



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Ph.D. DISSERTATION

Private Information Retrieval  
with Information Leakage  
under KL Divergence and JS Divergence

KL 발산 및 JS 발산에 따른  
정보 누출이 있는 개인 정보 검색

BY

JUN-WOO TAK

FEBRUARY 2021

DEPARTMENT OF ELECTRICAL AND  
COMPUTER ENGINEERING  
COLLEGE OF ENGINEERING  
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Private Information Retrieval  
with Information Leakage  
under KL Divergence and JS Divergence

KL 발산 및 JS 발산에 따른  
정보 누출이 있는 개인 정보 검색

BY

JUN-WOO TAK

FEBRUARY 2021

DEPARTMENT OF ELECTRICAL AND  
COMPUTER ENGINEERING  
COLLEGE OF ENGINEERING  
SEOUL NATIONAL UNIVERSITY

# Private Information Retrieval with Information Leakage under KL Divergence and JS Divergence

KL 발산 및 JS 발산에 따른  
정보 누출이 있는 개인 정보 검색

지도교수 노 종 선  
이 논문을 공학박사 학위논문으로 제출함

2021년 2월

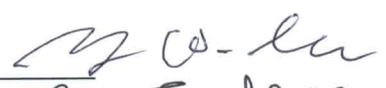




서울대학교 대학원

전기 컴퓨터 공학부

탁 준 우

탁준우의 공학박사 학위 논문을 인준함

2021년 2월

위 원 장: 이 정 우   
부위원장: 노 종 선   
위 원: 최 장   
위 원: 김 상 호   
위 원: 김 영 식 

# Abstract

In this dissertation, two main contributions are given as;

- Private information retrieval with information leakage under the Kullback-Leibler divergence is formulated and solved.
- Private information retrieval with information leakage under the Jensen-Shannon divergence is formulated and solved.

First, the private information retrieval (PIR) problem with information leakage is proposed with the Kullback-Leibler (KL) divergence. The amount of information leakage is measured by the KL divergence. The divergence is from the given reference probability distribution causing no information leakage in the PIR system to an arbitrary probability distribution of user's choice. Information leakage can be helpful in terms of the performance of the PIR system, that is, the download cost. In other words, allowing information leakage enables us to reduce the download cost of the PIR problem. We want to restrict the problem as efficiently as possible, and thus, the optimal tradeoff between the information leakage and the download cost is being considered. The problem is formulated as an optimization problem and solved using convex optimization. Furthermore, we propose an alternative PIR scheme with less message length that shows a better tradeoff than the existing PIR scheme in some tradeoff intervals.

Second, the same private information retrieval problem with information leakage is proposed but with the Jensen-Shannon (JS) divergence. The JS divergence is based on the KL divergence. The divergence occurs from the difference in probability distributions among the user's desired messages. Similar to the KL divergence, it captures the dissimilarity among the probability distributions but with some desirable features. One of the advantages it gives is that it can measure the dissimilarity of more than two probability distributions, which makes the problem more general. More specif-

ically, the problem formulated with JS divergence does not need the given reference probability distribution causing no information leakage in the PIR system. The tradeoff between the information leakage measured by the JS divergence and the download cost is formulated as a convex optimization problem and solved with numerical solutions.

**keywords:** Convex optimization, download cost, information leakage, information theory, Jensen-Shannon (JS) divergence, Kullback–Leibler (KL) divergence, private information retrieval (PIR).

**student number:** 2015-21002

# Contents

<b>Abstract</b>	<b>i</b>
<b>Contents</b>	<b>iii</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Overview of Dissertation . . . . .	2
1.3 Notations . . . . .	4
<b>2 Preliminaries</b>	<b>5</b>
2.1 Private Information Retrieval . . . . .	5
2.2 Information Leakage in PIR . . . . .	9
2.3 Convex Optimization . . . . .	11
<b>3 PIR with Information Leakage under the Kullback-Leibler Divergence</b>	<b>13</b>
3.1 Introduction . . . . .	13
3.2 Problem Formulation under the Kullback-Leibler Divergence . . . . .	14
3.3 Achievable Scheme under the Kullback-Leibler Divergence . . . . .	16
3.3.1 Probabilistic Query Generation . . . . .	16

3.3.2	Example of Symmetric TSC Scheme with $N = 2, K = 2$ . . .	20
3.3.3	Example of Symmetric TSC Scheme with $N = 3, K = 2$ . . .	23
3.3.4	Example of Symmetric TSC Scheme with $N = 3, K = 3$ . . .	24
3.3.5	Probabilistic PIR Scheme with General $N, K$ . . . . .	27
3.4	Optimal Tradeoff Between Information Leakage and Download Cost under the Kullback-Leibler Divergence . . . . .	29
3.4.1	Optimization of Probability Distribution . . . . .	30
3.4.2	Optimal Tradeoff Between Information Leakage and Down- load Cost . . . . .	35
3.4.3	Numerical Analysis with Examples . . . . .	35
3.5	Alternative Probabilistic PIR Scheme . . . . .	37
3.5.1	The Proposed Alternative PIR Scheme . . . . .	44
3.5.2	Alternative Optimal Tradeoff Between Information Leakage and Download Cost . . . . .	50
3.5.3	Numerical Analysis of the Proposed Alternative Scheme . . .	51
<b>4</b>	<b>PIR with Information Leakage under the Jensen-Shannon Divergence</b>	<b>59</b>
4.1	Introduction . . . . .	59
4.2	Problem Formulation under the Jensen-Shannon Divergence . . . . .	62
4.3	Achievable Scheme under the Jensen-Shannon Divergence . . . . .	64
4.3.1	Probabilistic Query Generation . . . . .	64
4.3.2	Example of Symmetric TSC Scheme with $N = 2, K = 2$ . . .	65
4.3.3	Example of Symmetric TSC Scheme with $N = 3, K = 2$ . . .	69
4.4	Optimal Tradeoff Between Information Leakage and Download Cost under the Jensen-Shannon Divergence . . . . .	69
4.4.1	Optimization Problem with General $N, K$ . . . . .	69
4.4.2	Numerical Analysis with Examples . . . . .	73
<b>5</b>	<b>Conclusions</b>	<b>77</b>





# List of Tables

2.1	The query of the classical PIR scheme with $N = 2, K = 2$ . . . . .	8
2.2	A probabilistic PIR query structure to retrieve $W_\theta$ . . . . .	10
3.1	The probabilistic query structure of symmetric TSC scheme with $N =$ $2, K = 2$ . . . . .	21
3.2	The probabilistic query structure of symmetric TSC scheme with $N =$ $3, K = 2$ . . . . .	25
3.3	The probabilistic query structure of symmetric TSC scheme with $N =$ $3, K = 3$ to retrieve $W_1$ . . . . .	26
3.4	A sketch of the probabilistic query structure of symmetric TSC scheme with $N$ databases and $K$ messages to retrieve $W_1$ . . . . .	28
3.5	The probabilistic query structure of alternative PIR scheme with $N =$ $3, K = 2$ to retrieve $W_1$ . . . . .	45
3.6	The probabilistic query structure of alternative PIR scheme with $N =$ $3, K = 3$ to retrieve $W_1$ . . . . .	48
3.7	The probabilistic query structure of alternative PIR scheme with $N =$ $4, K = 2$ to retrieve $W_1$ . . . . .	49
4.1	A probabilistic PIR query structure to retrieve $W_\theta$ . . . . .	63
4.2	The probabilistic query structure of symmetric TSC scheme with $N =$ $2, K = 2$ . . . . .	66

4.3	The dissimilarity of probability distribution of queries according to the desired message index seen at database 1 for $N = 2, K = 2$ . . . . .	67
4.4	The probabilistic query structure of PIR scheme with $N = 3, K = 2$ . . . . .	70
4.5	The dissimilarity of probability distribution of queries according to the desired message index seen at database 1 for $N = 3, K = 2$ . . . . .	71

# List of Figures

2.1	Classical PIR model with $N$ databases and $K$ messages. . . . .	6
3.1	Optimal tradeoff between information leakage and the normalized download cost of symmetric TSC scheme with $N = 2, K = 2$ . . . . .	38
3.2	Optimal tradeoff between information leakage and the normalized download cost of symmetric TSC scheme with $N = 3, K = 2$ . . . . .	39
3.3	Optimal tradeoff between information leakage and the normalized download cost of symmetric TSC scheme with $N = 3, K = 3$ . . . . .	40
3.4	Optimal information leakage and download cost for the probability of lower download cost of symmetric TSC scheme with $N = 2, K = 2$ . . . . .	41
3.5	Optimal information leakage and download cost for the probability of lower download cost of symmetric TSC scheme with $N = 3, K = 2$ . . . . .	42
3.6	Optimal information leakage and download cost for the probability of lower download cost of symmetric TSC scheme with $N = 3, K = 3$ . . . . .	43
3.7	Optimal tradeoff between information leakage and the download cost for two PIR schemes with $N = 3, K = 2$ . . . . .	53
3.8	Optimal tradeoff between normalized information leakage and the download cost for two PIR schemes with $N = 3, K = 2$ . . . . .	54
3.9	Optimal tradeoff between information leakage and the download cost for two PIR schemes with $N = 3, K = 3$ . . . . .	55

3.10	Optimal tradeoff between normalized information leakage and the download cost for two PIR schemes with $N = 3, K = 3$ . . . . .	56
3.11	Optimal tradeoff between information leakage and the download cost for two PIR schemes with $N = 4, K = 2$ . . . . .	57
3.12	Optimal tradeoff between normalized information leakage and the download cost for two PIR schemes with $N = 4, K = 2$ . . . . .	58
4.1	Optimal tradeoff between information leakage by the JS divergence and the normalized download cost of PIR scheme with $N = 2, K = 2$ .	75
4.2	Optimal tradeoff between information leakage by the JS divergence and the normalized download cost of PIR scheme with $N = 3, K = 2$ .	76

# Chapter 1

## Introduction

### 1.1 Background

In an era of information and data, telecommunication operators and IT service providers build data centers for reliable and independent services. Besides, in a communication environment that is more entangled than ever before, the privacy issue has become a serious consideration. There are many ways to implement privacy in different disciplines in the communication environment. Among them, technologies that can protect individual privacy from databases are in the spotlight. One of the ways to protect personal privacy is private information retrieval (PIR). PIR is a privacy problem model consisting of multiple databases, messages stored therein, and a single user. The user wants to download the desired message using several databases and does not want to let all databases know about his desired message index. PIR mainly deals with the problem of accessing sensitive data, but it is a model applicable to many applications that want to hide user preferences.

Initially studied by computer scientists, it was mainly focused on improving computational complexity problems. In the last years, it has been known what the maximum performance can be achieved with this PIR model through information-theoretic problem setting and approach. Since then, it has attracted significant attention, and

active research has been carried out to the present by a wide variety of models and assumptions. An example of solving a problem originally studied by computer engineers through an information-theoretic approach can also be found in index coding. This problem was also originally suggested by computer engineers in 1998, but after a long time, research through information theory has been actively conducted.

Solving the PIR problem by the information-theoretic approach means an approach from a rigorous perspective. Recently, Sun and Jafar [2] presented the capacity or the upper bound of information-theoretic performance of PIR and an achievable scheme to meet the capacity. Since then, many studies on various assumptions and environments have been actively conducted until now [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18].

Among the several derivatives of the classical PIR, recent studies [19, 20, 21] introduced information leakage in the PIR problem. By allowing information leakage, unlike the classical PIR, the privacy requirement is relaxed, and a certain amount of information leakage is allowed in exchange for the improved PIR performance (the download cost), which is even better than the capacity with no information leakage. Therefore, in practical operation, one can think of finding an appropriate compromise between the performance and perfect privacy. However, the tradeoff between information leakage and performance shown in recent studies does not, by itself, mean PIR capacity with information leakage. In [19], they showed the upper bound and the lower bound on capacity, but there exists a gap between them.

## **1.2 Overview of Dissertation**

This dissertation is organized as follows.

In Chapter 2, some preliminaries of PIR are briefly overviewed. Basic concepts of PIR and related researches are introduced, especially about the problem with informa-

tion leakage. The convex optimization is also introduced.

In Chapter 3, the PIR problem with information leakage under the Kullback-Leibler divergence is proposed. We adopt a PIR scheme with the probabilistic query structure as the target of optimization. By introducing the reference probability distribution indicating no information leakage in the PIR system, KL divergence measures how far an arbitrary probability distribution of user's choice diverges. Information leakage establishes the tradeoff relationship with the performance measure of the system, the download cost. The information leakage measured by KL divergence is minimized using a convex optimization problem. By applying to the given probabilistic query structure, an analytic solution to the optimal tradeoff is found. Furthermore, we observe that there can be alternative schemes that show the better tradeoff than currently known schemes. As an example, we present an alternative PIR scheme that achieves the more desirable tradeoff in some operational range.

