information in Theorem 5.2.

Theorem 5.2. *Consider $Q_1,Q_2,\ldots,Q_N$ are queries generated for demand index W_0 with side information indices S_0 from a valid multi-server single-message PIR with side information coding scheme. If the virtual side information indices for any index $W_1 \notin W_0\cup S_0$ are different for different queries, i.e., $\exists\, j_1 \neq j_2 \in \{1,\ldots,N\}$ such that $S_{1,j_1} \neq S_{1,j_2}$, there exist a common virtual side information S_1 such that $\forall\, j \in \{1,\ldots,N\}$:*

$$H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0\cup S_0}) \geq H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]} = \ddot{Q}_j^{[W_1,S_1]}, \mathbf{X}_{W_0\cup S_0}). \tag{5.104}$$

Proof. The proof is given in the Appendix 5.5.1. □

Before we move on to the proof of converse, it is instructive to notice that:

- The original query Q_j for $j \in \{1,\ldots,N\}$ has the same virtual side information S_0 for W_0, but has different virtual side information $S_{1,j}$ for W_1. The constructed query $\ddot{Q}_j$ for $j \in \{1,\ldots,N\}$ has the same virtual side information S_0 for W_0, and also has the same virtual side information S_1 for W_1.

- The constructed queries $\ddot{Q}_1,\ldots,\ddot{Q}_N$ collectively may not permit the decoding of X_{W_1} with X_{S_1} as side information messages. This is because the constructed queries $\ddot{Q}_1,\ldots,\ddot{Q}_N$ is obtained from the original query $Q_1,\ldots,Q_N$, which are generated for decoding X_{W_0} with X_{S_0} as side information messages. Since the virtual side information $S_{1,j}$'s for W_1 may be different in different query Q_j's, it is not guaranteed that X_{W_1} can be decoded given some side information messages.

Now we can continue the proof of the converse for Theorem 5.1. Taking summation over all $j \in \{1,\ldots,N\}$ at both sides of Equation (5.102), we have

$$ND \geq NL + \sum_{j=1}^{N} H(\mathbf{A}_j^{[W_0,S_0]} | \mathbf{Q}_j^{[W_0,S_0]} = Q_j^{[W_0,S_0]}, \mathbf{X}_{W_0 \cup S_0}) \tag{5.105}$$

$$\geq NL + \sum_{j=1}^{N} H(\mathbf{A}_j^{[W_1,S_{1,j}]} | \mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0 \cup S_0}) \tag{5.106}$$

$$\geq NL + \sum_{j=1}^{N} H(\mathbf{A}_j^{[W_1,S_1]} | \mathbf{Q}_j^{[W_1,S_1]} = \ddot{Q}_j^{[W_1,S_1]}, \mathbf{X}_{W_0 \cup S_0}) \quad [\text{for some } S_1] \tag{5.107}$$

$$\geq NL + \min_{S_1} \sum_{j=1}^{N} H(\mathbf{A}_j^{[W_1,S_1]} | \mathbf{Q}_j^{[W_1,S_1]} = \ddot{Q}_j^{[W_1,S_1]}, \mathbf{X}_{W_0 \cup S_0}), \tag{5.108}$$

where Equation (5.107) is from Theorem 5.2 and Equation (5.108) is because we minimize the sum of conditional entropy terms over all possible choices of S_1.

As we mentioned above, the answer strings for $\ddot{Q}_{1:N}^{[W_1,S_1]}$ collectively may not enable the user to decode $\mathbf{X}_{W_1}$ with $\mathbf{X}_{S_1}$ as side information. Nevertheless, the existence of queries $\hat{Q}_{1:N}^{[W_1,S_1]}$ such that the answer strings of them can be used to decode $\mathbf{X}_{W_1}$ with $\mathbf{X}_{S_1}$ as side information is proved by the following Lemma.

Lemma 5.7. *For any group of queries $Q_1,\ldots,Q_N$ generated for demand index W_0 and side information indices S_0 from any multi-server single-message PIR with side information coding scheme, if the virtual side information of $W_1 \notin W_0 \cup S_0$ in all $Q_1,\ldots,Q_N$ is the same, denoted by S_1, and the corresponding answer strings $\mathbf{A}_1,\ldots,\mathbf{A}_N$ cannot be used to decode X_{W_1} with X_{S_1} as side information messages, i.e.,*

$$H(X_{W_1} | \mathbf{A}_{1:N}, \mathbf{Q}_{1:N} = Q_{1:N}, \mathbf{X}_{S_1}) \neq 0, \tag{5.109}$$

there must exist another group of queries $\hat{Q}_1,\ldots,\hat{Q}_N$ such that the corresponding answer strings $\hat{\mathbf{A}}_1,\ldots,\hat{\mathbf{A}}_N$ can be used to decode X_{W_1} with X_{S_1} as side information, i.e.,

$$H(X_{W_1} | \hat{\mathbf{A}}_{1:N}, \hat{\mathbf{Q}}_{1:N} = \hat{Q}_{1:N}, \mathbf{X}_{S_1}) = 0, \tag{5.110}$$

and for any $j \in \{1,\ldots,N\}$, they satisfy

$$H(\mathbf{A}_j | \mathbf{Q}_j = Q_j, \mathbf{X}_{W_0 \cup S_0}) = H(\hat{\mathbf{A}}_j | \hat{\mathbf{Q}}_j = \hat{Q}_j, \mathbf{X}_{W_0 \cup S_0}). \tag{5.111}$$

Proof. The proof is presented in Appendix 5.5.2. □

For Lemma 5.7, we note that

- For both original queries $Q_1,\ldots,Q_N$ and the new queries $\hat{Q}_1,\ldots,\hat{Q}_N$, the virtual side information indices of W_0 and W_1 are S_0 and S_1, respectively. Hence, for any $j \in \{1,\ldots,N\}$,

Q_j and $\hat{Q}_j$ can be written as $Q_j^{[W_1,S_1]}$ and $\hat{Q}_j^{[W_1,S_1]}$, respectively.

- The answer strings corresponding to $\hat{Q}_1,\ldots,\hat{Q}_N$ can be used to decode $\mathbf{X}_{W_1}$ with $\mathbf{X}_{S_1}$ as side information messages. But they may not permit the decoding of $\mathbf{X}_{W_0}$ as $\mathbf{X}_{S_0}$ as side information messages.

- Both groups of queries download the same number of bits from the servers.

According to Lemma 5.7, we can replace $\ddot{Q}_1^{[W_1,S_1]},\ldots,\ddot{Q}_N^{[W_1,S_2]}$ with $\hat{Q}_1^{[W_1,S_1]},\ldots,\hat{Q}_N^{[W_1,S_2]}$, which satisfies

$$H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]}=\ddot{Q}_j^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0})=H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]}=\hat{Q}_j^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}). \tag{5.112}$$

Therefore, the total number of download bits can be further lower bounded by

$$\begin{aligned}
ND \geq & NL+\min_{S_1}\sum_{j=1}^{N} H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]}=\hat{Q}_j^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}) && (5.113)\\
\geq & NL+\min_{S_1} H(\mathbf{A}_{1:N}^{[W_1,S_1]}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}) && (5.114)\\
= & NL+\min_{S_1} H(\mathbf{X}_{W_1,S_1},\mathbf{A}_{1:N}^{[W_1,S_1]}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0})\\
& -H(\mathbf{X}_{W_1,S_1}|\mathbf{A}_{1:N}^{[W_1,S_1]},\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}) && (5.115)\\
= & NL+\min_{S_1} H(\mathbf{A}_{1:N}^{[W_1,S_1]}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0^1\cup S_0^1})\\
& +H(\mathbf{X}_{W_1\cup S_1}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0})\\
& -H(\mathbf{X}_{S_1}|\mathbf{A}_{1:N}^{[W_1,S_1]},\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}). && (5.116)
\end{aligned}$$

Dividing by N on both sides, we get

$$\begin{aligned}
D \geq & L+\min_{S_1}\frac{1}{N}[H(\mathbf{A}_{1:N}^{[W_1,S_1]}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0^1\cup S_0^1})\\
& +H(\mathbf{X}_{W_1\cup S_1}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0})\\
& -H(\mathbf{X}_{S_1}|\mathbf{A}_{1:N}^{[W_1,S_1]},\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0})]. && (5.117)
\end{aligned}$$

For $i\in\{1,\ldots,K-1\}$, define D_i as follows.

$$\begin{aligned}
D_i = & H(\mathbf{X}_{W_i\cup S_i}|\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}})\\
& -H(\mathbf{X}_{S_i}|\mathbf{A}_{1:N}^{[W_i,S_i]},\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}). && (5.118)
\end{aligned}$$

Then we can rewrite Equation (5.117) as

$$D\geq L+\min_{S_1}\frac{1}{N}H(\mathbf{A}_{1:N}^{[W_1,S_1]}|\mathbf{Q}_{1:N}^{[W_1,S_1]}=\hat{Q}_{1:N}^{[W_1,S_1]},\mathbf{X}_{W_0^1\cup S_0^1})+\frac{D_1}{N}. \tag{5.119}$$

Iteratively, we can use $\{W_{i+1}, S_{i+1}\}$ to replace $\{W_i, S_i\}$ such that $W_{i+1} \notin W_0^i \cup S_0^i$. Accordingly, Theorem 5.2 and Lemma 5.7 can be easily extended to the cases where the conditioning part is $\mathbf{X}_{W_0^{i-1} \cup S_0^{i-1}}$, i.e.,

$$H(\mathbf{A}_j^{[W_i,S_{i,j}]}|\mathbf{Q}_j^{[W_i,S_{i,j}]} = Q_j^{[W_i,S_{i,j}]}, \mathbf{X}_{W_0^{i-1} \cup S_0^{i-1}}) \geq H(\mathbf{A}_j^{[W_i,S_i]}|\mathbf{Q}_j^{[W_i,S_i]} = \ddot{Q}_j^{[W_i,S_i]}, \mathbf{X}_{W_0^{i-1} \cup S_0^{i-1}}), \tag{5.120}$$

and

$$H(\mathbf{A}_j^{[W_i,S_i]}|\mathbf{Q}_j^{[W_i,S_i]} = \ddot{Q}_j^{[W_i,S_i]}, \mathbf{X}_{W_0^{i-1} \cup S_0^{i-1}}) = H(\mathbf{A}_j^{[W_i,S_i]}|\mathbf{Q}_j^{[W_i,S_i]} = \hat{Q}_j^{[W_i,S_i]}, \mathbf{X}_{W_0^{i-1} \cup S_0^{i-1}}). \tag{5.121}$$

Then, after T iterations, the total number of download bits can be bounded by

$$D \geq L + \min_{S_1} \frac{1}{N} H(\mathbf{A}_{1:N}^{[W_1,S_1]}|\mathbf{Q}_{1:N}^{[W_1,S_1]} = \hat{Q}_{1:N}^{[W_1,S_1]}, \mathbf{X}_{W_0^1 \cup S_0^1}) + \frac{D_1}{N} \tag{5.122}$$

$$\geq L + \min_{S_1,S_2} \frac{1}{N^2} H(\mathbf{A}_{1:N}^{[W_2,S_2]}|\mathbf{Q}_{1:N}^{[W_2,S_2]} = \hat{Q}_{1:N}^{[W_2,S_2]}, \mathbf{X}_{W_0^2 \cup S_0^2}) + \frac{D_2}{N^2} + \frac{D_1}{N} \tag{5.123}$$

$$\dots \tag{5.124}$$

$$\geq L + \min_{S_1,\dots,S_T} \frac{1}{N^T} H(\mathbf{A}_{1:N}^{[W_T,S_T]}|\mathbf{Q}_{1:N}^{[W_T,S_T]} = \hat{Q}_{1:N}^{[W_T,S_T]}, \mathbf{X}_{W_0^T \cup S_0^T}) + \frac{D_T}{N^T} + \frac{D_{T-1}}{N^{T-1}} + \cdots + \frac{D_1}{N}. \tag{5.125}$$