In Chapter 4, another PIR problem with information leakage under the Jensen-Shannon divergence is proposed. For the same problem settings with probabilistic query structure, the divergence between the probability distributions in queries that depend on the identity of the desired message is measured with the JS divergence. The JS divergence is advantageous since unlike other commonly used dissimilarity measures, it can capture the dissimilarity of more than two distributions which is desirable in the PIR system with an arbitrary number of messages. The tradeoff between the information leakage taken by the JS divergence and the download cost is solved by using a convex optimization formulation. Finally, the concluding remarks are given in Chapter 5.



### 1.3 Notations

For a positive integer  $a$  and  $b$ , we denote  $[a] \triangleq \{1, 2, \dots, a\}$  and  $[a : b] \triangleq \{a, a + 1, \dots, b\}$ . We use the notation  $A_{[a:b]} \triangleq \{A_a, A_{a+1}, \dots, A_b\}$  if  $a \leq b$ , and null set, otherwise. We use uppercase letters for random variables (RVs),  $X$  for scalar and  $\mathbf{X}$  for vector.  $\mathbb{E}_X[\cdot]$  denotes the expectation with respect to RV  $X$ .  $H(X)$  denotes entropy of  $X$  and  $I(X; Y)$  denotes the mutual information (MI) between  $X$  and  $Y$ .  $D_{KL}(P \parallel Q)$  represents the Kullback-Leibler divergence from a probability distribution  $Q$  to a probability distribution  $P$ .

## **Chapter 2**

### **Preliminaries**

In this chapter, some preliminaries needed for this dissertation are introduced. First, the basic problem setup and concepts of PIR are described. Second, some related results of recent research on information leakage for PIR are discussed. The principle of maximum entropy and the principle of minimum cross-entropy are explained.

#### **2.1 Private Information Retrieval**

The PIR problem introduced in the seminal paper [1] deals with a security protocol required for communication between a user and databases. The user wants to download the desired information from databases while hiding the identity of the information being requested from all databases. A trivial solution to this problem is to make the user download all the information in the databases, which is very inefficient and not desirable.

The original PIR problem was mainly studied by computer engineers. In the recent work, however, the PIR problem was analyzed from the viewpoint of information theory [2]. The upload cost (communication cost from the user to databases) is regarded as negligible compared to the download cost (the amount of information flow from

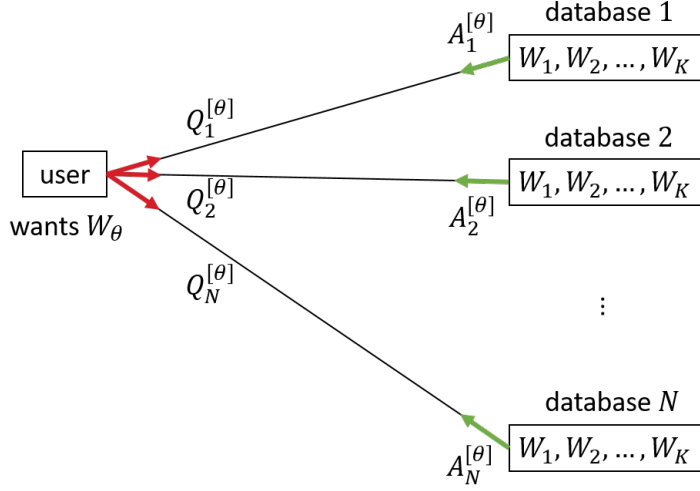


Figure 2.1: Classical PIR model with  $N$  databases and  $K$  messages.

databases to the user). By considering only the size of the desired message and the size of the required download, the performance measure of the PIR problem, that is, the rate is defined as

$$\text{Rate} = \frac{(\text{desired message size})}{(\text{required download cost})}.$$

The maximum possible rate achievable for the given PIR problem is called the information theoretic capacity.

The classical PIR model is depicted in Figure 2.1. It consists of  $N$  databases and statistically independent  $K$  messages, all replicated and stored identically therein. For this simple scenario, databases cannot communicate with each other, and all message sizes are equal to  $L$ . By using the Shannon entropy notation, the following equations hold

$$H(W_1, W_2, \dots, W_K) = H(W_1) + H(W_2) + \dots + H(W_K),$$

$$H(W_1) = H(W_2) = \dots = H(W_K) = L.$$

The user wants to retrieve any one of the messages  $W_\theta$  uniformly, while the identity of the message or the user's preference of the message  $\theta \in [K]$  is hidden from all

databases. The user generates and forwards query  $Q_n^{[\theta]}$  to each database  $n \in [N]$ . There are two conditions that should be satisfied in the classical PIR problem, that is, the privacy condition and the correctness condition. First, the privacy condition hides the preference of the user, and each query should not contain any information about the index  $\theta$ . The privacy condition is often represented as

$$I(\theta; Q_n^{[\theta]}) = 0, \theta \in [K], n \in [N].$$

Once each database that has been queried as  $Q_n^{[\theta]}$  must send back the corresponding answer,  $A_n^{[\theta]}$ . The answering process is deterministic with the query and messages in the database, that is,

$$H(A_n^{[\theta]} | Q_n^{[\theta]}, W_1, \dots, W_K) = 0, \theta \in [K], n \in [N].$$

Second, the correctness condition is that the user has the collection of answers from  $N$  databases and must be able to correctly decode the desired message  $W_\theta$ , that is,

$$H(W_\theta | Q_{[1:N]}^{[\theta]}, A_{[1:N]}^{[\theta]}) = 0.$$

The authors in [2] proved that the exact capacity of the PIR problem with  $N$  databases and  $K$  messages is given as

$$C = \frac{1}{1 + \frac{1}{N} + \frac{1}{N^2} + \dots + \frac{1}{N^{K-1}}}. \quad (2.1)$$

The capacity result can be interpreted that  $\frac{1}{N} + \frac{1}{N^2} + \dots + \frac{1}{N^{K-1}}$  extra bits of the download cost are needed per desired message bit to ensure the privacy of the user. Also, one can easily observe that the capacity increases with more  $N$  and less  $K$ . The simplest classical PIR example for the case of  $N = 2, K = 2$  from [2] achieving the capacity result is presented in Table 2.1 to retrieve  $W_1$  and  $W_2$ .

In this example, each message has the size of  $L = 4$ .  $W_1$  and  $W_2$  are represented by  $[a_1, a_2, a_3, a_4]$  and  $[b_1, b_2, b_3, b_4]$ , respectively. For both cases of retrieving  $W_1$  and

Table 2.1: The query of the classical PIR scheme with  $N = 2, K = 2$

(a) To retrieve  $W_1$

Database 1	Database 2
$a_1$	$a_2$
$b_1$	$b_2$
$a_3 + b_2$	$a_4 + b_1$

(b) To retrieve  $W_2$

Database 1	Database 2
$a_1$	$a_2$
$b_1$	$b_2$
$a_2 + b_3$	$a_1 + b_4$

$W_2$ , symbols required from each database form symmetric structures from the perspective of individual database, and thus, the message being retrieved is indistinguishable. Inevitably, unnecessary symbols are also requested for privacy expense, for example,  $b_1$  from database 1 in Table 2.1 (a). This symbol is used in a way that is downloaded in a summation to the desired symbol from another database that is not requested, for this case,  $a_4 + b_1$  from database 2.

## 2.2 Information Leakage in PIR

Information leakage can be introduced in private information retrieval system if the perfect privacy requirement is relaxed. The relaxation gives a gain in rate, and therefore, the tradeoff between privacy cost and retrieval cost is of our interest. In a practical way, obtaining perfect privacy in PIR accompanies inefficiency in terms of communication costs, especially when there is a large number of messages stored in databases. It is reasonable if a user can selectively expose the privacy to some allowable extent. With this idea, recent PIR studies have dealt with rate gains that can be obtained at the expense of perfect privacy.

There are several representative kinds of research with information leakage in the PIR problem. In their studies, information leakage was defined in different ways. There were studies about finding information theoretic capacity of PIR with information leakage, and also, there were studies about finding the numerical solution of the problem. Most of the studies have in common that they used probabilistic query models. Therefore, before we cover the PIR studies about information leakage, it is worth mentioning the work in [23].

In [23], there was an attempt to reduce the conventional message size  $L = N^K$  to its optimal value  $N - 1$ . In that study, a probability-based PIR scheme was proposed rather than the conventional deterministic PIR scheme, and the user could implement

Table 2.2: A probabilistic PIR query structure to retrieve  $W_\theta$

Option	DB 1	...	DB $N$	Probability	Download cost
1	$Q_1^{[\theta]}(1)$	...	$Q_N^{[\theta]}(1)$	$p_1$	$d_1$
2	$Q_1^{[\theta]}(2)$	...	$Q_N^{[\theta]}(2)$	$p_2$	$d_2$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$M$	$Q_1^{[\theta]}(M)$	...	$Q_N^{[\theta]}(M)$	$p_M$	$d_M$

PIR statistically by using one of several options. Assume that there are  $M$  options for a user to retrieve an arbitrary desired message and any selected option gives the user the desired message equivalently. When the  $m$ -th option is used to retrieve the message  $W_\theta$ , we denote the query received at database  $n$  and the answer to that query as  $Q_n^{[\theta]}(m)$  and  $A_n^{[\theta]}(m)$ , respectively. Then the correctness condition becomes

$$H(W_\theta | Q_{[1:N]}^{[\theta]}(m), A_{[1:N]}^{[\theta]}(m)) = 0,$$

$$m \in [1 : M], \theta \in [1 : K].$$

A sketch of probabilistic query structure of PIR with options is shown in Table 2.2.  $M$  options of possible query sets are represented with their corresponding probabilities  $p_1, p_2, \dots, p_M$  and the download costs  $d_1, d_2, \dots, d_M$ .

In [23], the uniform distribution for  $M$  options is used for the capacity-achieving scheme. However, in leaky PIR (LPIR) [19],  $\epsilon$ -privacy is introduced where  $\epsilon$  defines the upper bound of the ratio of arbitrary two probabilities of queries sent to a database. More specifically, with non-negative  $\epsilon$ ,  $\epsilon$ -privacy is given as

$$\frac{Pr(Q_n^{[k_1]} = q)}{Pr(Q_n^{[k_2]} = q)} \leq e^\epsilon, \quad k_1, k_2 \in [1 : K], n \in [1 : N],$$

where  $\epsilon$  indicates the amount of information leakage allowance and  $q$  represents the possible realization of the query sent. Note that  $\epsilon$ -privacy is similar to the definition of differential privacy [22]. In the study of LPIR, they tried to find the capacity of

PIR with information leakage for the given  $\epsilon$  but only loose bound is given as a result. In another study of weakly PIR (WPIR) [20], the information leakage is defined in more information theoretic manner as the nonzero mutual information between desired message index and corresponding query. Therefore, their privacy condition is relaxed as

$$I(\theta; Q_n^{[\theta]}) \leq \rho, \theta \in [1 : K], n \in [1 : N],$$

where  $\rho$  indicates the amount of information leakage allowance. Also, in [21], the maximal information leakage metric is proposed as a measure of information leakage defined as

$$\mathcal{L}(\theta \rightarrow Q_n^{[\theta]}) = \log \sum_{q \in \mathcal{Q}} \max_{k \in [1:K]} Pr_{Q_n^{[k]}}(q),$$

where  $\mathcal{Q}$  is the query space or the set of possible queries that the user can ask. In their study, they found an optimal tradeoff between the information leakage and the download cost for a specific achievable scheme. The research in this dissertation resembles that of [21]. For a specific achievable scheme, we will apply the newly defined information leakage measures to the PIR problem and solve the optimization problems to find the optimal tradeoff between the information leakage measure and PIR performance.

## 2.3 Convex Optimization

Convex optimization is a special class of mathematical optimization that studies the problem of minimizing convex objective function over convex sets or maximizing concave objective function over convex sets. Compared to general mathematical optimization problems, many classes of convex optimization problems can be solved very efficiently in polynomial time.