Additionally, we assume that after T substitutions, we have

$$W_0^T \cup S_0^T = \{1, \dots, K\}. \tag{5.126}$$

Since $\mathbf{A}_{1:N}^{[W_T,S_T]}$ are deterministic given $\mathbf{Q}_{1:N}^{[W_T,S_T]} = \hat{Q}_{1:N}^{[W_T,S_T]}$ and all messages, $\{\mathbf{X}_1, \dots, \mathbf{X}_K\}$, we have

$$H(\mathbf{A}_{1:N}^{[W_T,S_T]}|\mathbf{Q}_{1:N}^{[W_T,S_T]} = \hat{Q}_{1:N}^{[W_T,S_T]}, \mathbf{X}_{W_0^T \cup S_0^T}) = 0. \tag{5.127}$$

Hence, the total number of download bits satisfies

$$D \geq L + \min_{S_1,\dots,S_T} \frac{D_T}{N^T} + \frac{D_{T-1}}{N^{T-1}} + \cdots + \frac{D_1}{N}. \tag{5.128}$$

To get the lower bound, we need to minimize each D_i. Note that we can get a lower bound on

each D_i as follows.

$$\begin{aligned}
D_i =& H(\mathbf{X}_{W_i\cup S_i}|\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}) \\
&- H(\mathbf{X}_{S_i}|\mathbf{A}_{1:N}^{[W_i,S_i]},\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}) && (5.129)\\
=& H(\mathbf{X}_{W_i}|\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i}}) \\
&+ H(\mathbf{X}_{S_i}|\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}) \\
&- H(\mathbf{X}_{S_i}|\mathbf{A}_{1:N}^{[W_i,S_i]},\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}) && (5.130)\\
=& L+ H(\mathbf{X}_{S_i}|\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}) \\
&- H(\mathbf{X}_{S_i}|\mathbf{A}_{1:N}^{[W_i,S_i]},\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i-1}}) && (5.131)\\
\geq& L, && (5.132)
\end{aligned}$$

where Equation (5.130) is because $H(\mathbf{X}_{W_i}|\mathbf{Q}_{1:N}^{[W_i,S_i]}=\hat{Q}_{1:N}^{[W_i,S_i]},\mathbf{X}_{W_0^{i-1}\cup S_0^{i}})=H(\mathbf{X}_{W_i})=L$ for each $W_i\notin W_0^{i-1}\cup S_0^{i-1}$; and Equation (5.132) is due to the fact that condition cannot increase the entropy. Thus, each D_i is a positive term and is lower bounded by L. In order to get the lower bound for D, we also need to minimize the number of terms D_i for $i=\{1,\dots,T\}$. It is equivalent to maximize the size of $W_0^i\cup S_0^i$ given W_0^i and S_0^{i-1}. Apparently, the optimal choice for S_i is M new indices which are not included in $W_0^i\cup S_0^{i-1}$. Since the total number of messages is K, to satisfy our assumption (5.126), we need

$$T\geq\left\lceil\frac{K-M-1}{M+1}\right\rceil=\left\lceil\frac{K}{M+1}\right\rceil-1. \tag{5.133}$$

Hence, the lower bound for the total number of download bits is

$$\begin{aligned}
D\geq& L\left(1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^T}\right) && (5.134)\\
=& L\left(1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^{\lceil\frac{K}{M+1}\rceil-1}}\right). && (5.135)
\end{aligned}$$

Thus, the capacity of multi-server single-message PIR with side information problem can be upper bounded by

$$C(K,M,N)=\sup\lim_{L\to\infty}\frac{L}{D}\leq\left(1+\frac{1}{N}+\frac{1}{N^2}+\cdots+\frac{1}{N^{\lceil\frac{K}{M+1}\rceil-1}}\right)^{-1}. \tag{5.136}$$

5.3.2 Achievability

In this section, we present a proof of the achievability of Theorem 5.1. We present a coding scheme which satisfies the privacy and retrieval conditions and the rate matches the information-theoretic converse bound (5.136).

In [42], Kadhe *et al.* present an achievability scheme, named Multi-Server W-PIR scheme, for

multi-server single-message private information retrieval with side information when the total number of messages K is divisible by $M+1$. We generalize this Multi-Server W-PIR scheme to the cases where K is not divisible by $M+1$, also see [49, Remark 5]. In line with the terminology in [42], we will refer to this scheme as the *Partition-and-Coding scheme for Multi-Server*. It can be broken down into the following steps:

1. The user first generates $\Theta = \lceil \frac{K}{M+1} \rceil$ empty subsets, denoted by $\wp_1, \ldots, \wp_\Theta$. The first $\Theta - 1$ subsets have size $M+1$ and the last subset has size $K - (\Theta-1)(M+1)$, i.e.,

$$|\wp_1| = |\wp_2| = \cdots = |\wp_{\Theta-1}| = M+1 \tag{5.137}$$

$$|\wp_\Theta| = K - (\Theta-1)(M+1) \tag{5.138}$$

2. The user randomly selects one subset to contain the demand message X_W with probability proportional to the size of the subsets.

$$\Pr(X_W \in \wp_i) = \frac{|\wp_i|}{K}, \forall i \in \{1, \ldots, \Theta\} \tag{5.139}$$

3. The user puts the side information messages in the selected subset until the subset is full[3].

4. The user randomly distributes the other messages to the other subsets.

5. For each subset, the user generates a *super-message* (following the terminology in [42]), which is simply the sum of all messages in the subset.

6. The user applies the PIR coding scheme for the case where there is no side information (exactly as in [22]) on the super-messages and sends the queries to the servers.

Theorem 5.3. *The Partition-and-Coding scheme for Multi-Server satisfies the Privacy and Retrieval conditions and achieves the maximum rate.*

Proof. To see that the rate of the Partition-and-Coding scheme for Multi-Server indeed matches the claimed formula, we start by observing that a PIR coding scheme without side information for Θ *super-messages* requires $1 + \frac{1}{N} + \cdots + \frac{1}{N^{\Theta-1}}$ transmissions [22]. According to step one, the number of *super-messages* always satisfies $\Theta = \lceil \frac{K}{M+1} \rceil$. Thus the rate R of the Partition-and-Coding scheme for Multi-Server satisfies

$$R = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{\lceil \frac{K}{M+1} \rceil - 1}}\right)^{-1}, \tag{5.140}$$

which matches the upper bound of the capacity previously shown in Equation (5.136). Hence, the Partition-and-Coding scheme for Multi-Server achieves the maximum rate.

[3] If the last subset is chosen to contain the demand message, then not all side information messages are required to be placed in the last subset. Otherwise, all side information messages should be placed in the chosen subset.

Next, we show that the *retrieval condition* is satisfied. The PIR coding scheme for multi-server without side information proposed in [22] guarantees that the *super-message* consisting of the demand message and the side information messages can be successfully decoded. Since the *super-message* is the sum of all messages in the subset, given the side information messages in the subset, the only unknown is the demand message. Hence, the user can decode the demand message, which satisfies the *retrieval condition*.

Finally, we show that the *privacy condition* is also satisfied. For *super-messages*, the PIR coding scheme for the multi-server without side information (exactly as in [22]) satisfies the privacy condition, which means the server cannot infer which *super-message* contains the demand message. Let $\wp_i$ denote the subset of messages for constructing *super-message* Y_i. Then, the probability for Y_i to contain the demand message X_W can be computed as

$$\Pr(X_{\mathbf{W}} \in \wp_i | \mathbf{A}, \mathbf{Q}) = \frac{|\wp_i|}{K}, \tag{5.141}$$

where $\mathbf{W}$ denotes the random variable for the index of the demand message and $\mathbf{A}$ and $\mathbf{Q}$ denote the answer strings and queries, respectively. Moreover, each message in the same subset has the same probability to be the demand message, i.e.,

$$\Pr(\mathbf{W} = j | X_{\mathbf{W}} \in \wp_i, \mathbf{A}, \mathbf{Q}) = \frac{1}{|\wp_i|}. \tag{5.142}$$

Hence, $\forall\, i \in \{1, \ldots, \Theta\}$ and $\forall\, X_j \in \wp_i$, the probability for message X_j to be the demand message given the answers and queries is

$$\begin{aligned}
\Pr(\mathbf{W} = j | \mathbf{A}, \mathbf{Q}) &= \Pr(\mathbf{W} = j, X_{\mathbf{W}} \in \wp_i | \mathbf{A}, \mathbf{Q}) && (5.143)\\
&= \Pr(\mathbf{W} = j | X_{\mathbf{W}} \in \wp_i, \mathbf{A}, \mathbf{Q}) \Pr(X_{\mathbf{W}} \in \wp_i | \mathbf{A}, \mathbf{Q}) && (5.144)\\
&= \frac{|\wp_i|}{K} \frac{1}{|\wp_i|} && (5.145)\\
&= \frac{1}{K}. && (5.146)
\end{aligned}$$

Therefore, from the server's perspective, each message has the same probability ($\frac{1}{K}$) to be the demand message, which satisfies the privacy condition. □

Example 5.4. *Consider the multi-server single-message PIR with side information for $K = 5$ messages, $N = 2$ servers and $M = 1$ side information message. The user wants to download X_1 and has X_2 as the side information message. We first show an asymmetric coding scheme that uses different virtual side information for different servers and hence suboptimal. Then we use the Partition-and-Coding scheme for multi-server to construct an optimal code.*

Specifically, consider the following asymmetric coding scheme:

Each message is divided into 4 chunks. It is easy to verify that given X_2 (and its chunks $X_{21}, X_{22}, X_{23}, X_{24}$), the demand message X_1 can be recovered from the answer strings. Two

Server 1	Server 2
$X_{11} + X_{21}$	$X_{12} + X_{22}$
$X_{31} + X_{41}$	$X_{32} + X_{52}$
X_{51}	X_{42}
$X_{13} + X_{23} + X_{32} + X_{42} + X_{52}$	$X_{14} + X_{24} + X_{31} + X_{41} + X_{51}$

chunks of each message are requested from each server and each server individually cannot infer any information about the demand index. The virtual side information for X_3 is different at the two servers. Specifically, for X_3, the virtual side information in Server 1's perspective is X_4, while the virtual side information in Server 2's perspective is X_5. For 4 demand bits, the total number of download bits is 8. Hence, the rate for this coding scheme is $\frac{4}{8} = \frac{1}{2}$, which is suboptimal.

By contrast, the Partition-and-Coding scheme for multi-server proceeds as follows:

1. *The user creates $\Theta = \lceil \frac{K}{M+1} \rceil = 3$ empty subsets, $\wp_1, \wp_2, \wp_3$ of size $2, 2,$ and 1, respectively:*

$$\wp_1 = \{*, *\}, \wp_2 = \{*, *\}, \wp_3 = \{*\}. \tag{5.147}$$

2. *The user randomly selects one subset from $\{\wp_1, \wp_2, \wp_3\}$ to contain the demand message X_1 with probability $\frac{2}{5}, \frac{2}{5}, \frac{1}{5}$, respectively. Suppose the user chooses $\wp_2$. Then, we have:*

$$\wp_1 = \{*, *\}, \wp_2 = \{X_1, *\}, \wp_3 = \{*\}. \tag{5.148}$$

3. *The user puts side information X_2 into the selected subset $\wp_2$, leading to:*

$$\wp_1 = \{*, *\}, \wp_2 = \{X_1, X_2\}, \wp_3 = \{*\}. \tag{5.149}$$

4. *The user randomly distributes the remaining messages to the other subsets. For example, suppose that the outcome of this process is*

$$\wp_1 = \{X_3, X_5\}, \wp_2 = \{X_1, X_2\}, \wp_3 = \{X_4\}. \tag{5.150}$$

5. *For each subset, a super-message is generated*

$$\wp_1 = \{X_3, X_5\}, \qquad Y_1 = X_3 + X_5, \tag{5.151}$$
$$\wp_2 = \{X_1, X_2\}, \qquad Y_2 = X_1 + X_2, \tag{5.152}$$
$$\wp_3 = \{X_4\}, \qquad Y_3 = X_4. \tag{5.153}$$

6. *For the current example with 2 servers and 3 super-messages, we now show that an optimal coding scheme can be implemented already with messages of length $L = 2^3 = 8$*

bits.[4] *That is, we have that* $\forall\, j \in \{1,\dots,8\}$:

$$Y_{1j} = X_{3j} + X_{5j}, \tag{5.154}$$

$$Y_{2j} = X_{1j} + X_{2j}, \tag{5.155}$$

$$Y_{3j} = X_{4j}. \tag{5.156}$$

Following the standard method proposed in [22], we can construct an optimal PIR scheme for multi-server without side information for Y_1, Y_2, Y_3. *The user sends the queries shown as follows to Server 1 and Server 2.*

Server 1	Server 2
Y_{11}, Y_{21}, Y_{31}	Y_{12}, Y_{22}, Y_{32}
$Y_{23} + Y_{12}$	$Y_{25} + Y_{11}$
$Y_{24} + Y_{32}$	$Y_{26} + Y_{31}$
$Y_{13} + Y_{33}$	$Y_{14} + Y_{34}$
$Y_{27} + Y_{14} + Y_{34}$	$Y_{28} + Y_{13} + Y_{33}$

It can be verified that $Y_{21}, \dots, Y_{28}$ *can be recovered from the coding scheme. Hence,* Y_2 *can be fully decoded. Given* X_2 *as the side information, the demand message* X_1 *can also be retrieved. For* 8 *bits message, the total number of download bits is* 14. *Hence, the rate is* $\frac{8}{14} = \frac{4}{7}$, *which is higher than that of the above asymmetric coding scheme and matches the capacity* $\left(1 + \frac{1}{2} + \frac{1}{2^2}\right)^{-1}$.

5.4 Discussions and Conclusion

5.4.1 Models for the Demand Index and Side Information Indices

In Section 5.1, we first define the distribution of the random variable of demand index **W** and then define the distribution of the random variable of side information indices **S** is defined as the conditional distribution given **W**. This way of definitions for **W** and **S** is somehow counter-intuitive and may cause the confusion that if the user can choose the side information indices from some distribution, why don't the user just choose the demand index as side information. For this misunderstanding, we would like to clarify that for any specific private information retrieval with side information problem, neither the demand index nor the side information indices can be chosen by the user. Both **W** and **S** are fixed as realizations W and S at beginning, and are used as the inputs to generate queries. The distributions are assumptions and used for the coding schemes instead of any specific query or answer string.

It is also beneficial to notice that from the distribution for **W** and conditional distribution for **S** given **W** defined in Equation (5.3) and (5.4), respectively, we can derive the marginal

[4] Note that Theorem 5.1 characterizes optimal performance in the limit as the message length L becomes large, as defined in Equation (5.10). For the example at hand, that same performance can be attained already for $L = 8$.

distribution for $\mathbf{S}$ as follows.

$$\Pr(\mathbf{S}=S) = \sum_{W\in\{1,\dots,K\}} \Pr(\mathbf{S}=S,\mathbf{W}=W) \tag{5.157}$$

$$= \sum_{W\in\{1,\dots,K\}} \Pr(\mathbf{S}=S|\mathbf{W}=W)\Pr(\mathbf{W}=W) \tag{5.158}$$

$$= \sum_{W\in\{1,\dots,K\}\setminus S} \Pr(\mathbf{S}=S|\mathbf{W}=W)\Pr(\mathbf{W}=W) + \sum_{W\in S} \Pr(\mathbf{S}=S|\mathbf{W}=W)\Pr(\mathbf{W}=W) \tag{5.159}$$

$$=(K-M)\times\frac{1}{\binom{K-1}{M}}\times\frac{1}{K} \tag{5.160}$$

$$=\frac{1}{\binom{K}{M}}, \tag{5.161}$$

where in Equation (5.159), $\Pr(\mathbf{S}=S|\mathbf{W}=W)=0$ for any $W\in S$. The conditionally distribution for $\mathbf{W}$ given $\mathbf{S}$ satisfies

$$\Pr(\mathbf{W}=W|\mathbf{S}=S) = \frac{\Pr(\mathbf{W}=W,\mathbf{S}=S)}{\Pr(\mathbf{S}=S)} \tag{5.162}$$

$$= \frac{\Pr(\mathbf{S}=S|\mathbf{W}=W)\Pr(\mathbf{W}=W)}{\Pr(\mathbf{S}=S)} \tag{5.163}$$

$$= \frac{\frac{1}{\binom{K-1}{M}}\times\frac{1}{K}}{\frac{1}{\binom{K}{M}}} \tag{5.164}$$

$$= \frac{1}{K-M}. \tag{5.165}$$

The distribution for $\mathbf{W}$ and conditional distribution for $\mathbf{W}$ given $\mathbf{S}$ are more intuitive and can be interpreted as the M side information messages X_S are chosen uniformly at random from K messages and then the demand message is chosen uniformly at random from the remaining $K-M$ messages. However, the two ways of defining the distributions are equivalent. The joint distribution of $\mathbf{W}$ and $\mathbf{S}$ satisfies

$$\Pr(\{\mathbf{W},\mathbf{S}\}=\{W,S\}) = \frac{1}{\binom{K}{M+1}}, \forall S\subset\{1,\dots,K\}, W\in\{1,\dots,K\}\setminus S. \tag{5.166}$$

5.4.2 Virtual Side Information in Multi-Server and Single-Server Cases

In the multi-server PIR with side information problem, we defined the virtual side information for each server and each query. The virtual side information just indicates that the query may possibly be generated for any demand index W with proper side information indices S. A similar concept for single-server PIR with side information is defined in [49], which can be interpreted as the decoding property. For every index $W\in\{1,\dots,K\}$, it is possible to find $S\subseteq\{1,\dots,K\}\setminus\{W\}$ with $|S|=M$ such that given side information $\mathbf{X}_S$, the message

$\mathbf{X}_W$ can be decoded. However, in multi-server cases, the virtual side information does not guarantee similar decoding property. Even when queries for all servers have the same virtual side information for every index, the corresponding answer string may still not permit the decoding of the messages which are neither the demand message nor the side information messages. This is the main difference between multi-server and single-server PIR with side information problems.

Moreover, in multi-server cases, the answer string for each server may not include all messages. This is because the missing message in answer string generated by server i can be used in answer string generated by server j. However, in single-server cases, the answer string must cover all messages. Otherwise, the missing message can be excluded from being the demand message, which violates the *privacy condition*.

5.4.3 Conclusion

In this chapter, we studied the multi-server single-message private information retrieval with side information problem. We characterized the capacity of this problem by presenting the proof of the converse bound for the total number of download bits per demand bit and proposing an achievability scheme to construct optimal codes which satisfy both the *retrieval condition* and *privacy condition*. We introduced the conception, *virtual side information*, which can be utilized in the proof of the converse bound. The tricky part of this problem is that for queries generated for different servers, the virtual side information for those undemanded indices can be different. We have shown that for each group of queries with different virtual side information, it is always possible to generate another group of queries with the same virtual side information and downloads no more bits. The proposed achievability scheme is a linear coding scheme, which implies that linear coding schemes are sufficient to optimally solve multi-server single-message private information retrieval with side information problem.

5.5 Appendix

5.5.1 Proof of Theorem 5.2

In this section, we present the proof for Theorem 5.2.

Proof. For index W_1, denote the corresponding virtual side information at Server j by $S_{1,j}$ for $j \in \{1,\dots,N\}$. According to Lemma 5.5, without loss of optimality, we can assume that $S_{1,j} \cap (W_0 \cup S_0) = \emptyset$, $\forall\, j \in \{1,\dots,N\}$. Without loss of generality, let us assume that $|S_{1,1}| \geq |S_{1,2}| \geq \cdots \geq |S_{1,N}|$. Let the common virtual side information be $S_1 = S_{1,1}$.

For any $S_{1,j}$ with $j \neq 1$, according to Lemma 5.6, it is possible to find a query $\hat{Q}_j$ with virtual

side information indices $\hat{S}_{1,j}$ for W_1 such that

$$\hat{S}_{1,j} \subseteq S_1, \tag{5.167}$$

$$|\hat{S}_{1,j}| = |S_{1,j}|, \tag{5.168}$$

$$H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0\cup S_0}) = H(\mathbf{A}_j^{[W_1,\hat{S}_{1,j}]}|\mathbf{Q}_j^{[W_1,\hat{S}_{1,j}]} = \hat{Q}_j^{[W_1,\hat{S}_{1,j}]}, \mathbf{X}_{W_0\cup S_0}). \tag{5.169}$$

Note that the virtual side information for one index in a query and the corresponding answer string are not necessarily unique. For answer string $\mathbf{A}_j^{[W_1,\hat{S}_{1,j}]}$ and corresponding query realization $Q_j^{[W_1,\hat{S}_{1,j}]}$, since the virtual side information for W_1 satisfies $\hat{S}_{1,j} \subseteq S_1$, S_1 is also a possible virtual side information for W_1. Thus, there must exist query realization $\ddot{Q}_j^{[W_1,S_1]}$ such that

$$H(\mathbf{A}_j^{[W_1,\hat{S}_{1,j}]}|\mathbf{Q}_j^{[W_1,\hat{S}_{1,j}]} = \hat{Q}_j^{[W_1,\hat{S}_{1,j}]}, \mathbf{X}_{W_0\cup S_0}) \geq H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]} = \ddot{Q}_j^{[W_1,S_1]}, \mathbf{X}_{W_0\cup S_0}), \tag{5.170}$$

where the equality holds when $\ddot{Q}_j^{[W_1,S_1]} = \hat{Q}_j^{[W_1,\hat{S}_{1,j}]}$. Therefore, it is always possible to find the common virtual side information S_1 such that $\forall\, j \in \{1,\ldots,N\}$:

$$H(\mathbf{A}_j^{[W_1,S_{1,j}]}|\mathbf{Q}_j^{[W_1,S_{1,j}]} = Q_j^{[W_1,S_{1,j}]}, \mathbf{X}_{W_0\cup S_0}) \geq H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]} = \ddot{Q}_j^{[W_1,S_1]}, \mathbf{X}_{W_0\cup S_0}). \tag{5.171}$$

□

5.5.2 Proof of Lemma 5.7

In this section, we present the proof of Lemma 5.7.