A set  $C$  is said to be convex if for  $x_1, x_2 \in C$  and  $0 \leq \theta \leq 1$ , we have

$$\theta x_1 + (1 - \theta)x_2 \in C,$$



and a function  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  is said to be convex if its domain is convex set and for  $x, y$  and  $0 \leq \theta \leq 1$ , we have

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y). \quad (2.2)$$

A convex optimization problem in its standard form is given as follows:

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{subject to} && f_i(x) \leq 0, \quad i = 1, \dots, m, \\ & && h_i(x) = 0, \quad i = 1, \dots, p, \end{aligned}$$

where  $x \in \mathbf{R}^n$  is the optimization variable and convex function  $f_0 : \mathbf{R}^n \rightarrow \mathbf{R}$  is the objective function. Inequalities with convex functions  $f_i(x) : \mathbf{R}^n \rightarrow \mathbf{R}$  are inequality constraints and equalities with affine functions  $h_i(x) : \mathbf{R}^n \rightarrow \mathbf{R}$  are equality constraints. When the objective function is strictly convex, that is, only inequality holds in (2.2), solving a convex optimization problem gives at most one optimal point or global optimum point.

## Chapter 3

# PIR with Information Leakage under the Kullback-Leibler Divergence

### 3.1 Introduction

The classical PIR problem [2] and most of its related studies[ . . . ] use deterministic query structures as their achievable schemes. This means that in order to download the message the user wants, a fixed query structure must always be used. However, some recent papers suggested the use of probabilistically generated queries to deal with different semantics in messages [24] or to reduce the message size and the upload cost [23]. In the probabilistic query model, the user can choose one of several query options to download the desired message. It is necessary that the desired message can be obtained no matter which option is used.

Furthermore, the probabilistic query model is also used in the studies on the PIR problem with information leakage mentioned as in Section 2.2. In this case, the amount of information leakage generated can be changed by adjusting the probability allocation. In this dissertation, the Kullback-Leibler (KL) divergence is introduced to describe how the adjustment is made. The KL divergence or relative entropy is a measure of how two probability distributions differ, which will be used as an information

leakage in this dissertation. Since its introduction [25] in 1951, it has been popularly used in a wide variety of fields and applications. Consider that we have two probability distributions  $P$  and  $Q$ . For discrete distributions, the KL divergence from  $Q$  to  $P$  is defined as follows.

**Definition 3.1.** *The Kullback-Leibler (KL) divergence from a probability distribution  $Q$  to a probability distribution  $P$  defined on the probability space  $\mathcal{X}$  is defined as*

$$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

In the formula, the logarithm with base 2 or base  $e$  is used if the unit of information being measured is in bits or in nats, respectively. Note that KL divergence is always nonnegative and equal to zero if and only if  $P = Q$ .

## 3.2 Problem Formulation under the Kullback-Leibler Divergence

We have a PIR scenario in Figure 2.1, where  $K$  messages are stored identically in  $N$  databases without collusion. The messages are denoted as  $W_1, W_2, \dots, W_K$  and equally sized by  $H(W_k) = L, k \in [1 : K]$ . Consider a probabilistic query structure with  $M$  options as shown in Table 2.2. The download costs for each option  $d_1, d_2, \dots, d_M$  are computed as the sum of the answer sizes received from databases. The download cost  $d_m$  for the  $m$ -th option is given as the summation of the answer size across the databases given as

$$d_m = \sum_{n=1}^N H(A_n(m)).$$

The performance of this probabilistic PIR model is measured by the expectation of the download cost normalized by the message size  $L$  given as

$$D = \frac{1}{L} \sum_{m=1}^M p_m d_m. \quad (3.1)$$

In fact,  $D$  is the amount of download needed per retrieval of unit message size. The achievable rate is simply the reciprocal of it.

Essentially, without any information leakage, the upper bound of achievable rate  $1/D$  must match the capacity  $C$  in (2.1). However, with allowed information leakage, we can reduce the download cost in (3.1), and thus we can achieve a gain in rate. The range that the download cost can take is between two extreme cases, no information leakage case and full information leakage case. When there is no information leakage, the user should download  $1/C$  per unit message size. Otherwise, when we allow full information leakage, the user directly downloads the desired message only, and the download cost per unit message size is 1. Therefore,  $D$  has the following range,

$$D \in \left[ 1, \frac{1}{C} \right]. \quad (3.2)$$

Consider a probabilistic PIR query structure in Table 2.2. Let  $U = (u_1, u_2, \dots, u_M)$  represent the probability assignment to  $P = (p_1, p_2, \dots, p_M)$  for the case with no information leakage, or perfect privacy in statistical manner. Then the information leakage denoted by  $\rho_{KL}$  when distribution  $P$  is used instead of  $U$  is measured as the Kullback-Leibler divergence

$$\rho_{KL} = D_{KL}(P \parallel U) \quad (3.3)$$

$$\begin{aligned} &= \sum_{q \in \mathcal{Q}} P(q) \log \frac{P(q)}{U(q)} \\ &= \sum_{m=1}^M p_m \log \frac{p_m}{u_m}, \end{aligned} \quad (3.4)$$

where the query space  $\mathcal{Q}$  has the cardinality of  $|\mathcal{Q}| = M$ . Note that if and only if  $P = U$ , then the information leakage equals zero.

The goal of this dissertation is to find the most efficient probability distribution  $P$  that minimizes the information leakage measured by the Kullback-Leibler divergence given a probability distribution  $U$  and a certain amount of the target download cost  $D$ . By doing so, the optimal tradeoff between the download cost and the information leakage will be established. The problem can be formulated in a convex optimization problem as follows:

$$\begin{aligned}
& \text{minimize} && \rho_{KL} = D_{KL}(P \parallel U) \\
& \text{subject to} && \frac{1}{L} \sum_{m=1}^M p_m d_m = D, \\
& && \sum_{m=1}^M p_m = 1.
\end{aligned} \tag{3.5}$$

### 3.3 Achievable Scheme under the Kullback-Leibler Divergence

#### 3.3.1 Probabilistic Query Generation

Throughout the dissertation, the probabilistic query structure of Scheme 2 in [24] is adopted, which corresponds to the TSC scheme proposed in [23] with database symmetry. Therefore we will refer to this scheme as the symmetric TSC scheme. Unlike Scheme 2 in [24], no different semantics in messages are assumed, and the message length  $L$  is fixed to its minimum value  $N - 1$ . Let  $W_\theta$  be the desired message size  $N - 1$  given as

$$W_\theta = [W_\theta(1), W_\theta(2), \dots, W_\theta(N - 1)], \theta \in [1 : K].$$

The probabilistic query structure of PIR is explained as follows [23, 24] and described in Table 3.4.

- (Step 1) Use first  $N - 1$  databases to download desired message symbols

$$W_\theta(1), W_\theta(2), \dots, W_\theta(N - 1),$$

respectively. Enumerate its cyclic shifts across databases. This step builds  $N$  query options.

- (Step 2) Download  $W_i(1)$  from the first database, where  $i \in [1 : K] \setminus \{\theta\}$ . Use the other  $N - 1$  databases to download desired message symbols added to  $W_i(1)$ , that is,

$$\begin{aligned} &W_\theta(1) + W_i(1) \\ &W_\theta(2) + W_i(1) \\ &\vdots \\ &W_\theta(N - 1) + W_i(1), \end{aligned}$$

respectively. Enumerate its cyclic shifts across databases. Up to the cyclic shifts we have  $N$  query options. Repeat for other symbols in  $W_i$ , which are  $W_i(2), \dots, W_i(N - 1)$ . Repeat for the other  $i \in [1 : K] \setminus \{\theta\}$ . This step builds  $N(N - 1) \binom{K-1}{1}$  query options.

- (Step 3) Download  $W_i(1) + W_j(1)$  from the first database, where  $i, j \in [1 : K] \setminus \{\theta\}$  and  $i \neq j$ . Use the other  $N - 1$  databases to download desired message symbols added to  $W_i(1) + W_j(1)$ , that is,

$$\begin{aligned} &W_\theta(1) + W_i(1) + W_j(1) \\ &W_\theta(2) + W_i(1) + W_j(1) \\ &\vdots \\ &W_\theta(N - 1) + W_i(1) + W_j(1), \end{aligned}$$

respectively. Enumerate its cyclic shifts across databases. Up to the cyclic shifts we have  $N$  query options. Repeat for other symbols in  $W_i$  and  $W_j$ , which are in

total  $(N - 1)^2$  multiple cases. Repeat for the other  $i, j$ , where  $i, j \in [1 : K] \setminus \{\theta\}$  and  $i \neq j$ . This step builds  $N(N - 1)^2 \binom{K-1}{2}$  query options.

- Repeat the steps with the same procedure until it reaches Step  $K$ . Step  $K$  builds  $N(N - 1)^{K-1} \binom{K-1}{K-1}$  query options.

In the query structure, the queries generated in Step 1 trivially request the desired symbols only. It is obvious that the user can have the desired message directly. For the queries generated from Step 2 to Step  $K$ , the user can subtract the undesired symbol or the sum of the undesired symbols from received symbols and then recover the desired message.

By adding up the number of query options built from each step, the number of possible options can be calculated as

$$\begin{aligned}
& N + N(N - 1) \binom{K - 1}{1} + \dots + N(N - 1)^{K-1} \binom{K - 1}{K - 1} \\
&= N \sum_{k=0}^{K-1} (N - 1)^k \binom{K - 1}{k} \\
&= N \cdot N^{K-1} \\
&= N^K.
\end{aligned} \tag{3.6}$$

Therefore, there are  $N^K$  options that the user can take.

**Remark 1.** Refer to the *k-sum* terminology in [2], where a *k-sum* symbol is the sum of  $k$  distinct symbols, each drawn from  $k$  different messages. Note that in the probabilistic query structure explained above, all possible *k-sums* appear just once in each database for every  $k = 1, \dots, K$ . The structure also includes one *0-sum* symbol for each database, which means downloading nothing. By adding up the number of all types of *k-sum* queries that the user can send to an individual database, the size of universal query space  $\mathcal{Q}$  can be calculated as follows.

- The number of *0-sum*: 1

- The number of 1-sums:  $(N - 1) \binom{K}{1}$
- The number of 2-sums:  $(N - 1)^2 \binom{K}{2}$
- ⋮
- The number of  $K$ -sums:  $(N - 1)^K \binom{K}{K}$

Adding all, we have

$$\begin{aligned}
& 1 + (N - 1) \binom{K}{1} + (N - 1)^2 \binom{K}{2} + \cdots + (N - 1)^K \binom{K}{K} \\
&= \sum_{k=0}^K (N - 1)^k \binom{K}{k} \\
&= N^K,
\end{aligned}$$

which is identical to the number of possible query options. Therefore, the probabilistic query structure can be understood as a proper permutation of the elements in the query space.

**Remark 2.** The query structure is symmetric within each database from Remark 1. By allocating the uniform probability to  $N^K$  query options, the PIR scheme achieves perfect privacy. Therefore  $U = (u_1, u_2, \dots, u_M)$  in (3.3) is the uniform distribution with probability mass function  $1/N^K$  for this PIR scheme and (3.4) can be rewritten as

$$\begin{aligned}
\sum_{m=1}^M p_m \log \frac{p_m}{u_m} &= \sum_{m=1}^M p_m (\log p_m - \log u_m) \\
&= \sum_{m=1}^M p_m (\log p_m - \log \frac{1}{N^K}) \\
&= \sum_{m=1}^M p_m \log p_m + \log N^K \\
&= H(U) - H(P).
\end{aligned} \tag{3.7}$$



### 3.3.2 Example of Symmetric TSC Scheme with $N = 2, K = 2$

In this subsection, the simplest example of the probabilistic PIR scheme with  $N = 2$  databases and  $K = 2$  messages is introduced. The message size is  $L = N - 1 = 1$  and two messages are simply  $W_1 = W_1(1)$  and  $W_2 = W_2(1)$ . The number of query options that the user can choose is  $M = N^K = 4$ . The query structures to retrieve  $W_1$  and  $W_2$  are shown in Table 3.1. The notation  $\phi$  in the tables means that the user requests nothing, and thus no symbol is downloaded. Without loss of generality, we assume that  $W_1$  is wanted, and the PIR scheme is analyzed hereafter.

The queries are generated by the following steps.

- (Step 1) Use the first database to download the desired message symbol  $W_1(1)$ , which is the option 1. Use its cyclic shift as the option 2.
- (Step 2) Download  $W_2(1)$  from the first database. Let the second database download desired message symbol  $W_1(1)$  added to it,  $W_1(1) + W_2(1)$ . This forms the option 3 and use its cyclic shift as the option 4.

If probability  $(p_1, p_2, p_3, p_4)$  is equiprobable, the queries are symmetric within each database. All possible 0-sum, 1-sum, and 2-sum symbols are generated over possible options in both databases. For database 1, each symbol from two messages is requested equally likely at the option 1 and the option 3, respectively. The option 2 and the option 4 leaves no information about desired message index 1 since the option 2 requests no message symbol and the option 4 requests the sum of both message symbols. Therefore database 1 cannot tell which one is the user's interest between  $W_1$  and  $W_2$ . A similar observation is found in database 2. Consequently, perfect privacy is guaranteed. In this example, the expectation of the download cost normalized by message size is calculated as

$$\begin{aligned}
 D &= \frac{1}{L} \sum_{m=1}^M p_m d_m \\
 &= p_1 \cdot 1 + p_2 \cdot 1 + p_3 \cdot 2 + p_4 \cdot 2.
 \end{aligned} \tag{3.8}$$

Table 3.1: The probabilistic query structure of symmetric TSC scheme with  $N = 2, K = 2$

(a) To retrieve  $W_1$

Option	Database 1	Database 2	Probability	Download cost
1	$W_1(1)$	$\phi$	$p_1$	1
2	$\phi$	$W_1(1)$	$p_2$	1
3	$W_2(1)$	$W_1(1) + W_2(1)$	$p_3$	2
4	$W_1(1) + W_2(1)$	$W_2(1)$	$p_4$	2

(b) To retrieve  $W_2$

Option	Database 1	Database 2	Probability	Download cost
1	$W_2(1)$	$\phi$	$p_1$	1
2	$\phi$	$W_2(1)$	$p_2$	1
3	$W_1(1)$	$W_1(1) + W_2(1)$	$p_3$	2
4	$W_1(1) + W_2(1)$	$W_1(1)$	$p_4$	2

Since  $(p_1, p_2, p_3, p_4)$  is equiprobable, (3.8) becomes

$$D = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{3}{2}, \quad (3.9)$$

where its reciprocal matches the PIR capacity result in (2.1) with  $N = 2$  and  $K = 2$ ,

$$C = \frac{1}{1 + 1/2} = \frac{2}{3}.$$

Now information leakage is introduced. Information leakage is generated by allowing higher probabilities for the options with lower download costs. In the above example, if the option 1 and the option 2 are more frequently used than the option 3 and the option 4, the average download cost can be reduced. For example, input probability distribution  $(p_1, p_2, p_3, p_4) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$  in (3.8) gives

$$D = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 2 = \frac{4}{3},$$

which is definitely lower than the cost with no information leakage in (3.9),  $\frac{3}{2}$ .

The amount of information leakage  $\rho_{KL}$ , compared with the perfect privacy is measured by the Kullback-Leibler divergence. Input probability distribution  $P = (p_1, p_2, p_3, p_4) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$  and  $U = (u_1, u_2, u_3, u_4) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$  in (3.4) gives the divergence from  $U$  to  $P$ ,

$$\begin{aligned} \rho_{KL} &= D_{KL}(P \parallel U) \\ &= \frac{1}{3} \log_2 \frac{\frac{1}{3}}{\frac{1}{4}} + \frac{1}{3} \log_2 \frac{\frac{1}{3}}{\frac{1}{4}} + \frac{1}{6} \log_2 \frac{\frac{1}{6}}{\frac{1}{4}} + \frac{1}{6} \log_2 \frac{\frac{1}{6}}{\frac{1}{4}} \\ &= 5/3 - \log_2 3 \approx 0.0817, \end{aligned}$$

or simply, from (3.7), we have

$$\begin{aligned} \rho_{KL} &= H(U) - H(P) \\ &= H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) - H\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right) \\ &\approx 2 - 1.9183 \\ &= 0.0817. \end{aligned}$$

In the computation, a logarithm with base 2 is used throughout the dissertation.

**Remark 3.** *Intuitively, allowing more information leakage will lower the download cost. At this moment we only have two download cost-information leakage pairs of  $(\frac{3}{2}, 0)$  and  $(\frac{4}{3}, 0.0817)$ . Furthermore, if we allow maximum information leakage, the download cost can be further reduced to its extreme point. Suppose the user only uses the option 1 or the option 2 equally likely to retrieve  $W_1$  or  $W_2$ , and then the download cost becomes 1. It is obvious that the databases will notice what message the user wants for sure. From (3.7), information leakage is calculated as*

$$\begin{aligned}
 \rho_{KL} &= H(U) - H(P) \\
 &= H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) - H\left(\frac{1}{2}, \frac{1}{2}, 0, 0\right) \\
 &= 2 - 1 \\
 &= 1.
 \end{aligned}$$

*We now have one more download cost-information leakage pair of (1,1). We can say that these pairs are achievable in the regime of information leakage measured by the Kullback-Leibler divergence. With more achievable pairs, the download cost-information leakage tradeoff can be described. However, it is not sure yet if these pairs will draw the optimal curve. In Section 3.4 we will formally characterize the optimal tradeoff between the download cost and information leakage.*

In the next three subsections, two more examples of the query generation with small  $N$  and  $K$  are presented without considering information leakage. Additionally, a sketch of the probabilistic query structure of symmetric TSC scheme with general  $N$  and  $K$  is followed.

### 3.3.3 Example of Symmetric TSC Scheme with $N = 3, K = 2$

The example of the query structure of probabilistic PIR scheme with  $N = 3$  databases and  $K = 2$  messages is demonstrated. Each message consists of  $L = N - 1 = 2$  sym-

bols, namely  $W_1 = [W_1(1), W_1(2)]$  and  $W_2 = [W_2(1), W_2(2)]$ , respectively. The user has  $M = N^K = 9$  query options in total. If the options are chosen uniformly, perfect privacy is achieved statistically. Otherwise, the model will suffer a certain amount of information leakage. Specific query structures to retrieve  $W_1$  and  $W_2$  are presented in Table 3.2.

Note that if the query options are chosen equally likely, then the expected download cost normalized by message size  $L = 2$  is calculated as

$$\begin{aligned}
 D &= \frac{1}{L} \sum_{m=1}^M p_m d_m & (3.10) \\
 &= \frac{1}{2} (p_1 \cdot 2 + p_2 \cdot 2 + p_3 \cdot 2 + p_4 \cdot 3 + p_5 \cdot 3 + p_6 \cdot 3 + p_7 \cdot 3 + p_8 \cdot 3 + p_9 \cdot 3) \\
 &= \frac{1}{2} \cdot \frac{1}{9} (2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 3) = \frac{24}{18} = \frac{4}{3},
 \end{aligned}$$

where its reciprocal matches the PIR capacity result in (2.1) with  $N = 3$  and  $K = 2$ ,

$$C = \frac{1}{1 + 1/3} = \frac{3}{4}.$$

Now information leakage can be introduced, like the previous example with  $N = 2$ ,  $K = 2$ . By allowing higher probabilities for the options with lower download costs, the average download cost can be reduced. In this example, the options 1, 2, and 3 correspond to the options with lower download cost.

### 3.3.4 Example of Symmetric TSC Scheme with $N = 3$ , $K = 3$

Likewise, the example of the query generation with  $N = 3$  databases and  $K = 3$  messages is demonstrated. Each message consists of  $L = N - 1 = 2$  symbols, namely  $W_1 = [W_1(1), W_1(2)]$ ,  $W_2 = [W_2(1), W_2(2)]$  and  $W_3 = [W_3(1), W_3(2)]$ , respectively. The number of query options the user can choose is  $M = N^K = 27$ . The query structure to retrieve  $W_1$  is shown in the Table 3.3. The retrieval of  $W_2$  or  $W_3$  is identical with the case of  $W_1$  after modifications in the subscripts.

Table 3.2: The probabilistic query structure of symmetric TSC scheme with  $N = 3, K = 2$

(a) To retrieve  $W_1$

Option	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_1(1)$	$W_1(2)$	$\phi$	$p_1$	2
2	$\phi$	$W_1(1)$	$W_1(2)$	$p_2$	2
3	$W_1(2)$	$\phi$	$W_1(1)$	$p_3$	2
4	$W_2(1)$	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$p_4$	3
5	$W_1(2) + W_2(1)$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_5$	3
6	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$W_2(1)$	$p_6$	3
7	$W_2(2)$	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$p_7$	3
8	$W_1(2) + W_2(2)$	$W_2(2)$	$W_1(1) + W_2(2)$	$p_8$	3
9	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$W_2(2)$	$p_9$	3

(b) To retrieve  $W_2$

Option	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_2(1)$	$W_2(2)$	$\phi$	$p_1$	2
2	$\phi$	$W_2(1)$	$W_2(2)$	$p_2$	2
3	$W_2(2)$	$\phi$	$W_2(1)$	$p_3$	2
4	$W_1(1)$	$W_1(1) + W_2(1)$	$W_1(1) + W_2(2)$	$p_4$	3
5	$W_1(1) + W_2(2)$	$W_1(1)$	$W_1(1) + W_2(1)$	$p_5$	3
6	$W_1(1) + W_2(1)$	$W_1(1) + W_2(2)$	$W_1(1)$	$p_6$	3
7	$W_1(2)$	$W_1(2) + W_2(1)$	$W_1(2) + W_2(2)$	$p_7$	3
8	$W_1(2) + W_2(2)$	$W_1(2)$	$W_1(2) + W_2(1)$	$p_8$	3
9	$W_1(2) + W_2(1)$	$W_1(2) + W_2(2)$	$W_1(2)$	$p_9$	3

Table 3.3: The probabilistic query structure of symmetric TSC scheme with  $N = 3$ ,  $K = 3$  to retrieve  $W_1$

Opt.	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_1(1)$	$W_1(2)$	$\phi$	$p_1$	2
2	$\phi$	$W_1(1)$	$W_1(2)$	$p_2$	2
3	$W_1(2)$	$\phi$	$W_1(1)$	$p_3$	2
4	$W_2(1)$	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$p_4$	3
5	$W_1(2) + W_2(1)$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_5$	3
6	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$W_2(1)$	$p_6$	3
7	$W_2(2)$	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$p_7$	3
8	$W_1(2) + W_2(2)$	$W_2(2)$	$W_1(1) + W_2(2)$	$p_8$	3
9	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$W_2(2)$	$p_9$	3
10	$W_3(1)$	$W_1(1) + W_3(1)$	$W_1(2) + W_3(1)$	$p_{10}$	3
11	$W_1(2) + W_3(1)$	$W_3(1)$	$W_1(1) + W_3(1)$	$p_{11}$	3
12	$W_1(1) + W_3(1)$	$W_1(2) + W_3(1)$	$W_3(1)$	$p_{12}$	3
13	$W_3(2)$	$W_1(1) + W_3(2)$	$W_1(2) + W_3(2)$	$p_{13}$	3
14	$W_1(2) + W_3(2)$	$W_3(2)$	$W_1(1) + W_3(2)$	$p_{14}$	3
15	$W_1(1) + W_3(2)$	$W_1(2) + W_3(2)$	$W_3(2)$	$p_{15}$	3
16	$W_2(1) + W_3(1)$	$W_1(1) + W_2(1) + W_3(1)$	$W_1(2) + W_2(1) + W_3(1)$	$p_{16}$	3
17	$W_1(2) + W_2(1) + W_3(1)$	$W_2(1) + W_3(1)$	$W_1(1) + W_2(1) + W_3(1)$	$p_{17}$	3
18	$W_1(1) + W_2(1) + W_3(1)$	$W_1(2) + W_2(1) + W_3(1)$	$W_2(1) + W_3(1)$	$p_{18}$	3
19	$W_2(1) + W_3(2)$	$W_1(1) + W_2(1) + W_3(2)$	$W_1(2) + W_2(1) + W_3(2)$	$p_{19}$	3
20	$W_1(2) + W_2(1) + W_3(2)$	$W_2(1) + W_3(2)$	$W_1(1) + W_2(1) + W_3(2)$	$p_{20}$	3
21	$W_1(1) + W_2(1) + W_3(2)$	$W_1(2) + W_2(1) + W_3(2)$	$W_2(1) + W_3(2)$	$p_{21}$	3
22	$W_2(2) + W_3(1)$	$W_1(1) + W_2(2) + W_3(1)$	$W_1(2) + W_2(2) + W_3(1)$	$p_{22}$	3
23	$W_1(2) + W_2(2) + W_3(1)$	$W_2(2) + W_3(1)$	$W_1(1) + W_2(2) + W_3(1)$	$p_{23}$	3
24	$W_1(1) + W_2(2) + W_3(1)$	$W_1(2) + W_2(2) + W_3(1)$	$W_2(2) + W_3(1)$	$p_{24}$	3
25	$W_2(2) + W_3(2)$	$W_1(1) + W_2(2) + W_3(2)$	$W_1(2) + W_2(2) + W_3(2)$	$p_{25}$	3
26	$W_1(2) + W_2(2) + W_3(2)$	$W_2(2) + W_3(2)$	$W_1(1) + W_2(2) + W_3(2)$	$p_{26}$	3
27	$W_1(1) + W_2(2) + W_3(2)$	$W_1(2) + W_2(2) + W_3(2)$	$W_2(2) + W_3(2)$	$p_{27}$	3

Note that if the query options are chosen equally likely, then the expected download cost normalized by message size  $L = 2$  is calculated as

$$\begin{aligned}
 D &= \frac{1}{L} \sum_{m=1}^M p_m d_m \\
 &= \frac{1}{2} \left( p_1 \cdot 2 + p_2 \cdot 2 + p_3 \cdot 2 + \sum_{m=4}^{27} (p_m \cdot 3) \right) \\
 &= \frac{1}{2} \cdot \frac{1}{27} (2 + 2 + 2 + 3 \cdot 24) = \frac{13}{9},
 \end{aligned}$$

where its reciprocal matches the PIR capacity result of (2.1) with  $N = 3$  and  $K = 3$ ,

$$C = \frac{1}{1 + \frac{1}{3} + \frac{1}{3^2}} = \frac{9}{13}.$$

### 3.3.5 Probabilistic PIR Scheme with General $N, K$

In this subsection, a sketch of the probabilistic TSC PIR query structure with general  $N$  databases and  $K$  messages is presented. Each message consists of  $L = N - 1$  symbols, and there are  $M = N^K$  query options that the user can choose. As shown in the previous examples, the first  $N$  options have the download cost of  $N - 1$  since each of them uses  $N - 1$  databases for downloading a single symbol, respectively. There are remaining  $N^K - N$  options with the download cost of  $N$ . They use  $N$  databases for downloading a single symbol or a sum of symbols, respectively. Rather than specific query realization, a sketch of query structure is presented in Table 3.4.