Proof. By assumption, the both groups of answer strings, $\mathbf{A}_1,\ldots,\mathbf{A}_N$ and $\hat{\mathbf{A}}_1,\ldots,\hat{\mathbf{A}}_N$, have virtual side information S_0 for W_1 and S_1 for W_1 in all queries Q_j for $j \in \{1,\ldots,N\}$. According to Lemma 5.1, $\forall\, j \in \{1,\ldots,N\}$:

$$H(\mathbf{A}_j^{[W_0,S_0]}|\mathbf{Q}_j^{[W_0,S_0]} = Q_j^{[W_0,S_0]}, \mathbf{X}_{W_0\cup S_0}) = H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]} = Q_j^{[W_1,S_1]}, \mathbf{X}_{W_0\cup S_0}) \tag{5.172}$$

where $Q_j^{[W_0,S_0]} = Q_j^{[W_1,S_1]} = Q_j$ are the same query. Similarly,

$$H(\hat{\mathbf{A}}_j^{[W_1,S_1]}|\hat{\mathbf{Q}}_j^{[W_1,S_1]} = \hat{Q}_j^{[W_1,S_1]}, \mathbf{X}_{W_0\cup S_0}) = H(\hat{\mathbf{A}}_j^{[W_0,S_0]}|\hat{\mathbf{Q}}_j^{[W_0,S_0]} = \hat{Q}_j^{[W_0,S_0]}, \mathbf{X}_{W_0\cup S_0}) \tag{5.173}$$

where $\hat{Q}_j^{[W_1,S_1]} = \hat{Q}_j^{[W_0,S_0]} = \hat{Q}_j$ are the same query. From Server j's perspective, both Q_j and $\hat{Q}_j$ can possibly be generated for (W_0,S_0) or (W_1,S_1). Thus, Q_j and $\hat{Q}_j$ ($\forall\, j \in \{1,\ldots,N\}$) are actually from the same subset of queries, denoted by $\mathcal{Q}_j^{[W_0,S_0],[W_1,S_1]}$, which has virtual side information S_0 and S_1 for W_0 and W_1, respectively. By taking average over all queries in $\mathcal{Q}_j^{[W_0,S_0],[W_1,S_1]}$, we

can obtain

$$
\begin{aligned}
& H(\mathbf{A}_j^{[W_0,S_0]}|\mathbf{Q}_j^{[W_0,S_0]},\mathbf{X}_{W_0\cup S_0}) \\
=& \mathbb{E}_{Q_j^{[W_0,S_0]}\in\mathcal{Q}_j^{[W_0,S_0],[W_1,S_1]}}\left[H(\mathbf{A}_j^{[W_0,S_0]}|\mathbf{Q}_j^{[W_0,S_0]}=Q_j^{[W_0,S_0]},\mathbf{X}_{W_0\cup S_0})\right] && (5.174)\\
=& \mathbb{E}_{Q_j^{[W_1,S_1]}\in\mathcal{Q}_j^{[W_0,S_0],[W_1,S_1]}}\left[H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]}=Q_j^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0})\right] && (5.175)\\
=& H(\mathbf{A}_j^{[W_1,S_1]}|\mathbf{Q}_j^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}) && (5.176)\\
=& H(\hat{\mathbf{A}}_j^{[W_1,S_1]}|\hat{\mathbf{Q}}_j^{[W_1,S_1]},\mathbf{X}_{W_0\cup S_0}) && (5.177)
\end{aligned}
$$

For the queries in $\mathcal{Q}_j^{[W_0,S_0],[W_1,S_1]}$, Server j can only infer the virtual side information for W_0 and W_1, but should not be able to infer which one of W_0 and W_1 is the real demand index and which one of S_0 and S_1 is the real side information indices. Thus, without loss of optimality, we can assume that for every $Q_1,\ldots,Q_N$ there exist $\hat{Q}_1,\ldots,\hat{Q}_N$ which satisfy Equation (5.111). □

6 Multi-User Private Information Retrieval with Side Information

In Chapter 4 and Chapter 5, we studied the private information retrieval with side information for multi-messages and multi-server extensions, respectively. In this chapter, we investigate another extension, the multi-user private information retrieval with side information. We consider the scenario where multiple users cooperatively download one message from the single server while keeping the index of the demand message private from the server. We consider the linear cases where answer string is assumed to be linear combinations of the messages. We establish the capacity by providing the proof for converse and proposing an achievability scheme.

6.1 Problem Statement

In the multi-user private information retrieval with side information problem, there is a database which consists of K messages, denoted by $X_{1:K} = \{X_1, \ldots, X_K\}$, and is stored in a single server. The random variables of the messages, denoted by $\mathbf{X}_i$'s, for $i \in \{1, \ldots, K\}$, are assumed to be independent from each other, i.e.,

$$H(\mathbf{X}_1, \ldots, \mathbf{X}_K) = H(\mathbf{X}_1) + \cdots + H(\mathbf{X}_K). \tag{6.1}$$

We assume that there are N users that want to cooperatively retrieve a common message $X_W \in \{X_1, \ldots, X_K\}$. We refer to $W \in \{1, \ldots, K\}$ as the demand index and X_W as the demand message. Let $\mathbf{W}$ denote the random variable of the demand index W. $\mathbf{W}$ is assumed to be uniformly distributed over $\{1, \ldots, K\}$, i.e.,

$$\Pr(\mathbf{W} = W) = \frac{1}{K}, \qquad \forall W \in \{1, \ldots, K\}. \tag{6.2}$$

Additionally, it is assumed that each user initially has a subset of messages as side information. For each user $i \in \{1, \ldots, N\}$, let S_i denote the set of the indices of the side information messages of user i. We refer to $S_i \subseteq \{1, \ldots, K\} \setminus \{W\}$ as the side information indices of user i and $X_{S_i} = \{X_j : j \in S_i\}$ as the side information messages of user i. Let M_i denote the number of side

information messages of user i, i.e., $|S_i| = M_i$. Let $\mathbf{S}_i$ denote the random variable for S_i for each $i \in \{1,\dots,,N\}$, which is assumed to be conditionally uniformly distributed over all subsets of $\{1,\dots,K\} \setminus \{W\}$ with size M_i, i.e.,

$$\Pr(\mathbf{S}_i = S_i) = \frac{1}{\binom{K-1}{M_i}}, \qquad \forall S_i \subseteq \{1,\dots,K\} \setminus \{W\}, |S_i| = M_i. \tag{6.3}$$

We note that different users can have different numbers of side information messages. We assume that the server only knows the numbers of the side information messages of all users $(M_1,\dots,M_N)$, but does not know the actual indices of side information messages $(S_1,\dots,S_N)$.

The goal of the users is to download the demand message X_W from the server while revealing no information about W to the server. To achieve the goal, the users jointly generate and send a query Q to the server. Let $\mathbf{Q}$ denote the random variable for query realization Q which is generated for retrieving message X_W while having $X_{S_1},\dots,X_{S_N}$ as side information at each user, respectively. Following the literature, we assume that $\mathbf{Q}$ is a (stochastic) function of the demand index W and all side information indices $S_1,\dots,S_N$, but does not depend on the contents of any of the messages, i.e.,

$$H(\mathbf{Q}|\mathbf{X}_{1:K}) = H(\mathbf{Q}). \tag{6.4}$$

After the server receives the query Q, it generates and replies the corresponding answer string A to the users. Let $\mathbf{A}$ denote the random variable for the answer string realization A, which is a deterministic function of query $\mathbf{Q}$ and all messages $\mathbf{X}_1,\dots,\mathbf{X}_K$, i.e.,

$$H(\mathbf{A}|\mathbf{Q},\mathbf{X}_{1:K}) = 0. \tag{6.5}$$

We only consider the linear code scheme, where the answer string is assumed to be linear combinations of messages. The query $\mathbf{Q}$ is chosen from an alphabet $\mathscr{Q}$ and the answer string $\mathbf{A}$ is from a corresponding alphabet $\mathscr{A}$. The PIR scheme is the set of queries and answer strings.

6.1.1 Retrieval and Privacy Conditions

For any fixed W and $S_1,\dots,S_N$, the user jointly generate one query Q (from potentially multiple queries) and request the corresponding answer string $\mathbf{A}$ from the server. In order to let every user successfully recover the demand message $\mathbf{X}_W$, the answer string $\mathbf{A}$ and query $\mathbf{Q}$ must satisfy:

$$H(\mathbf{X}_W|\mathbf{A},\mathbf{Q},\mathbf{X}_{S_i}) = 0, \qquad \forall i \in \{1,\dots,N\} \tag{6.6}$$

We refer to Condition (6.6) as the *retrieval condition* for multi-user private information retrieval with side information.

The private information retrieval also requires that the server should not be able to infer any

information about the index of the demand message from the received query. Thus, the query $\mathbf{Q}$ must satisfy:

$$I(\mathbf{W};\mathbf{Q}) = 0. \tag{6.7}$$

As we have shown in Chapter 5, Equation (6.7) can be used to derive the following equation.

$$I(\mathbf{W};\mathbf{A},\mathbf{Q},\mathbf{X}_{1:K}) = 0. \tag{6.8}$$

We refer to Condition (6.8) as the *privacy condition* for multi-user private information retrieval.

6.1.2 Definitions and Useful Lemma

For each specific multi-user PIR with side information problem, we can use a matrix to represent all the information that we need.

Definition 6.1 (**Characterization Matrix**)**.** *For the multi-user private information retrieval with side information problem with demand index W and side information indices $S_1,\ldots,S_N$, define the characterization matrix C with entry $C_{i,j}$ $(\forall\, i \in \{1,\ldots,N\}, j \in \{1,\ldots,K\})$:*

$$C_{i,j} = \begin{cases} 1, & \text{if } j = W, \\ \alpha, & \text{if } j \in S_i, \\ 0, & \text{otherwise.} \end{cases} \tag{6.9}$$

For $V \subseteq \{1,\ldots,K\}$, let C^V denote the submatrix of C with columns indexed by V. Let $w_\alpha(C(i,:))$ denote the number of α's in the i-th row vector of C.

The characterization matrix contains 3 entries, which are 1, α and 0 representing demand index, side information indices and other indices, respectively. The number of columns is equal to the total number of messages K. The number of rows is equal to the number of users N. Each row carries the information for the corresponding user.

Example 6.1. *Consider a multi-user private information retrieval with side information problem with the setting: $K = 7$, $N = 3$, $W = 1$, $S_1 = \{2,3\}$, $S_2 = \{3,4,5\}$, and $S_3 = \{2,4\}$. Then we can use the following matrix to represent the problem with this specific setting.*

$$C = \begin{bmatrix} 1 & \alpha & \alpha & 0 & 0 & 0 & 0 \\ 1 & 0 & \alpha & \alpha & \alpha & 0 & 0 \\ 1 & \alpha & 0 & \alpha & 0 & 0 & 0 \end{bmatrix}. \tag{6.10}$$

Since we only consider the linear coding schemes, the answer string is the set of linear combinations of messages. Suppose there are R linear combinations in answer string $\mathbf{A} = \{\mathbf{T}_1,\ldots,\mathbf{T}_R\}$. We note that each linear combination T_r for $r \in \{1,\ldots,R\}$ may not use all messages. In other

words, the coefficients of some messages may be zeros. Hence, we define the coding subspace for the messages which are used to generate each linear combination as follows.

Definition 6.2 (**Coding Subspace**). *For any linear answer string* $\mathbf{A} = \{\mathbf{T}_1, \ldots, \mathbf{T}_R\}$*, let* $supp(\mathbf{T}_r)$ *denote the messages which are used to generated* $\mathbf{T}_r$ *for any* $r \in \{1, \ldots, R\}$*. Define the partition of the messages according to the answer string* $\mathbf{A}$ *as* $\mathscr{P}(\mathbf{A}) = \{\wp_1, \ldots\}$ *such that* $\forall \mathbf{T}_r \in \{\mathbf{T}_1, \ldots, \mathbf{T}_R\}$, $!\exists \wp_j \in \mathscr{P}(\mathbf{A})$ *such that* $supp(\mathbf{T}_i) \subseteq \wp_j$*. We call the subspace spanned by messages in* $\wp_j$ *a coding subspace, and in slight abuse of notation, use the same symbol* $\wp_j$ *to denote this subspace. Let* $\mathbf{T}(\wp_j) = \{\mathbf{T}_r \in \{\mathbf{T}_1, \ldots, \mathbf{T}_R\} : supp(\mathbf{T}_r) \subseteq \wp_j\}$.

Regarding the concept of coding subspace, it is good to notice that:

1. For each linear combination $\mathbf{T}_r \in \{\mathbf{T}_1, \ldots, \mathbf{T}_R\}$, we can easily identify which messages are used to generate this linear combination. Hence, it is always possible to partition the messages into subsets of messages such that each linear combination only consists of messages from a single subset.
2. The set of all messages is always a valid coding subspace for all linear answer strings.
3. Linear combinations in different coding subspace have no commonly used messages. They are completely independent of each other and cannot help each other to decode any messages.

Definition 6.3 (**Decoding Pattern**). *The set of side information messages* $\mathbf{X}_S$ *is called a* decoding pattern of message $\mathbf{X}_i$ *if* $\mathbf{X}_S$ *and* $\mathbf{X}_i$ *belong to the same coding subspace (*$\wp$*) and given* $\mathbf{X}_S$*,* $\mathbf{X}_i$ *can be decoded from the answer string requested by the users. Hence S should satisfy* $H(\mathbf{X}_i|\mathbf{T}(\wp), \mathbf{X}_S) = 0$.