Table 3.4: A sketch of the probabilistic query structure of symmetric TSC scheme with  $N$  databases and  $K$  messages to retrieve  $W_1$

Option	Database 1	Database 2	...	Database $N$	Prob.	Cost
1	$W_1(1)$	$W_1(2)$	...	$\phi$	$p_1$	$N-1$
2	$\phi$	$W_2(1)$	...	$W_1(N-1)$	$p_2$	$N-1$
3	$W_1(N-1)$	$\phi$	...	$W_1(N-2)$	$p_3$	$N-1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	$W_1(2)$	$W_1(3)$	...	$W_1(1)$	$p_N$	$N-1$
$N+1$	$W_2(1)$	$W_1(1) + W_2(1)$	...	$W_1(N-1) + W_2(1)$	$p_{N+1}$	$N$
$N+2$	$W_1(N-1) + W_2(1)$	$W_2(1)$	...	$W_1(N-2) + W_2(1)$	$p_{N+2}$	$N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N^K - 1$	$W_1(2) + \sum_{i \in [2:K]} W_i(N-1)$	$W_1(3) + \sum_{i \in [2:K]} W_i(N-1)$	...	$W_1(1) + \sum_{i \in [2:K]} W_i(N-1)$	$p_{N^K-1}$	$N$
$N^K$	$W_1(1) + \sum_{i \in [2:K]} W_i(N-1)$	$W_1(2) + \sum_{i \in [2:K]} W_i(N-1)$	...	$\sum_{i \in [2:K]} W_i(N-1)$	$p_{N^K}$	$N$

Note that if the query options are chosen equally likely, then the expected download cost normalized by message size  $L = N - 1$  is calculated as

$$\begin{aligned}
D &= \frac{1}{L} \sum_{m=1}^M p_m d_m \\
&= \frac{1}{N-1} \left( \sum_{m=1}^N (N-1)p_m + \sum_{m=N+1}^{N^K} Np_m \right) \\
&= \frac{1}{N-1} \left( N \cdot \frac{N-1}{N^K} + (N^K - N) \cdot \frac{N}{N^K} \right) \\
&= \frac{1}{N^{K-1}} + (N + N^2 + \dots + N^{K-2} + N^{K-1}) \cdot \frac{N}{N^K} \\
&= \frac{1}{N^{K-1}} + \frac{1}{N^{K-2}} + \frac{1}{N^{K-3}} + \dots + \frac{1}{N} + 1,
\end{aligned}$$

where its reciprocal matches the PIR capacity result of (2.1) with general  $N$  and  $K$  given as

$$C = \frac{1}{1 + \frac{1}{N} + \frac{1}{N^2} + \dots + \frac{1}{N^{K-1}}}.$$

### 3.4 Optimal Tradeoff Between Information Leakage and Download Cost under the Kullback-Leibler Divergence

In this section, the examples from the previous section are revisited and formulated by convex optimization problems. By solving them, we will find the optimal tradeoff between the download cost and information leakage. Problems will be designed in the form of the problem in (3.5) as

$$\begin{aligned}
&\text{minimize} && \rho_{KL} = D_{KL}(P \parallel U) \\
&\text{subject to} && \frac{1}{L} \sum_{m=1}^M p_m d_m = D, \\
&&& \sum_{m=1}^M p_m = 1.
\end{aligned}$$

This is well-known cross-entropy minimization problem introduced by Kullback [28]. Since the divergence is measured from the uniform distribution  $U$  in the considering query structure, the cross-entropy minimization problem above is reduced to the following problem,

$$\begin{aligned}
& \text{minimize} && \rho_{KL} = H(U) - H(P) && (3.11) \\
& \text{subject to} && \frac{1}{L} \sum_{m=1}^M p_m d_m = D, \\
& && \sum_{m=1}^M p_m = 1.
\end{aligned}$$

Since the objective function in (3.11) is a summation of the negative entropy and a constant, it is convex on its domain, and the problem has a unique solution. In fact, this problem is the well-known entropy maximization problem. We solve the optimization problem for the probabilistic PIR scheme with arbitrary  $N$  and  $K$  and derive the analytic solution. After then the numerical result of the examples with graphical analysis will be given.

### 3.4.1 Optimization of Probability Distribution

Before we give the optimal tradeoff as a main result, we will solve the optimization problem in (3.5) for the case in Table 3.4. From the fact that the symmetric TSC scheme has  $N^K$  options as in (3.6), achieves perfect privacy by using the uniform distribution  $U$ , and has message length of  $L = N - 1$ , the problem in (3.5) is rewritten as

$$\begin{aligned}
& \text{minimize} && \sum_{m=1}^{N^K} p_m \log_2 p_m + \log_2 N^K \\
& \text{subject to} && \frac{1}{N-1} \sum_{m=1}^{N^K} p_m d_m = D, && (3.12) \\
& && \sum_{m=1}^{N^K} p_m = 1.
\end{aligned}$$

Since the objective function in (3.12) is a summation of the negative entropy and a constant, it is convex on its domain and has a unique solution. The Lagrangian taken from (3.12) is given as

$$\begin{aligned} \mathcal{L}(P, \lambda, \nu) &= \sum_{m=1}^{N^K} p_m \log_2 p_m + \log N^K \\ &+ \lambda \left( \frac{1}{N-1} \sum_{m=1}^{N^K} p_m d_m - D \right) + \nu \left( \sum_{m=1}^{N^K} p_m - 1 \right), \end{aligned}$$

where  $P = (p_1, \dots, p_{N^K})$  is the probability vector and  $\lambda$  and  $\nu$  are the Lagrange multipliers. Taking partial derivatives, we obtain

$$\frac{\partial \mathcal{L}}{\partial p_m} = \log_2 p_m + \frac{1}{\ln 2} + \frac{1}{N-1} \lambda d_m + \nu, m \in [1 : N^K], \quad (3.13)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \frac{1}{N-1} \sum_{m=1}^{N^K} p_m d_m - D,$$

$$\frac{\partial \mathcal{L}}{\partial \nu} = \sum_{m=1}^{N^K} p_m - 1. \quad (3.14)$$

Equating (3.13)-(3.14) with zero gives the solution to  $P$  as

$$p_m = \frac{1}{e^{(\frac{1}{N-1} \lambda d_m + \nu) \ln 2 + 1}}, m \in [1 : N^K], \quad (3.15)$$

where  $P = (p_1, \dots, p_{N^K})$  satisfies

$$\begin{aligned} \frac{1}{N-1} \sum_{m=1}^{N^K} \frac{d_m}{e^{(\frac{1}{N-1} \lambda d_m + \nu) \ln 2 + 1}} &= D \\ \Leftrightarrow \frac{1}{N-1} \sum_{m=1}^{N^K} \frac{d_m}{e^{\frac{1}{N-1} \lambda d_m \ln 2}} &= D e^{\nu \ln 2 + 1} \end{aligned} \quad (3.16)$$

and

$$\begin{aligned} \sum_{m=1}^{N^K} \frac{1}{e^{(\frac{1}{N-1} \lambda d_m + \nu) \ln 2 + 1}} &= 1 \\ \Leftrightarrow \sum_{m=1}^{N^K} \frac{1}{e^{\frac{1}{N-1} \lambda d_m \ln 2}} &= e^{\nu \ln 2 + 1}, \end{aligned} \quad (3.17)$$

respectively. For a given set of the possible download cost  $\{d_1, \dots, d_M\}$  and the target download cost  $D$ , we can solve (3.16) and (3.17) to find  $\lambda$  and  $\nu$ . After then, the optimal probability allocation (3.15) and corresponding information leakage  $\rho_{KL}$  in (3.3) can be obtained.

In general, solving for  $P = (p_1, \dots, p_{N^K})$  from the above procedure is not solvable in a closed-form equation explicitly. This observation is related to the fact that there exists no algebraic solution to the general quintic equation with arbitrary polynomial coefficients. Specifically, if  $N^K \geq 5$  with arbitrary  $d_1, \dots, d_{N^K}$ , we cannot solve (3.16) and (3.17) for  $\lambda$  and  $\nu$ , in a closed-form explicitly. Therefore, a numerical method can be a good alternative option.

However, since the probabilistic query structure of the symmetric TSC scheme has only two download costs, it is possible to solve the optimization problem explicitly. In the scheme, note that Step 1 has  $N$  options with downloads of  $N - 1$  symbols. Steps 2 to  $K$  have  $N^K - N$  options in total with downloads of  $N$  symbols. Therefore, we have

$$d_m = \begin{cases} N - 1, & m \in [1 : N] \\ N, & m \in [N + 1 : N^K]. \end{cases}$$

Now substituting (3.17) into (3.16), we have

$$\frac{1}{N - 1} \sum_{m=1}^{N^K} \frac{d_m}{e^{\frac{1}{N-1} \lambda d_m \ln 2}} = D \sum_{m=1}^{N^K} \frac{1}{e^{\frac{1}{N-1} \lambda d_m \ln 2}}. \quad (3.18)$$

Then from (3.18), we have

$$\begin{aligned}
& \frac{1}{N-1} \left\{ N \cdot \frac{N-1}{e^{\lambda \ln 2}} + (N^K - N) \cdot \frac{N}{e^{\frac{1}{N-1} \lambda N \ln 2}} \right\} \\
&= D \left\{ N \cdot \frac{1}{e^{\lambda \ln 2}} + (N^K - N) \cdot \frac{1}{e^{\frac{1}{N-1} \lambda N \ln 2}} \right\} \\
&\Leftrightarrow (N^K - N) \cdot \frac{N - (N-1)D}{e^{\frac{1}{N-1} \lambda N \ln 2}} = N \cdot \frac{(N-1)D - N + 1}{e^{\lambda \ln 2}} \\
&\Leftrightarrow (N^{K-1} - 1) \cdot \frac{1 - (N-1)(D-1)}{e^{\frac{1}{N-1} \lambda N \ln 2}} = \frac{(N-1)(D-1)}{e^{\lambda \ln 2}} \\
&\Leftrightarrow e^{\frac{1}{N-1} \lambda \ln 2} = \frac{(N^{K-1} - 1) \{1 - (N-1)(D-1)\}}{(N-1)(D-1)} \triangleq \Lambda, \tag{3.19}
\end{aligned}$$

where (3.19) forms a closed-form solution to  $\lambda$ . Note that  $\Lambda$  is used for ease of the notation. Further, we proceed to find  $\nu$ . Substituting (3.19) into (3.17) gives

$$\begin{aligned}
& \sum_{m=1}^M \frac{1}{e^{\frac{1}{N-1} \lambda d_m \ln 2}} = e^{\nu \ln 2 + 1} \\
&\Leftrightarrow \sum_{m=1}^M \frac{1}{\Lambda^{d_m}} = e^{\nu \ln 2 + 1} \\
&\Leftrightarrow \frac{N}{\Lambda^{N-1}} + \frac{N^K - N}{\Lambda^N} = e^{\nu \ln 2 + 1}, \tag{3.20}
\end{aligned}$$

where (3.20) forms a closed-form solution to  $\nu$  as well.