For any answer string $\mathbf{A}$ satisfying the *privacy condition*, it is always possible to find one (or multiple) decoding pattern(s) for every message. Otherwise, if one message has no decoding pattern, which implies that message cannot be the demand message and violates the *privacy condition*.

Definition 6.4 (**MDS-Condition**). *A linear answers string* $\mathbf{A}$ *satisfies the MDS-Condition in coding subspace* $\wp_i \in \mathscr{P}(\mathbf{A})$ *if either of the following two conditions is satisfied:*

(i) The normalized number of download bits is equal to the size of the coding subspace, i.e.,

$$|\mathbf{T}(\wp_i)| = |\wp_i|. \tag{6.11}$$

(ii) The normalized number of download bits is equal to the size of coding subspace minus M_i*, i.e.,* $\exists M_i \in \{1, \ldots, |\wp_i| - 1\}$ *such that*

$$|\mathbf{T}(\wp_i)| = |\wp_i| - M_i, \tag{6.12}$$

and given any subset of messages in $\wp_i$ with size M_i, the other messages in $\wp_i$ can be fully decoded, i.e.,

$$H(\wp_i|\mathbf{T}(\wp_i),\mathbf{X}_V)=0,\forall \mathbf{X}_V\subset\wp_i,|\mathbf{X}_V|=M_i. \tag{6.13}$$

Additionally, no message in $\wp_i$ can be decoded, given less than M_i side information, i.e., $\forall \mathbf{X}_U\subset\wp_i$ with $|\mathbf{X}_U|\leq M_i-1$, $\forall \mathbf{X}_j\in\wp_i\setminus\mathbf{X}_U$:

$$H(\mathbf{X}_j|\mathbf{T}(\wp_i),\mathbf{X}_U)\neq 0. \tag{6.14}$$

We note that the MDS-Condition is defined for an answer string and one of its coding subspaces. It can be conveniently verified for any answer string $\mathbf{A}$ and coding subspace $\wp_i$. If $\mathbf{A}$ satisfies the condition (i), then the linear combinations $\mathbf{T}(\wp_i)$ are equivalent to sending each message in $\wp_i$ individually without any coding. If $\mathbf{A}$ satisfies the condition (ii), then the linear combinations $\mathbf{T}(\wp_i)$ can be used to decode any messages in $\wp_i$ as long as M_i messages are given as side information. This property is closely related to Maximum Distance Separable (MDS) codes. So we name it as the MDS-Condition.

Definition 6.5. *For any characterization matrix C and any $L\in\{1,\ldots,K\}$, define $\mathscr{R}(L)$ as*

$$\mathscr{R}_C(L)=L-\max_{V\subseteq\{1,\ldots,K\}\setminus\{W\},|V|=L-1}\ \min_{i\in\{1,\ldots,N\}} w_\alpha(C^V(i,:)), \tag{6.15}$$

where W is the index of the demand message.

For any fixed L, we select $L-1$ of columns and the demand message column from C to form a submatrix. For the selected submatrix of C, we can compute the number of α's for any row i and find the row with the minimum number of α's. The optimal selection of such $L-1$ columns should maximize the minimum number of α's in any row vector of the submatrix. And $\mathscr{R}_C(L)$ is defined as L minus the maximized minimal number of α's. We note that the optimal choices for V and i are not necessarily unique.

Example 6.2. *Consider the single-server multi-user private information retrieval with side information problem with the following characterization matrix:*

$$C=\begin{bmatrix}1&\alpha&\alpha&0&0&0&0\\1&0&\alpha&\alpha&\alpha&0&0\\1&\alpha&0&\alpha&0&0&0\end{bmatrix}. \tag{6.16}$$

We can try all possible selections of columns for any $L\in\{1,\ldots,7\}$ and get the following results

1. *$L=1:\mathscr{R}_C(L)=L$.*
2. *$L=2:\mathscr{R}_C(L)=L$ with optimal $V=\{2\}$ and $i=2$.*
3. *$L=3:\mathscr{R}_C(L)=L-1$ with optimal $V=\{2,3\}$ and $i=2$.*

4. $L \geq 4 : \mathscr{R}_C(L) = L-2$ *with optimal* $V \supseteq \{2,3,4\}$ *and* $i = 1$.

However, this traversal method requires high complexity. We show the detail of how to efficiently compute $\mathscr{R}_C(L)$ *in Section 6.3. Let us consider the following linear answer string* $\mathbf{A} = \{\mathbf{T}_1, \ldots, \mathbf{T}_4\}$:

$$\mathbf{T}_1 = \mathbf{X}_2 + \mathbf{X}_5 + \mathbf{X}_6 + \mathbf{X}_7, \tag{6.17}$$

$$\mathbf{T}_2 = \mathbf{X}_2 + 2\mathbf{X}_5 + 3\mathbf{X}_6 + 4\mathbf{X}_7, \tag{6.18}$$

$$\mathbf{T}_3 = \mathbf{X}_1 + \mathbf{X}_3 + \mathbf{X}_4, \tag{6.19}$$

$$\mathbf{T}_4 = \mathbf{X}_1 + 2\mathbf{X}_3 + 3\mathbf{X}_4. \tag{6.20}$$

It can be verified that the answer string $\mathbf{A}$ *satisfies the MDS-Condition for coding subspaces* $\wp_1 = \{\mathbf{X}_2, \mathbf{X}_5, \mathbf{X}_6, \mathbf{X}_7\}$ *and* $\wp_2 = \{\mathbf{X}_1, \mathbf{X}_3, \mathbf{X}_4\}$. *From* $\mathbf{T}_3$ *and* $\mathbf{T}_4$, *given any one message in* $\wp_2$, *the other two messages can be decoded. Since users have at least one message of* $\{\mathbf{X}_3, \mathbf{X}_4\}$, *they can successfully decode the demand message* $\mathbf{X}_1$. *Hence, answer string* $\mathbf{A}$ *satisfies the retrieval condition. Also, it can be computed that* $\mathscr{R}_C(4) = L-2 = 2$ *and* $\mathscr{R}_C(3) = L-1 = 2$, *which are the number of linear combinations in* $\wp_1$ *and* $\wp_2$, *respectively.*

In each coding subspace, the number of decoding patterns for each message is the same. But since the server has no information about the side information messages of the users, the demand message can be possible in either of the two coding subspaces. By using the randomized coding technique discussed in Section 6.2.2, the probability of the demand message to be placed in $\wp_i$ *is* $|\wp_i|/K$. *Then, every message can be the demand message with equal probability which is* $1/K$. *Hence, the privacy condition is also satisfied.*

As one may notice, only $\mathbf{T}_3$ *and* $\mathbf{T}_4$ *are useful in decoding the demand message* $\mathbf{X}_1$. *Why is it necessary to have two transmissions in* $\wp_1$*? The reason is* $\mathscr{R}_C(|\wp_1|) = 2$ *and we show that* $|\mathbf{T}(\wp_1)| \geq \mathscr{R}_C(|\wp_1|)$ *is a necessary condition for users to put the demand message in* $\wp_i$ *in Lemma 6.2.*

6.2 The Capacity

Theorem 6.1. *For the single-server multi-user private information retrieval with side information problem where all users want the same message but have different side information, the minimal number of required linear combinations satisfies*

$$R^* = \min_{\mathscr{L} \in \Pi(K)} \sum_{l \in \mathscr{L}} \mathscr{R}_C(l), \tag{6.21}$$

where $\Pi(K)$ *denotes the set of partitions of* K.

Proof. We present the proof of the converse and achievability for Theorem 6.1 in Section 6.2.1 and Section 6.2.2, respectively. □

We have the following remarks regarding Theorem 6.1.

1. The minimum number of required linear combinations can be obtained by solving an optimization problem over all partitions of the total number of messages, K. Although the number of partitions of K grows exponentially with K, the optimization problem can be solved by a dynamic programming algorithm shown in Section 6.3 with polynomial complexity.
2. The minimum number of required linear combinations is the sum of the number of linear combinations in each coding subspace.

6.2.1 Converse

In this section, we present the proof for the converse for Theorem 6.1. We need to show that

$$R^* \geq \min_{\mathscr{L}\in\Pi(K)} \sum_{l\in\mathscr{L}} \mathscr{R}_C(l). \tag{6.22}$$

For any answer string which is a set of linear combinations of messages from $\{\mathbf{X}_1,\ldots,\mathbf{X}_K\}$, it is always possible to find the corresponding coding subspaces according to Definition 6.2. To satisfy the *privacy condition*, a necessary condition can be derived for each coding subspace, which is stated by the following Lemma.

Lemma 6.1. *For any linear answer string which satisfies the* privacy condition *of single-server multi-user private information retrieval with side information, without loss of optimality, the MDS-Condition should be satisfied in* every *coding subspace.*

Proof. For any answer string $\mathbf{A}$ from a linear PIR coding scheme which satisfies the *privacy condition*, the server should not be able to infer any information about the demand index from the answer string and query. Equation (6.8) is equivalent to

$$H(\mathbf{W}|\mathbf{A},\mathbf{Q},\mathbf{X}_{1:K}) = H(\mathbf{W}). \tag{6.23}$$

Since $\mathbf{W}$ is uniformly distributed over all indices $\{1,\ldots,K\}$, the entropy $H(\mathbf{W})$ achieves the maximum entropy. Each message should have the same probability to be the demand message. For any $W\in\{1,\ldots,K\}$, we have

$$\Pr(\mathbf{W}=W|\mathbf{A}=A,\mathbf{Q}=Q,\mathbf{X}_{1:K}=X_{1:K}) = \Pr(\mathbf{W}=W) = \frac{1}{K}. \tag{6.24}$$

Therefore, for any coding subspace $\wp\in\mathscr{P}(\mathbf{A})$, the messages in $\wp$ should also have the same probability to be the demand message.

Suppose the MDS-condition is not satisfied in coding subspace $\wp$, then there must exist one message $\mathbf{X}_i\in\wp$, such that $\mathbf{X}_i$ can be decoded from $\mathbf{T}(\wp)$ given side information messages either $\mathbf{X}_{S_a}$ or $\mathbf{X}_{S_b}$. Additionally given $\mathbf{X}_{S_a}$, $\mathbf{X}_{S_b}$ cannot be decoded from $\mathbf{T}(\wp)$.

In such cases, if all users have $\mathbf{X}_{S_a}$ as side information messages, which permits successful decoding of $\mathbf{X}_i$, then the messages in $\wp$ which cannot be decoded from $\mathbf{T}(\wp)$ given $\mathbf{X}_{S_a}$ can be separated from $\wp$ to form another coding subspace by setting the coefficients of the messages that can be decoded given $\mathbf{X}_{S_a}$ (including $\mathbf{X}_{S_a}$ themselves) to be zeros in the corresponding linear combinations. And in the split two coding subspaces, the MDS-condition is satisfied. If not all users have $\mathbf{X}_{S_a}$ as side information messages, we cannot split $\wp$ into two coding subspaces, since some users have to use the side information messages $\mathbf{X}_{S_b}$ to decode the message $\mathbf{X}_i$. Hence, the server can infer that $\mathbf{X}_{S_a}$ and $\mathbf{S}_{S_b}$ are side information messages for different users and $\mathbf{X}_i$ is the demand message, which violates the *privacy condition*. Therefore, without loss of optimality, we can assume that the MDS-condition is satisfied in every coding subspace. □

According to Lemma 6.1, it is sufficient to only consider the linear coding schemes with answer strings satisfying MDS-Condition in every coding subspace.

Example 6.3. *Consider a two user private information retrieval with side information problem with setting:* $W = 3$ *and* $S_1 = \{1,2\}$, $S_2 = \{4,5\}$. *If we want to generate linear answer string in coding subspace* $\wp = \{1,2,3,4,5\}$ *which satisfies the retrieval condition for both users, we need at least two linear combinations:*

$$\mathbf{T}_1 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 \tag{6.25}$$

$$\mathbf{T}_2 = \mathbf{X}_3 + \mathbf{X}_4 + \mathbf{X}_5 \tag{6.26}$$

It can be verified that $\mathbf{T}_1, \mathbf{T}_2$ *does not satisfy the MDS-Condition in* $\wp$. *After receiving the query for requesting* $\mathbf{T}_1, \mathbf{T}_2$, *the server knows that the decoding patterns for* $\mathbf{X}_3$ *are* $\{\mathbf{X}_1, \mathbf{X}_2\}$ *and* $\{\mathbf{X}_4, \mathbf{X}_5\}$. *If the both users use decoding pattern* $\{\mathbf{X}_1, \mathbf{X}_2\}$ *to decode* $\mathbf{X}_3$, *then the following linear combinations should also satisfy the retrieval condition*

$$\mathbf{T}'_1 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 \tag{6.27}$$

$$\mathbf{T}'_2 = \mathbf{X}_4 + \mathbf{X}_5 \tag{6.28}$$

It can be verified that $\mathbf{T}'_1$ *and* $\mathbf{T}'_2$ *satisfy the MDS-Condition in coding subspaces* $\wp_1 = \{1,2,3\}$ *and* $\wp_2 = \{4,5\}$, *respectively. If two users use different coding patterns to decode* $\mathbf{X}_3$, *then* $\mathbf{T}'_1$ *and* $\mathbf{T}'_2$ *do not satisfy the MDS-Condition any more, since user* 2 *cannot decode* $\mathbf{X}_3$. *From the server's perspective, it can infer that the two users have different side information and* $\mathbf{X}_3$ *is the demand message, which violates the privacy condition. We can also generate the linear combinations which satisfy the MDS-Condition in coding subspace* $\wp = \{1,2,3,4,5\}$ *as follows.*

$$\mathbf{T}''_1 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 \tag{6.29}$$

$$\mathbf{T}''_2 = \mathbf{X}_1 + 2\mathbf{X}_2 + \mathbf{X}_4 \tag{6.30}$$

$$\mathbf{T}''_3 = \mathbf{X}_1 + 3\mathbf{X}_2 + \mathbf{X}_5 \tag{6.31}$$

The MDS-Condition guarantees that messages in the same coding subspaces have equal probability to be the demand message and have the same number of decoding patterns. Messages in coding subspaces with different dimensions may still have different numbers of decoding patterns. Nevertheless, it is not necessary for them to have the same number of decoding patterns as long as the demand message can be possibly put in any coding subspace. Since only one coding subspace can contain the demand message, if the demand message can be randomly placed into any coding subspace with the probability proportional to the dimension of the coding subspace, the probability for every message to be placed in any subspace is the same. To make sure that the users can randomly put the demand message in any coding subspace, we need the coding scheme to satisfy the condition stated as the following lemma.

Lemma 6.2. *To generate a linear answer string* $\mathbf{A}$ *which satisfies MDS-Condition in coding subspace* $\wp_i \in \mathscr{P}(\mathbf{A})$ *with* $|\wp_i| = L$*, if* $H(\mathbf{X}_W|\mathbf{T}(\wp_i),\mathbf{X}_{S_j}) = 0$ *for all* $j \in \{1,\ldots,N\}$*, then* $|\mathbf{T}(\wp_i)| \geq \mathscr{R}_C(L)$.

Proof. If $\forall\, j \in \{1,\ldots,N\}$: $H(\mathbf{X}_W|\mathbf{T}(\wp_i),\mathbf{X}_{S_j}) = 0$, then all users can decode the demand message from the linear combinations consisting of messages in coding subspace $\wp_i$. It implies that $\mathbf{X}_W \in \wp_i$. Since the answer string $\mathbf{A}$ satisfies MDS-condition in coding subspace $\wp_i$, there are two cases to discuss. If $|\mathbf{T}(\wp_i)| = |\wp_i| = L$, then the number of linear combinations is equal to the number of messages in the coding subspace. Hence, $H(\mathbf{X}_W|\mathbf{T}(\wp_i),\mathbf{X}_{S_j}) = 0$ is always satisfied and $|\mathbf{T}(\wp_i)| \geq \mathscr{R}(L)$, since $\mathscr{R}(L) \leq L$. If $|\mathbf{T}(\wp_i)| = |\wp_i| - M_i < |\wp_i|$, to satisfy $H(\mathbf{X}_W|\mathbf{T}(\wp_i),\mathbf{X}_{S_j}) = 0$, $\forall\, j \in \{1,\ldots,N\}$, we need $|\mathbf{X}_{S_j} \cap \wp_i| \geq M_i$, $\forall\, j \in \{1,\ldots,N\}$. It is equivalent to $\min_{j\in\{1,\ldots,N\}} |\mathbf{X}_{S_j} \cap \wp_i| \geq M_i$. Let V denote the indices of messages in $\wp_i$, then $\min_{j\in\{1,\ldots,N\}} |\mathbf{X}_{S_j} \cap \wp_i| = \min_{j\in\{1,\ldots,N\}} w_H(C^V(j,:))$ is the minimum number of α's in any row vector of the submatrix C^V. Hence,

$$\begin{aligned}
\mathscr{R}_C(L) &= L - \max_{V\subseteq\{1,\ldots,K\}\setminus W, |V|=L-1} \min_{j\in\{1,\ldots,N\}} w_H(C^V(j,:)) && (6.32)\\
&\leq L - \min_{j\in\{1,\ldots,N\}} w_H(C^V(j,:)) && (6.33)\\
&= |\wp_i| - \min_{j\in\{1,\ldots,N\}} |\mathbf{X}_{S_j} \cap \wp_i| && (6.34)\\
&\leq |\mathbf{T}(\wp_i)|. && (6.35)
\end{aligned}$$

By selecting the columns V to be the set of indices of the optimal coding subspace and $M_i = \min_{j\in\{1,\ldots,N\}} |\mathbf{X}_{S_j} \cap \wp_i|$, both inequalities are tight. □

Corollary 6.1 (Converse for Theorem 6.1)**.** *For the single-server multi-message private information retrieval with side information problem, the minimum number of required linear combinations is lower bounded by* $\min_{\mathscr{L}\in\Pi(K)} \sum_{l\in\mathscr{L}} \mathscr{R}_C(l)$.

Proof. According to Lemma 6.1, for any linear coding scheme which satisfies the *privacy condition*, without loss of optimality, we can assume that it also satisfies the MDS-Condition

in every coding subspace. According to Lemma 6.2, for linear combinations consisting of messages in coding subspace with size L, the minimum number of required linear combinations is $\mathscr{R}_C(L)$. Although only one coding subspace can contain the demand message, in order to guarantee that every coding subspace can possibly be used to contain the demand message, the number of linear combinations in any coding subspace with dimension L should be at least $\mathscr{R}(L)$ even if they are not used to transmit the demand message. Therefore, for any partition $\mathscr{L} \in \Pi(K)$, the $R^*(\mathscr{L}) = \sum_{l\in\mathscr{L}} \mathscr{R}_C(l)$. Hence, $R^* \geq \min_{\mathscr{L}\in\Pi(K)} \sum_{l\in\mathscr{L}} \mathscr{R}_C(l)$. □

6.2.2 Achievability

In this section, we present the proof of the achievability for Theorem 6.1 by constructing a linear coding scheme with R^* transmissions which satisfies both the *retrieval condition* and the *privacy condition*.

For any $\{l_1,\ldots,l_G\} \in \Pi(K)$, the users can compute $\mathscr{R}_C(l_i)$ and the $V^*(l_i)$ according to Definition 6.5 for all $i \in \{1,\ldots,G\}$.

Step 1: The users create a set of G subsets, denoted by $\{\wp_1,\ldots,\wp_G\}$, where $|\wp_i| = l_i$.

Step 2: The users randomly pick one subset (e.g. $\wp_i$) to contain the demand message with probability proportional to the sizes of the subsets, i.e.,

$$\Pr(\mathbf{X}_W \in \wp_i) = \frac{|\wp_i|}{K}, \qquad \forall \wp_i \in \{\wp_1,\ldots,\wp_G\}. \tag{6.36}$$

And the users fill up subset $\wp_i$ with side information messages indexed by $V^*(|\wp_i|)$.

Step 3: The users uniformly and randomly distribute other messages into the unchosen subsets.

Step 4: The users send query to the server to ask for the coding scheme which satisfies MDS-Condition in each coding subspace $\wp_j$ ($\forall\, j \in \{1,\ldots,G\}$) with $\mathscr{R}_C(|\wp_j|)$ linear combinations.

We name the above coding scheme as Partition-and-MDS-Coding scheme.

Remark 6.1. *In [42], Kadhe et al. proposed a Partition and Coding PIR scheme for single-sever single-user private information retrieval with side information. In their scheme, messages are partitioned into subsets with size $M+1$[1], where M is the number of side information messages. For each coding subspace, the server sends only one linear combination which is the sum of all messages in the coding subspace. Our coding scheme is an extended coding scheme designed for multi-user cases and includes their coding scheme as the special cases for single-user.*

[1] In the cases where K is not divided by $M+1$, the last subset has size less than $M+1$.

Lemma 6.3 (Achievability)**.** *For any $\mathscr{L} \in \Pi(K)$, the Partition-and-MDS-Coding scheme satisfies th retrieval condition and privacy condition and the number of required linear combinations is $\sum_{l\in\mathscr{L}} \mathscr{R}_C(l)$.*

Proof. In our Partition-and-MDS-Coding scheme, all users can decode the demand message from linear combinations in the coding subspace which is chosen to contain the demand message. Hence the *retrieval condition* is satisfied. To show the *privacy condition* is also satisfied, it is sufficient to show that the conditional probability of one message to be the demand message given the query and answer string is equal to the prior probability. Due to the random choice of the subset to carry the demand message, we have $\Pr(\mathbf{X}_W \in \wp_i | Q(W, S_1, S_2, \ldots, S_N)) = |\wp_i|/K$. Since in each coding subspace, the linear combinations satisfy the MDS-Condition, every message in the same coding subspace is the demand message with equal probability. $\Pr(\mathbf{W} = W | \mathbf{X}_W \in \wp_i, Q(W, S_1, \ldots, S_N)) = \frac{1}{|\wp_i|}$. By conditional probability, we have

$$
\begin{aligned}
&\Pr(\mathbf{W} = W | Q(W, S_1, \ldots, S_N)) \\
=&\Pr(\mathbf{W} = W, \mathbf{X}_W \in \wp_i | Q(W, S_1, \ldots, S_N)) && (6.37)\\
=&\Pr(\mathbf{W} = W | \mathbf{X}_W \in \wp_i, Q(W, S_1, \ldots, S_N)) \Pr(\mathbf{X}_W \in \wp_i | Q(W, S_1, \ldots, S_N)) \\
=&\frac{|\wp_i|}{K} \frac{1}{|\wp_i|} && (6.38)\\
=&\frac{1}{K}. && (6.39)
\end{aligned}
$$

Therefore, the *privacy condition* is also satisfied. The required number of linear combinations is the sum number of the linear combinations in each coding subspace, which is $\sum_{l\in\mathscr{L}} \mathscr{R}_C(l)$.