Now we are ready to show the solution to the optimal probability in (3.15). First, for  $m = 1, \dots, N$ , we have

$$\begin{aligned}
p_m &= \frac{1}{e^{(\frac{1}{N-1} \lambda d_m + \nu) \ln 2 + 1}} \\
&= \frac{1}{e^{(\lambda + \nu) \ln 2 + 1}} \\
&= \frac{1}{\Lambda^{N-1} \left( \frac{N}{\Lambda^{N-1}} + \frac{N^K - N}{\Lambda^N} \right)} \\
&= \frac{1}{N + \frac{N^K - N}{\Lambda}} \\
&= \frac{1}{N + N \frac{(N-1)(D-1)}{1 - (N-1)(D-1)}} \\
&= \frac{1 - (N-1)(D-1)}{N}
\end{aligned}$$

and for  $m = N + 1, \dots, N^K$ , we have

$$\begin{aligned}
p_m &= \frac{1}{e^{(\frac{1}{N-1}\lambda d_m + \nu) \ln 2 + 1}} \\
&= \frac{1}{e^{(\frac{N}{N-1}\lambda + \nu) \ln 2 + 1}} \\
&= \frac{1}{\Lambda^N \left( \frac{N}{\Lambda^{N-1}} + \frac{N^K - N}{\Lambda^N} \right)} \\
&= \frac{1}{\Lambda N + N^K - N} \\
&= \frac{1}{\frac{(N^K - N)\{1 - (N-1)(D-1)\}}{(N-1)(D-1)} + N^K - N} \\
&= \frac{(N-1)(D-1)}{N^K - N}.
\end{aligned}$$

The above findings are summarized in the following lemma.

**Lemma 3.1.** *The optimal solution to the optimization problem (3.5) on the symmetric TSC scheme is given as*

$$p_m = \begin{cases} \frac{1 - (N-1)(D-1)}{N}, & m \in [1 : N], \\ \frac{(N-1)(D-1)}{N^K - N}, & m \in [N + 1 : N^K]. \end{cases}$$

**Remark 4.** *The solution is a valid probability mass function. Trivially, one can easily find that  $\sum_m^{N^K} p_m = 1$ . The probability is always non-negative since as mentioned in (3.2), the range of  $D$  covers between two extreme cases, no information leakage and full information leakage. Therefore,  $1 \leq D \leq \frac{1}{C}$  and  $p_m$  is always non-negative.*

**Remark 5.** *The result we found in Lemma 3.1 has essentially the same meaning with the optimal probability distribution found by Theorem 1 in [21]. In fact, since we considered the symmetric version of the TSC scheme with  $N$  cyclic shifts, compared with [21], we obtain  $N$  times smaller probability with  $N$  times many queries. In other words, if we solve the optimization problem using KL divergence as the information leakage measure according to the problem setting they solved, exactly the same optimal probability distribution will be obtained as their theorem.*

### 3.4.2 Optimal Tradeoff Between Information Leakage and Download Cost

Now, we are ready to present the optimal tradeoff in PIR. The following theorem gives the optimal tradeoff between the information leakage measured in KL divergence  $\rho_{KL}$  and the expectation of normalized download cost  $D$  on the symmetric TSC scheme.

**Theorem 3.1.** *The optimal tradeoff between the information leakage measured by KL divergence  $\rho_{KL}$  and the expected normalized download cost  $D$  on the symmetric TSC PIR scheme with arbitrary  $N$  databases and  $K$  messages is given as*

$$\begin{aligned} \rho_{KL} = & \{1 - (N - 1)(D - 1)\} \log_2 \frac{1 - (N - 1)(D - 1)}{N} \\ & + (N - 1)(D - 1) \log_2 \frac{(N - 1)(D - 1)}{N^K - N} \\ & + \log_2 N^K, \end{aligned}$$

where the range of the expected normalized download cost is

$$1 \leq D \leq \frac{1}{C}.$$

Then the information leakage-download cost pairs establish the optimal tradeoff.

*Proof.* The proof is straightforward by substituting Lemma 3.1 in the objective function (3.12). The optimality of the tradeoff is obtained from the fact that the optimization problem we designed in (3.12) has its global optimum since it has the convex objective function and affine constraint functions.  $\square$

### 3.4.3 Numerical Analysis with Examples

In this subsection, we present graphical analyses on the numerical results of some examples previously considered,  $N = 2, K = 2$  case,  $N = 3, K = 2$  case, and  $N = 3, K = 3$  case. Curves showing the tradeoffs between information leakage and the download cost are shown in the Figures 3.1, 3.2, and 3.3. For all cases, the download costs shown in the figures are normalized with the desired message size  $L$ .



The two extreme points are located at the bottom right corner and the top left corner of each graph. The bottom right corner corresponds to the case where there is no information leakage, and the download cost for this point is the same as that obtained from the classical PIR result given as reciprocal of (2.1),

$$\frac{1}{C} = 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}.$$

In the three cases of examples,  $1/C$  are computed as

$$\begin{aligned} 1 + \frac{1}{2} &= \frac{3}{2} = 1.5, \\ 1 + \frac{1}{3} &= \frac{4}{3} \approx 1.3333, \\ 1 + \frac{1}{3} + \frac{1}{9} &= \frac{13}{9} \approx 1.4444, \end{aligned}$$

respectively, which agree with the graphical results.

The top left corner corresponds to the case where direct downloading is used without the need for privacy. Because direct downloading is used, the download cost for this point is 1 and information leakage shows its maximum value. Information leakage expressed in (3.7),  $H(U) - H(P)$  can be used to verify whether the information leakage occurring at this point is optimal, that is, its minimum possible value. Since direct downloading is the only option for the user to achieve the cost 1, the user has to decide which options to choose and with what probability to use among all possible direct downloadable options. They are  $N$  direct downloadable options out of the total  $N^K$  options. The solution to minimizing  $H(U) - H(P)$  is to have the uniform probability distribution for  $P$  with probability space having maximum cardinality. Therefore let

$$p_m = \begin{cases} \frac{1}{N}, & m = 1, \dots, N \\ 0, & \text{otherwise,} \end{cases}$$

and then  $H(U) - H(P)$  has its minimum value given as

$$\rho_{KL} = \log_2 N^K - \log_2 N,$$

which agrees with Theorem 3.1 when  $D = 1$ . Also, in the three cases of examples,  $\rho_{KL}$  are computed as

$$\begin{aligned}\log_2 2^2 - \log_2 2 &= 1, \\ \log_2 3^2 - \log_2 3 &= \log_2 3 \approx 1.5850, \\ \log_2 3^3 - \log_2 3 &= \log_2 9 \approx 3.1699,\end{aligned}$$

respectively, which agree with the graphical results.

We give different numerical analyses for the three cases of examples in Figures 3.4, 3.5, and 3.6. In the figures, the optimal information leakage and download cost for the probability of options with lower download cost are shown in the same graph. The probability is given in Lemma 3.1 and corresponds to the case for  $m \in [1 : N]$ . As the probability increases, we observe that the download cost decreases and information leakage increases. The range of the probability is from  $1/N^K$  to  $1/N$ , which corresponds to the case of no leakage case and maximum leakage, respectively.

### 3.5 Alternative Probabilistic PIR Scheme

In this section, we present an alternative probabilistic PIR scheme achieving a better tradeoff between information leakage and download cost within a certain range. The proposed alternative PIR scheme resembles the symmetric TSC scheme but with a smaller option size and a shorter message length. Since we have the symmetric TSC scheme with  $N = 3, K = 2$  from the previous section, We begin with the alternative PIR example with the case of  $N = 3, K = 2$  as in Table 3.5. For comparison with the conventional TSC scheme, refer to Table 3.2. Note that the total number of options is reduced from 9 to 6, and the message length  $L$  is reduced from 2 to 1. With uniform distribution to  $P = (p_1, \dots, p_6)$ , perfect privacy will be achieved since databases will be accessed with unbiased queries. However, unlike the symmetric TSC scheme, the

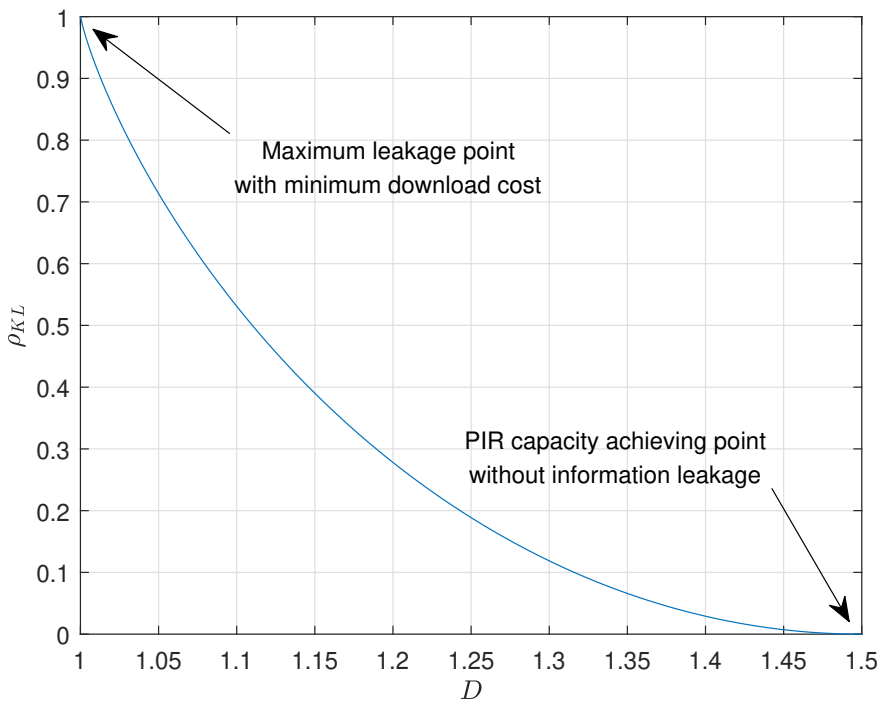


Figure 3.1: Optimal tradeoff between information leakage and the normalized download cost of symmetric TSC scheme with  $N = 2$ ,  $K = 2$ .

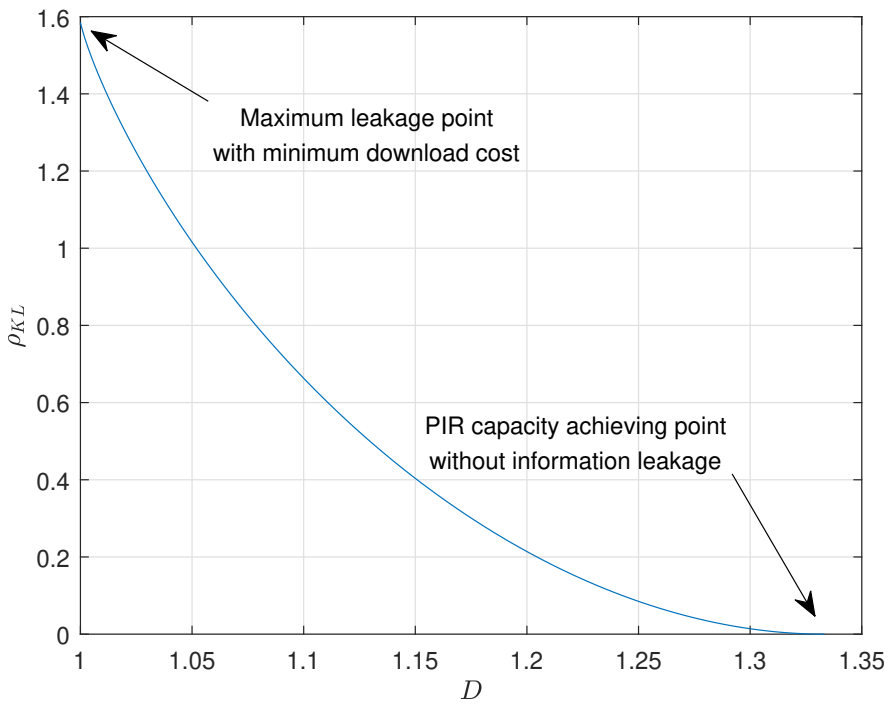


Figure 3.2: Optimal tradeoff between information leakage and the normalized download cost of symmetric TSC scheme with  $N = 3$ ,  $K = 2$ .