□

Example 6.4. *Consider the single-server multi-user private information retrieval with side information problem with the following characterization matrix:*

$$
C = \begin{bmatrix} 1 & \alpha & \alpha & 0 & 0 & 0 & 0 \\ 1 & 0 & \alpha & \alpha & \alpha & 0 & 0 \\ 1 & \alpha & 0 & \alpha & 0 & 0 & 0 \end{bmatrix}, \tag{6.40}
$$

Let us generate the Partition-and-MDS-Coding scheme for partition $\mathscr{L} = \{4, 3\}$. According to Equation (6.15)*, we can obtain*

$$
\begin{aligned}
\mathscr{R}_C(4) =& 2 \textit{ with} && V^*(4) = \{2, 3, 4\} && (6.41)\\
\mathscr{R}_C(3) =& 2 \textit{ with} && V^*(3) = \{3, 4\} && (6.42)
\end{aligned}
$$

We note that $V^*(3)$ is not unique. It can also be $\{2, 3\}$. We create two subsets $\wp_1$ and $\wp_2$ with size 4 and 3, respectively. Then we can randomly choose $\wp_1$ with probability $\frac{4}{7}$ or $\wp_2$ with probability $\frac{3}{7}$ to contain the demand message $\mathbf{X}_1$. Suppose we choose the second subset $\wp_2$, then $\mathbf{X}_1$ and $\mathbf{X}_{V^*(3)} = \{\mathbf{X}_3, \mathbf{X}_4\}$ should be placed in $\wp_2$. For subset $\wp_1$, fill it up with the remaining messages.

Now we have two coding subspaces:

$$\wp_1 = \{X_2, X_5, X_6, X_7\}, \tag{6.43}$$
$$\wp_2 = \{X_1, X_3, X_4\}. \tag{6.44}$$

Then we can generate the query to ask for the following answer string.

$$T_1 = X_2 + X_5 + X_6 + X_7 \tag{6.45}$$
$$T_2 = X_2 + 2X_5 + 3X_6 + 4X_7 \tag{6.46}$$
$$T_3 = X_1 + X_3 + X_4 \tag{6.47}$$
$$T_4 = X_1 + 2X_3 + 3X_4, \tag{6.48}$$

It satisfies the MDS-Condition in $\wp_1$ with $\mathscr{R}_C(4)$ linear combinations and in $\wp_2$ with $\mathscr{R}_C(3)$ linear combinations.

From the server's perspective, $\Pr(\mathbf{X_W} \in \wp_1) = \frac{4}{7}$ *and* $\Pr(\mathbf{X_W} \in \wp_2) = \frac{3}{7}$. *Furthermore, in each coding subspace, every message is equally likely to be the demand message. Thus,* $\Pr(\mathbf{W} = W|Q(W, S_1, \ldots, S_N)) = \frac{1}{7} = \Pr(\mathbf{W} = W)$. *The server cannot infer any information about the demand index.*

6.3 Solving the Optimization

In Theorem 6.1, given $\mathscr{R}_C(L)$ for all $L \in \{1, \ldots, K\}$, the minimum number of required linear combinations can be obtained by solving the optimization problem over all partitions of the total number of messages. Instead of trying every possible partition, we can efficiently find the optimal one by using a dynamic programming algorithm. In this section, we present the algorithms for computing $\mathscr{R}_C(L)$ and searching for the optimal decomposition with the minimum number of request linear combinations.

6.3.1 Computing $\mathscr{R}_C(L)$

For any fixed $L \in \{1, \ldots, K\}$, to compute the $\mathscr{R}_C(L)$, we need to find the optimal subset of columns such that the minimal number of α's at any row is maximized. This is a set cover problem and cannot be solved by polynomial-time algorithms. However, it is not necessary to check all possible $\binom{K}{L}$ subset of columns. Let C_α denote the submatrix of C which consists of columns with α-entry and $\mathscr{K}(C_\alpha)$ denote the number of columns of C_α. For $L > \mathscr{K}(C_\alpha)$, $\mathscr{R}(L) = L - \min_{i \in [N]} w_H(C_\alpha(i,:))$. Hence, we only need to do traversal search for $L \leq \mathscr{K}(C_\alpha)$ and the complexity is bounded by $\mathscr{O}(2^{\mathscr{K}(C_\alpha)})$.

Example 6.5. *For the characterization matrix in Example 6.4. The submatrix C_α is*

$$C_\alpha = \begin{bmatrix} \alpha & \alpha & 0 & 0 \\ 0 & \alpha & \alpha & \alpha \\ \alpha & 0 & \alpha & 0 \end{bmatrix}. \tag{6.49}$$

We have

L	1	2	3	4	5	6	7
$\mathscr{R}_C(L)$	1	2	2	2	3	4	5

6.3.2 Searching for the Optimal Decomposition

Definition 6.6. *For PIR problem with characterization matrix C, definition the average cost of coding subspace with dimension L as*

$$E_C(L) = \frac{\mathscr{R}_C(L)}{L}. \tag{6.50}$$

As the privacy condition of PIR requires every message to be equally likely demanded by the users from the server's perspective, all messages must be used in the coding scheme. The average cost measures how many transmissions are required for each message if we partition the messages into a coding subspace with dimension L.

Given $\mathscr{R}_C(L)$ for $L \in \{1,\dots,K\}$, the optimization problem (6.21) can be formulated as:

$$\begin{aligned} &\text{minimize } \textstyle\sum_{l=1}^{K} \beta_l \mathscr{R}_C(l) \\ &\text{subject to } \textstyle\sum_{l=1}^{K} \beta_l l = K \end{aligned} \tag{6.51}$$

where β_l is the number of coding subspace with dimension l. This optimization problem is related to the Unbounded Knapsack Problem (UKP) [75]. We propose a Dynamic Programming Algorithm to solve the optimization problem (6.51).

In Algorithm 6, we search for the optimal partition for $k \in \{1,\dots,K\}$ by checking all possible size-2 partitions of k given that optimal partitions for all integers up to $k-1$ are obtained in previous rounds. The h_k^* stores the one component of the optimal partition for coding subspaces with dimension k for all $k \in \{1,\dots,K\}$. Once we get $Q(k)$, we know one component of the optimal size-2 partition for k is l_k^*. Hence, after we get all h_k^* for $k \in \{1,\dots,K\}$, we can start from the end to get the optimal partition for K by recursively partitioning it into 2 subspaces. The complexity of Algorithm 6 is bounded by $\mathscr{O}(K^2)$.

Example 6.6. *For the characterization matrix in Example 6.4, we can compute $\mathscr{R}_C(L)$ for $L \in \{1,\dots,K\}$ as follows. It would be intuitive to guess that the best decomposition is what we have shown in the previous example, $\{4,3\}$, since $E_C(4)$ is the smallest, which means coding*

Algorithm 6 Dynamic Programming Algorithm

Input: $\mathscr{R} = [\mathscr{R}_C(1), \ldots, \mathscr{R}_C(K)]$.
Output: The optimal partition vector $\mathscr{L}$, minimal number of required transmissions $Q(K)$.
Initialization: $\mathscr{L} = \emptyset$, $Q(0) = 0$.
for $k = 1, \ldots, K$ **do**
    for $l = 1, \ldots, k$ **do**
        $q_k(l) = \mathscr{R}(l) + Q(k-l)$
    end for
    $h_k^* = \arg\min_{l \le k} q_k(l)$.
    $Q(k) = q_k(h_k^*)$.
end for
while $K - \text{sum}(\mathscr{L}) > 0$ **do**
    $\mathscr{L} = \mathscr{L} \cup h^*_{K-\text{sum}(\mathscr{L})}$.
end while
Return $\mathscr{L}$ and $Q(K)$.

L	1	2	3	4	5	6	7
$\mathscr{R}_C(L)$	1	2	2	2	3	4	5
$E_C(L)$	1	1	0.67	0.5	0.6	0.67	0.71

subspace with 4 *dimension has lowest average cost.*

Apply Algorithm 6 on this example, we can get:

k	1	2	3	4	5	6	7
$Q(k)$	1	2	2	2	3	4	4
h_k^*	1	1	3	4	1	1	3

According to the table, we know that $h_7^* = 3$*, which means we can first partition the coding space into a subspace with dimension* 3*. Then, we still have* 4 *dimensions to decompose. However, since* $h_4^* = 4$*, we do not have to decompose it further. Therefore, we get the optimal decomposition, which is* $\{3,4\}$*. And the average cost of such decomposition is*

$$E_C(3,4) = \frac{\mathscr{R}_C(3) + \mathscr{R}_C(4)}{K} = \frac{4}{7} = 0.57. \tag{6.52}$$

Note that in this example, coding subspace with dimension $2,5,6,7$ *should never be used. Since* $h_k^* \neq k$ *for* $k \in \{2,5,6,7\}$*, for each* k*, there exists further partition, the number of required transmissions by which is equal to or smaller than putting* k *dimensions together. Thus, complexity of Algorithm 6 can be reduced by removing coding subspaces with dimension* $l \in \{1, \ldots, K\}$ *such that* $h_l^* < l$ *at the* 5*-th line of Algorithm 6.*

6.4 Conclusion

In this chapter, we study the single-server multi-user private information retrieval with side information problem for linear coding schemes. In this problem, it is assumed that all users want to download the same message from the single server and have different side information messages. We prove that for the linear coding schemes, the minimum number of required linear combinations can be obtained by solving an optimization problem over all partitions of the total number of messages. We also propose the Partition-and-MDS coding scheme to generate optimal linear coding schemes. Additionally, we have shown that the optimization problem can be solved by the dynamic programming algorithm without traversing all possible partitions.

Due to the assumption that all users want the same demand message and jointly generate the query, the multi-user effect can be interpreted as one user with various side information messages. The effective side information messages in such cases depend on the size of the coding subspace which is used to generate linear combinations. Specifically, when we choose different sizes of coding subspaces, the numbers of allowed side information may be different. This is different from the original single-user cases, where the side information messages are fixed and the number of side information is always the same for all coding subspaces.

There are two potential future working directions for this work. One direction is that we can release the linear coding scheme restriction and allow the coding schemes to be more general. Another direction is that we can release the assumption that all users want the same demand message.

7 Conclusion

In this thesis, we studied the cooperative data exchange problem and the private information retrieval problem. For Cooperative Data Exchange (CDE) problem in the fully connected network, we introduced the novel concept, (d,K)-Basis for simplifying the optimization problem without using submodular function minimization methods. Additionally, we proposed a polynomial-time deterministic algorithm based on the (d,K)-Basis to solve the CDE problem. We also show that our approach can be used to solve two generalized versions of the CDE problem, which are CDE with weighted cost and successive local omniscience. For the problem of Private Information retrieval (PIR) with side information, we investigated three generalized extensions, which are single-server multi-message PIR with side information, multi-server single-message PIR with side information, and single-server multi-user PIR with side information for linear coding schemes. For each extended problem, we proved the converse bound for the capacity and proposed achievability coding scheme. We introduced two useful tools, *conditioning answer string* and *virtual side information*, to help analyze the PIR with side information problem.

Possible future extensions include following:

- For CDE, the (d,K)-Basis is probably useful for other extensions of the CDE problem. For example, the helper problem, where some nodes only want to help other nodes and are not required to recover the common file. As the fully connected network is a strong assumption on topology, it would be great if we can relax this assumption and use (d,K)-Basis method to solve the CDE problem on general multi-hop networks.
- For PIR, we hope we can relax the assumption of linear coding schemes for single-server multi-user cases. In this problem, the users can be treated as a joint user but with various side information messages. Hence, the previous tools may be useful for analyzing non-linear cases. Another open problem would be the multi-server multi-message PIR with side information problem, which is the generalized version of both single-server multi-message and multi-server single-message PIR problems in this thesis.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] S. El Rouayheb, A. Sprintson, and P. Sadeghi, "On coding for cooperative data exchange," in *Information Theory (ITW 2010, Cairo), 2010 IEEE Information Theory Workshop on*. IEEE, 2010, pp. 1–5.

[3] I. Csiszar and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3047–3061, Dec 2004.

[4] A. Sprintson, P. Sadeghi, G. Booker, and S. El Rouayheb, "A randomized algorithm and performance bounds for coded cooperative data exchange," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1888–1892.

[5] ——, "Deterministic algorithm for coded cooperative data exchange," in *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*. Springer, 2010, pp. 282–289.

[6] N. Milosavljevic, S. Pawar, S. El Rouayheb, M. Gastpar, and K. Ramchandran, "Efficient algorithms for the data exchange problem," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1878–1896, 2016.