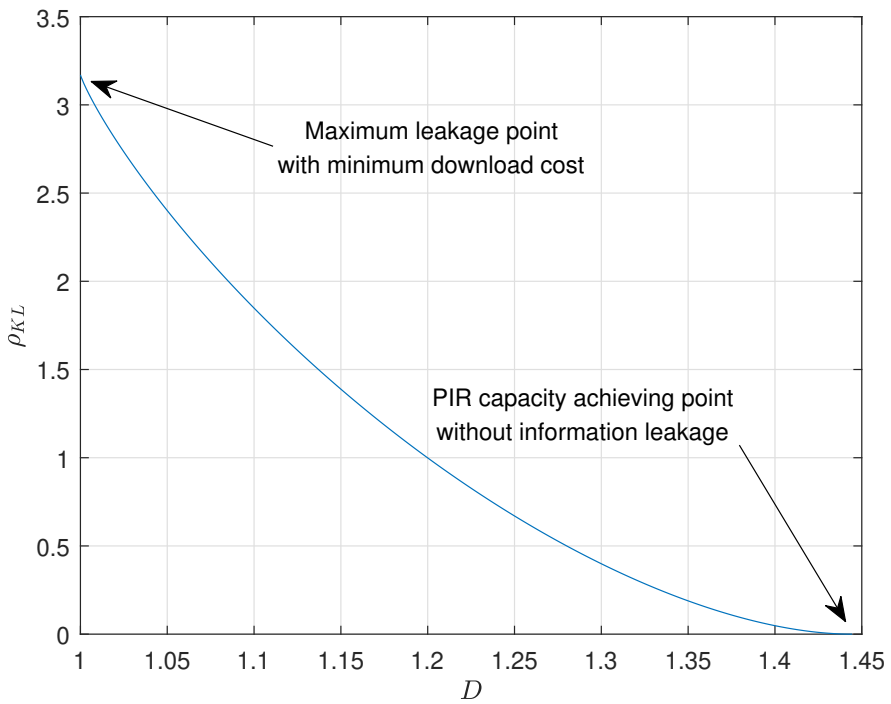


Figure 3.3: Optimal tradeoff between information leakage and the normalized download cost of symmetric TSC scheme with  $N = 3$ ,  $K = 3$ .

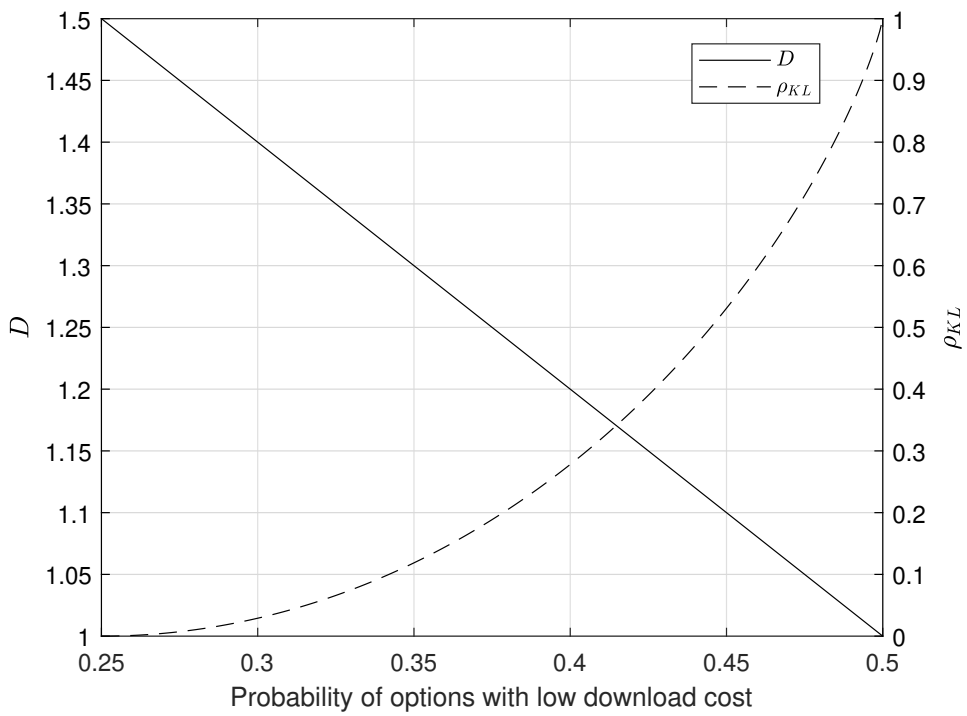


Figure 3.4: Optimal information leakage and download cost for the probability of lower download cost of symmetric TSC scheme with  $N = 2$ ,  $K = 2$ .

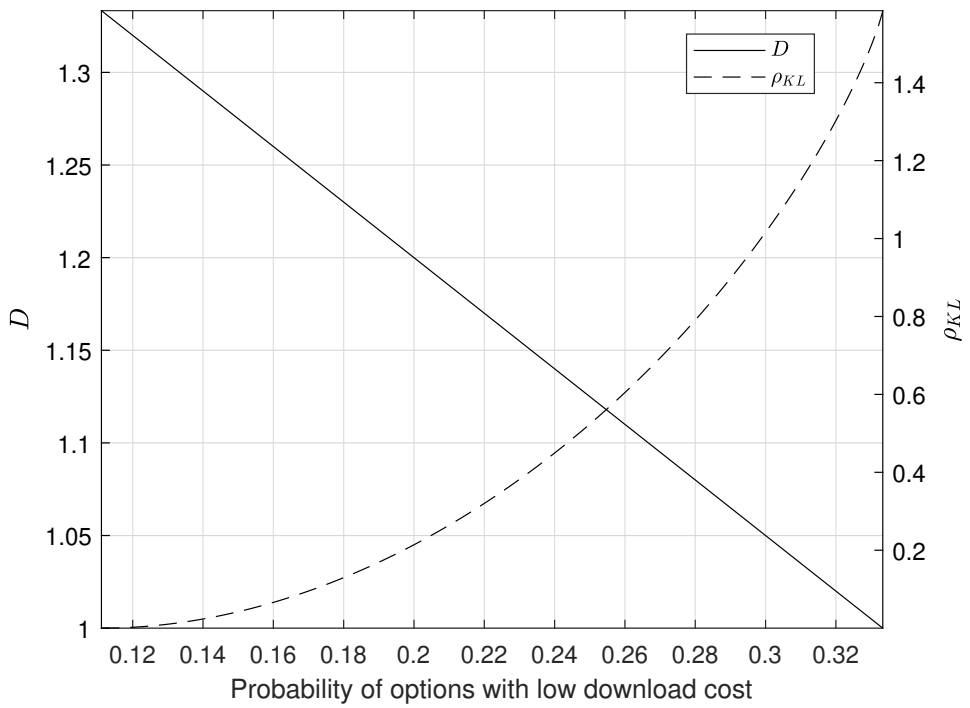


Figure 3.5: Optimal information leakage and download cost for the probability of lower download cost of symmetric TSC scheme with  $N = 3$ ,  $K = 2$ .

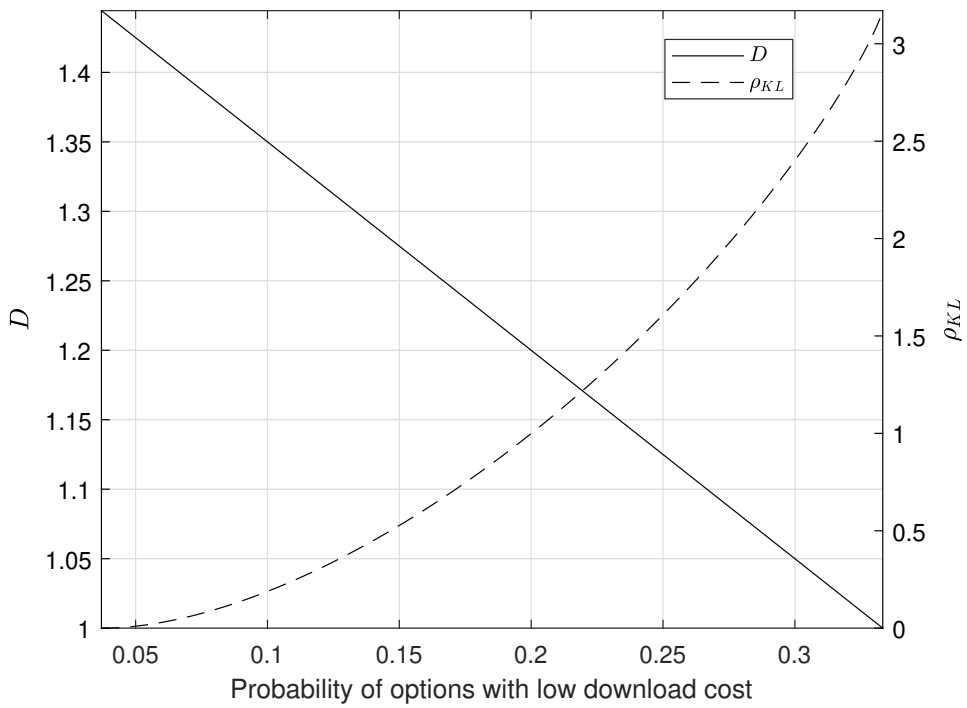


Figure 3.6: Optimal information leakage and download cost for the probability of lower download cost of symmetric TSC scheme with  $N = 3$ ,  $K = 3$ .



alternative PIR scheme with uniform distribution does not give the PIR capacity. This is simply verified as

$$D = \frac{1}{L} \sum_{m=1}^M p_m d_m = \frac{1}{6} \cdot (1 + 1 + 1 + 2 + 2 + 2) = \frac{3}{2},$$

which is strictly higher than  $D$  in (3.10). Therefore, the alternative PIR scheme seems to be undesirable. However, as the allowed information leakage increases, the alternative PIR scheme shows a more desirable tradeoff. We will first give a formal description of the alternative PIR scheme.

### 3.5.1 The Proposed Alternative PIR Scheme

The alternative PIR scheme we want to consider basically shares the same skeleton of the symmetric TSC scheme but has a smaller option size and a shorter message length. Let  $W_\theta$  be the desired message given as

$$W_\theta = [W_\theta(1), W_\theta(2), \dots, W_\theta(L)], \theta \in [1 : K],$$

where  $1 \leq L < N - 1$ . As in Section 3.3, the query structure of the alternative PIR scheme is explained as follows.

- (Step 1) Use first  $L$  databases to download desired message symbols

$$W_\theta(1), W_\theta(2), \dots, W_\theta(L),$$

respectively. Enumerate its cyclic shifts across databases. This step builds  $N$  query options.

- (Step 2) Download  $W_i(1)$  from the first database, where  $i \in [1 : K] \setminus \{\theta\}$ . Use the other  $L$  databases to download desired message symbols added to  $W_i(1)$ ,

Table 3.5: The probabilistic query structure of alternative PIR scheme with  $N = 3$ ,  $K = 2$  to retrieve  $W_1$

Option	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_1(1)$	$\phi$	$\phi$	$p_1$	1
2	$\phi$	$W_1(1)$	$\phi$	$p_2$	1
3	$\phi$	$\phi$	$W_1(1)$	$p_3$	1
4	$W_2(1)$	$W_1(1) + W_2(1)$	$\phi$	$p_4$	2
5	$\phi$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_5$	2
6	$W_1(1) + W_2(1)$	$\phi$	$W_2(1)$	$p_6$	2

that is,

$$\begin{aligned}
&W_\theta(1) + W_i(1) \\
&W_\theta(2) + W_i(1) \\
&\vdots \\
&W_\theta(L) + W_i(1),
\end{aligned}$$

respectively. Enumerate its cyclic shifts across databases. Up to the cyclic shifts we have  $N$  query options. Repeat for other symbols in  $W_i$  which are  $W_i(2), \dots, W_i(L)$ . Repeat for the other  $i \in [1 : K] \setminus \{\theta\}$ . This step builds  $NL \binom{K-1}{1}$  query options.

- (Step 3) Download  $W_i(1) + W_j(1)$  from the first database, where  $i, j \in [1 : K] \setminus \{\theta\}$  and  $i \neq j$ . Use the other  $L$  databases to download desired message symbols added to  $W_i(1) + W_j(1)$ , that is,

$$\begin{aligned}
&W_\theta(1) + W_i(1) + W_j(1) \\
&W_\theta(2) + W_i(1) + W_j(1) \\
&\vdots \\
&W_\theta(L) + W_i(1) + W_j(1),
\end{aligned}$$

respectively. Enumerate its cyclic shifts across databases. Up to the cyclic shifts we have  $N$  query options. Repeat for other symbols in  $W_i$  and  $W_j$  which are in total  $L^2$  multiple cases. Repeat for the other  $i, j$  where  $i, j \in [1 : K] \setminus \{\theta\}$  and  $i \neq j$ . This step builds  $NL^2 \binom{K-1}{2}$  query options.