[7] T. A. Courtade and R. D. Wesel, "Coded cooperative data exchange in multihop networks," *Information Theory, IEEE Transactions on*, vol. 60, no. 2, pp. 1136–1158, 2014.

[8] N. Ding, C. Chan, Q. Zhou, R. A. Kennedy, and P. Sadeghi, "Determining optimal rates for communication for omniscience," *IEEE Transactions on Information Theory*, vol. PP, no. 99, pp. 1–1, 2017.

[9] S. E. Tajbakhsh, P. Sadeghi, and R. Shams, "A generalized model for cost and fairness analysis in coded cooperative data exchange," in *Network Coding (NetCod), 2011 International Symposium on*. IEEE, 2011, pp. 1–6.

[10] M. Gonen and M. Langberg, "Coded cooperative data exchange problem for general topologies," *Information Theory, IEEE Transactions on*, vol. 61, no. 10, pp. 5656–5669, 2015.

[11] A. Heidarzadeh, M. Yan, and A. Sprintson, "Cooperative data exchange with priority classes," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2324–2328.

[12] C. Chan, A. Al-Bashabsheh, Q. Zhou, N. Ding, T. Liu, and A. Sprintson, "Successive omniscience," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3270–3289, June 2016.

[13] D. Ozgul and A. Sprintson, "An algorithm for cooperative data exchange with cost criterion," in *2011 Information Theory and Applications Workshop*, Feb 2011, pp. 1–4.

[14] A. Heidarzadeh and A. Sprintson, "Cooperative data exchange with unreliable clients," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 496–503.

[15] H. Tyagi and S. Watanabe, "Universal multiparty data exchange and secret key agreement," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4057–4074, July 2017.

[16] T. A. Courtade and T. R. Halford, "Coded cooperative data exchange for a secret key," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 3785–3795, July 2016.

[17] M. Yan and A. Sprintson, "Algorithms for weakly secure data exchange," in *2013 International Symposium on Network Coding (NetCod)*. IEEE, 2013, pp. 1–6.

[18] M. Yan, A. Sprintson, and I. Zelenko, "Weakly secure data exchange with generalized reed solomon codes," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 1366–1370.

[19] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proceedings of IEEE 36th Annual Foundations of Computer Science*, Oct 1995, pp. 41–50.

[20] S. Yekhanin, "Private information retrieval," *Commun. ACM*, vol. 53, no. 4, pp. 68–73, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721674

[21] A. Beimel, Y. Ishai, E. Kushilevitz, and J. F. Raymond, "Breaking the o(n1(2k-1)/) barrier for information-theoretic private information retrieval," in *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, 2002, pp. 261–270.

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

[23] ——, "The capacity of robust private information retrieval with colluding databases," *IEEE Transactions on Information Theory*, vol. PP, no. 99, pp. 1–1, 2017.

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

[25] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 856–860.

[26] T. H. Chan, S. W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 2842–2846.

[27] S. Kumar, E. Rosnes, and A. G. i Amat, "Private information retrieval in distributed storage systems using an arbitrary linear code," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1421–1425.

[28] Q. Wang and M. Skoglund, "Symmetric private information retrieval for mds coded distributed storage," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[29] K. Banawan and S. Ulukus, "Private information retrieval from coded databases," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[30] R. Tajeddine and S. E. Rouayheb, "Private information retrieval from mds coded data in distributed storage systems," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1411–1415.

[31] J. Li, D. Karpuk, and C. Hollanti, "Towards practical private information retrieval from mds array codes," *IEEE Transactions on Communications*, pp. 1–1, 2020.

[32] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, A.-L. Horlemann-Trautmann, D. Karpuk, and I. Kubjas, "t-private information retrieval schemes using transitive codes," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2107–2118, 2018.

[33] 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 *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1908–1912.

[34] K. Banawan and S. Ulukus, "The Capacity of Private Information Retrieval from Byzantine and Colluding Databases," *ArXiv e-prints*, Jun. 2017.

[35] Z. Jia, H. Sun, and S. A. Jafar, "Cross subspace alignment and the asymptotic capacity of x-secure t-private information retrieval," 2018.

[36] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti, "Robust private information retrieval from coded systems with byzantine and colluding servers," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2451–2455.

[37] L. Holzbaur, R. Freij-Hollanti, and C. Hollanti, "On the capacity of private information retrieval from coded, colluding, and adversarial servers," in *2019 IEEE Information Theory Workshop (ITW)*. IEEE, 2019, pp. 1–5.

[38] X. Yao, N. Liu, and W. Kang, "The capacity of multi-round private information retrieval from byzantine databases," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 2124–2128.

[39] Z. Wang, K. Banawan, and S. Ulukus, "Private set intersection: A multi-message symmetric private information retrieval perspective," *arXiv preprint arXiv:1912.13501*, 2019.

[40] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 322–329, 2018.

[41] Q. Wang, H. Sun, and M. Skoglund, "Symmetric private information retrieval with mismatched coded messages and randomness," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 365–369.

[42] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private Information Retrieval with Side Information," *ArXiv e-prints*, Aug. 2017.

[43] Z. Chen, Z. Wang, and S. Jafar, "The Capacity of Private Information Retrieval with Private Side Information," *ArXiv e-prints*, Sep. 2017.

[44] S. Li and M. Gastpar, "Single-server multi-user private information retrieval with side information," in *2018 IEEE International Symposium on Information Theory (ISIT) (ISIT'2018)*, Vail, USA, Jun. 2018.

[45] Z. Chen, Z. Wang, and S. Jafar, "The capacity of t-private information retrieval with private side information," *arXiv preprint arXiv:1709.03022*, 2017.

[46] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "The role of coded side information in single-server private information retrieval," 2019.

[47] F. Kazemi, E. Karimi, A. Heidarzadeh, and A. Sprintson, "Single-server single-message online private information retrieval with side information," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 350–354.

[48] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," *CoRR*, vol. abs/1809.09861, 2018. [Online]. Available: http://arxiv.org/abs/1809.09861

[49] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Transactions on Information Theory*, pp. 1–1, 2019.

[50] R. Tandon, "The Capacity of Cache Aided Private Information Retrieval," *ArXiv e-prints*, Jun. 2017.

[51] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental Limits of Cache-Aided Private Information Retrieval with Unknown and Uncoded Prefetching," *ArXiv e-prints*, Sep. 2017.

[52] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The capacity of private information retrieval from heterogeneous uncoded caching databases," *IEEE Transactions on Information Theory*, 2020.

[53] K. Banawan and S. Ulukus, "Multi-message private information retrieval," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1898–1902.

[54] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," *CoRR*, vol. abs/1805.11892, 2018. [Online]. Available: http://arxiv.org/abs/1805.11892

[55] S. Li and M. Gastpar, "Single-Server Multi-Message Private Information Retrieval with Side Information," *ArXiv e-prints*, Aug. 2018.

[56] A. Heidarzadeh, B. Garcia, S. Kadhe, S. El Rouayheb, and A. Sprintson, "On the Capacity of Single-Server Multi-Message Private Information Retrieval with Side Information," *ArXiv e-prints*, Jul. 2018.

[57] A. Heidarzadeh, S. Kadhe, S. El Rouayheb, and A. Sprintson, "Single-server multi-message individually-private information retrieval with side information," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1042–1046.

[58] A. Heidarzadeh and A. Sprintson, "Private computation with side information: The single-server case," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1657–1661.

[59] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," in *2018 Iran Workshop on Communication and Information Theory (IWCIT)*, 2018, pp. 1–6.

[60] F. Kazemi, E. Karimi, A. Heidarzadeh, and A. Sprintson, "Private information retrieval with private coded side information: The multi-server case," 2019.

[61] S. Kadhe, A. Heidarzadeh, A. Sprintson, and O. O. Koyluoglu, "On an equivalence between single-server pir with side information and locally recoverable codes," *arXiv preprint arXiv:1907.00598*, 2019.

[62] R. Singleton, "Maximum distanceq-nary codes," *IEEE Transactions on Information Theory*, vol. 10, no. 2, pp. 116–118, April 1964.

[63] J. B. Orlin, "A faster strongly polynomial time algorithm for submodular function minimization," *Mathematical Programming*, vol. 118, no. 2, pp. 237–251, 2009.

[64] T. A. Courtade, B. Xie, and R. D. Wesel, "Optimal exchange of packets for universal recovery in broadcast networks," in *Military Communications Conference, 2010-Milcom 2010*. IEEE, 2010, pp. 2250–2255.

[65] S. Jaggi, P. Sanders, P. A. Chou, M. Effros, S. Egner, K. Jain, and L. M. Tolhuizen, "Polynomial time algorithms for multicast network code construction," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1973–1982, 2005.

[66] P. Hall, "On representatives of subsets," *Journal of the London Mathematical Society*, vol. 1, no. 1, pp. 26–30, 1935.

[67] R. Motwani and P. Raghavan, *Randomized algorithms*. Chapman & Hall/CRC, 2010.

[68] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

[69] S. H. Dau, W. Song, Z. Dong, and C. Yuen, "Balanced sparsest generator matrices for mds codes," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1889–1893.

[70] S. H. Dau, W. Song, and C. Yuen, "On the existence of mds codes over small fields with constrained generator matrices," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 1787–1791.

[71] W. Song and K. Cai, "Generalized reed-solomon codes with sparsest and balanced generator matrices," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1–5.

[72] N. J. A. Harvey, "Deterministic network coding by matrix completion," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.

[73] S. Li and M. Gastpar, "Cooperative data exchange based on MDS codes," in *2017 IEEE International Symposium on Information Theory (ISIT) (ISIT'2017)*, Aachen, Germany, Jun. 2017, pp. 1411–1415.

[74] A. Heidarzadeh and A. Sprintson, "Successive local and successive global omniscience," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2313–2317.

[75] G. B. Mathews, "On the partition of numbers," *Proceedings of the London Mathematical Society*, vol. 1, no. 1, pp. 486–490, 1896.

Curriculum Vitae

Su Li

Education

École Polytechnique Fédéral de Lausanne, Switzerland 2015-2020
Docteur és sciences
Advisor: Prof. Michael C. Gastpar

École Polytechnique Fédéral de Lausanne, Switzerland 2012-2015
M.Sc. in Communication Systems
with specialization in wireless communication

ETH Zürich, Switzerland 2014-2015
Master Thesis
Advisor: Prof. Friedemann Mattern

University of Electronic Science and Technology of China, China 2008-2012
B.Sc. in Communication Engineering

Research Experience

Laboratory for Information in Networked Systems, EPFL 2016-2020
Doctoral Research Assistant

Distributed System Group, ETH Zürich 2015
Student Research Assistant

Publications

S. Li, M. Gastpar. “Single-server Multi-message Private Inforamtion Retrieval with Side Information: the General Cases". In 2020 IEEE International Symposium on Information Theory (ISIT), 2020.

S. Li, M. Gastpar. “Converse for Multi-server Single-message PIR with Side Information". In

54th Annual Conference on Information Sciences and Systems, 2020.

S. Li, M. Gastpar. “Single-server Multi-message Private Information Retrieval with Side Information". In 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2018.

S. Li, M. Gastpar.“Single-server Multi-user Private Information Retrieval with Side Information". In 2018 IEEE International Symposium on Information Theory (ISIT), 2018.

S. Li, A. Shah, M. Gastpar. “Cooperative Data Exchange with Weighted Cost based on Basis Construction". In 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2017.

S. Li, M. Gastpar. “Cooperative Data Exchange based on MDS codes". In 2017 IEEE International Symposium on Information Theory (ISIT), 2017.

A. Hithnawi, **S. Li**, H. Shafagh, J. Gross, S. Duquennoy. “CrossZig: Combating Cross-Technology Interference in Low-power Wireless Networks". In The 15th International Conference on Information Processing in Sensor Networks (IPSN), 2016.