- Repeat the steps with the same procedure until it reaches Step  $K$ . Step  $K$  builds  $NL^{K-1} \binom{K-1}{K-1}$  query options.

In the query structure, the queries generated in Step 1 trivially request the desired symbols only. It is obvious that the user can have the desired message directly. For the

queries generated from Step 2 to Step  $K$ , the user can subtract the undesired symbol or the sum of the undesired symbols from received symbols then recover the desired message.

By adding up the number of query options built from each step, the number of possible options for the proposed alternative PIR scheme can be calculated as

$$\begin{aligned}
 & N + NL \binom{K-1}{1} + \dots + NL^{K-1} \binom{K-1}{K-1} \\
 &= N \sum_{k=0}^{K-1} L^k \binom{K-1}{k} \\
 &= N(L+1)^{K-1}.
 \end{aligned}$$

Therefore, there are  $N(L+1)^{K-1}$  options that the user can take. Since we assume  $L < N-1$ , the number of total options in this scheme is less than  $N^K$ .

Another examples of the alternative PIR scheme to retrieve the first message  $W_1$  are shown in Tables 3.6 and 3.7 for  $N = 3, K = 3$  and  $N = 4, K = 2$ .  $L = 1$  is used for the case of  $N = 3, K = 3$  and  $W_1 = W_1(1)$ . Note that for the case of  $N = 4, K = 2$  in Table 3.7, alternative schemes for two possible message sizes  $L = 2$  and  $L = 1$  are shown.  $W_1$  is denoted as  $W_1 = [W_1(1), W_1(2)]$  for  $L = 2$  and  $W_1 = W_1(1)$  for  $L = 1$ , respectively.

Table 3.6: The probabilistic query structure of alternative PIR scheme with  $N = 3$ ,  $K = 3$  to retrieve  $W_1$

Opt.	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_1(1)$	$\phi$	$\phi$	$p_1$	1
2	$\phi$	$W_1(1)$	$\phi$	$p_2$	1
3	$\phi$	$\phi$	$W_1(1)$	$p_3$	1
4	$W_2(1)$	$W_1(1) + W_2(1)$	$\phi$	$p_4$	2
5	$\phi$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_5$	2
6	$W_1(1) + W_2(1)$	$\phi$	$W_2(1)$	$p_6$	2
7	$W_3(1)$	$W_1(1) + W_3(1)$	$\phi$	$p_7$	2
8	$\phi$	$W_3(1)$	$W_1(1) + W_3(1)$	$p_8$	2
9	$W_1(1) + W_3(1)$	$\phi$	$W_3(1)$	$p_9$	2
10	$W_2(1) + W_3(1)$	$W_1(1) + W_2(1) + W_3(1)$	$\phi$	$p_{10}$	2
11	$\phi$	$W_2(1) + W_3(1)$	$W_1(1) + W_2(1) + W_3(1)$	$p_{11}$	2
12	$W_1(1) + W_2(1) + W_3(1)$	$\phi$	$W_2(1) + W_3(1)$	$p_{12}$	2

Table 3.7: The probabilistic query structure of alternative PIR scheme with  $N = 4$ ,  $K = 2$  to retrieve  $W_1$

(a) With message size  $L = 2$

Opt.	Database 1	Database 2	Database 3	Database 4	Prob.	Cost
1	$W_1(1)$	$W_1(2)$	$\phi$	$\phi$	$p_1$	2
2	$\phi$	$W_1(1)$	$W_1(2)$	$\phi$	$p_2$	2
3	$\phi$	$\phi$	$W_1(1)$	$W_1(2)$	$p_3$	2
4	$W_1(2)$	$\phi$	$\phi$	$W_1(1)$	$p_4$	2
5	$W_2(1)$	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$\phi$	$p_5$	3
6	$W_2(2)$	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$\phi$	$p_6$	3
7	$\phi$	$W_2(1)$	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$p_7$	3
8	$\phi$	$W_2(2)$	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$p_8$	3
9	$W_1(2) + W_2(1)$	$\phi$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_9$	3
10	$W_1(2) + W_2(2)$	$\phi$	$W_2(2)$	$W_1(1) + W_2(2)$	$p_{10}$	3
11	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$\phi$	$W_2(1)$	$p_{11}$	3
12	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$\phi$	$W_2(2)$	$p_{12}$	3

(b) With message size  $L = 1$

Opt.	Database 1	Database 2	Database 3	Database 4	Prob.	Cost
1	$W_1(1)$	$\phi$	$\phi$	$\phi$	$p_1$	1
2	$\phi$	$W_1(1)$	$\phi$	$\phi$	$p_2$	1
3	$\phi$	$\phi$	$W_1(1)$	$\phi$	$p_3$	1
4	$\phi$	$\phi$	$\phi$	$W_1(1)$	$p_4$	1
5	$W_2(1)$	$W_1(1) + W_2(1)$	$\phi$	$\phi$	$p_5$	2
6	$\phi$	$W_2(1)$	$W_1(1) + W_2(1)$	$\phi$	$p_6$	2
7	$\phi$	$\phi$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_7$	2
8	$W_1(1) + W_2(1)$	$\phi$	$\phi$	$W_2(1)$	$p_8$	2

### 3.5.2 Alternative Optimal Tradeoff Between Information Leakage and Download Cost

Now, as in (3.12), we can solve the following optimization problem with a smaller option size and a shorter message length,

$$\begin{aligned}
 & \text{minimize} && \sum_{m=1}^{N(L+1)^{K-1}} p_m \log_2 p_m + \log_2 N(L+1)^{K-1} \\
 & \text{subject to} && \frac{1}{L} \sum_{m=1}^{N(L+1)^{K-1}} p_m d_m = D, \\
 & && \sum_{m=1}^{N(L+1)^{K-1}} p_m = 1.
 \end{aligned} \tag{3.21}$$

Solving the problem in (3.21) is similar to (3.12) but with different download cost for each option of

$$d_m = \begin{cases} L, & m \in [1 : N] \\ L+1, & m \in [N+1 : N(L+1)^{K-1}]. \end{cases}$$

We omit the detailed solving procedure for the optimal probability and present the following lemma.

**Lemma 3.2.** *The optimal solution to the optimization problem in (3.21) on the alternative PIR scheme is given as*

$$p_m = \begin{cases} \frac{1-L(D-1)}{N}, & m \in [1 : N] \\ \frac{L(D-1)}{N(L+1)^{K-1}-N}, & m \in [N+1 : N(L+1)^{K-1}], \end{cases}$$

where  $1 \leq L < N-1$ .

From Lemma 2, we have the following theorem for the alternative PIR scheme. We simply denote the information leakage in the alternative PIR scheme as  $\rho_{KL}^{alt}$ .

**Theorem 3.2.** *The optimal tradeoff between the information leakage measured by KL divergence  $\rho_{KL}^{alt}$  and the expected normalized download cost  $D$  on the alternative PIR scheme with arbitrary  $N$  databases and  $K$  messages is given as*

$$\begin{aligned}\rho_{KL}^{alt} = & \{1 - L(D - 1)\} \log_2 \frac{1 - L(D - 1)}{N} \\ & + L(D - 1) \log_2 \frac{L(D - 1)}{N(L + 1)^{K-1} - N} \\ & + \log_2 N(L + 1)^{K-1},\end{aligned}$$

where the range of  $D$  is

$$1 \leq D \leq \frac{1}{C},$$

and the message size  $L$  satisfies

$$1 \leq L < N - 1.$$

Then the information leakage-download cost pairs establish the optimal tradeoff.

*Proof.* The proof is straightforward as in Theorem 3.1. The optimality of the tradeoff is obtained by convex optimization as well.  $\square$

### 3.5.3 Numerical Analysis of the Proposed Alternative Scheme

By using Theorems 3.1 and 3.2, we compare the optimal tradeoffs achieved from two PIR schemes, that is, the symmetric TSC scheme and the alternative PIR scheme. Figure 3.7 shows the case of  $N = 3, K = 2$ . Note that there exists a specific range such that  $\rho_{KL}^{alt} < \rho_{KL}$  for the given download cost. In other words, alternative PIR scheme achieves lesser download cost than the conventional symmetric TSC scheme for the given information leakage. Figure shows that when download cost is prioritized over information leakage, the alternative PIR scheme is more desirable. However, since two PIR schemes we consider have different option sizes of  $N^K$  and  $N(L+1)^{K-1}$ , respectively, a normalized version of the KL divergence can be used for a fair comparison



between the symmetric TSC scheme and the alternative PIR scheme. Thus, (3.3) can be rewritten as

$$\bar{\rho}_{KL} = \frac{D_{KL}(P \| U)}{H(U)},$$

and then in this scheme,  $\bar{\rho}_{KL} = \rho_{KL}/\log_2 N^K$ . Similarly for the alternative PIR scheme,  $\bar{\rho}_{KL}^{alt} = \rho_{KL}^{alt}/\log_2 N(L+1)^{K-1}$  will be used. Figure 3.8 shows the same example of  $N = 3, K = 2$  with normalized information leakage. Note that there is still some range where the alternative PIR scheme is better in terms of information leakage, that is,  $\bar{\rho}_{KL}^{alt} < \bar{\rho}_{KL}$  for a fixed download cost. Figures 3.9 and 3.10 are represented for the case of  $N = 3, K = 3$  as well.

We present another example with  $N = 4, K = 2$ . As mentioned earlier, since  $1 \leq L < N - 1$ , there are two possible message sizes,  $L = 1$  and  $L = 2$ . We present the optimal tradeoffs in Figure 3.11 for both cases of the alternative PIR scheme with the conventional symmetric TSC scheme. Likewise, their normalized information leakage versions are shown in Figure 3.12. In both figures, there exist ranges in download costs such that  $\rho_{KL}^{alt} < \rho_{KL}$  and  $\bar{\rho}_{KL}^{alt} < \bar{\rho}_{KL}$ . Note that as the message size  $L$  becomes shorter, the alternative PIR scheme performs better when the amount of allowed information leakage is relaxed. In a practical scenario, therefore, a user might want to choose between  $N - 1$  possible schemes, including the conventional symmetric TSC scheme according to the specified information leakage allowance.

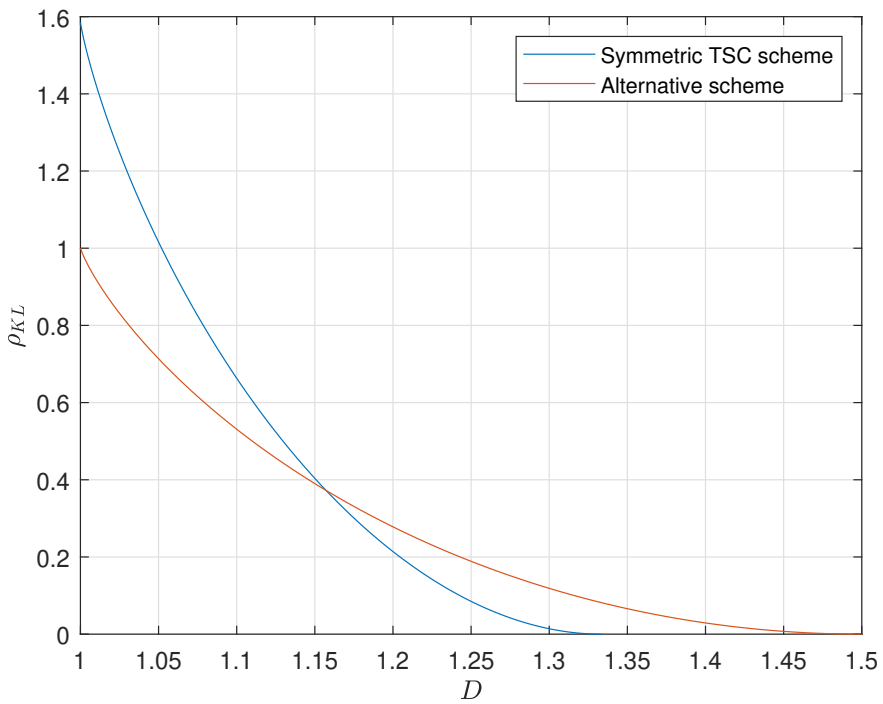


Figure 3.7: Optimal tradeoff between information leakage and the download cost for two PIR schemes with  $N = 3$ ,  $K = 2$ .

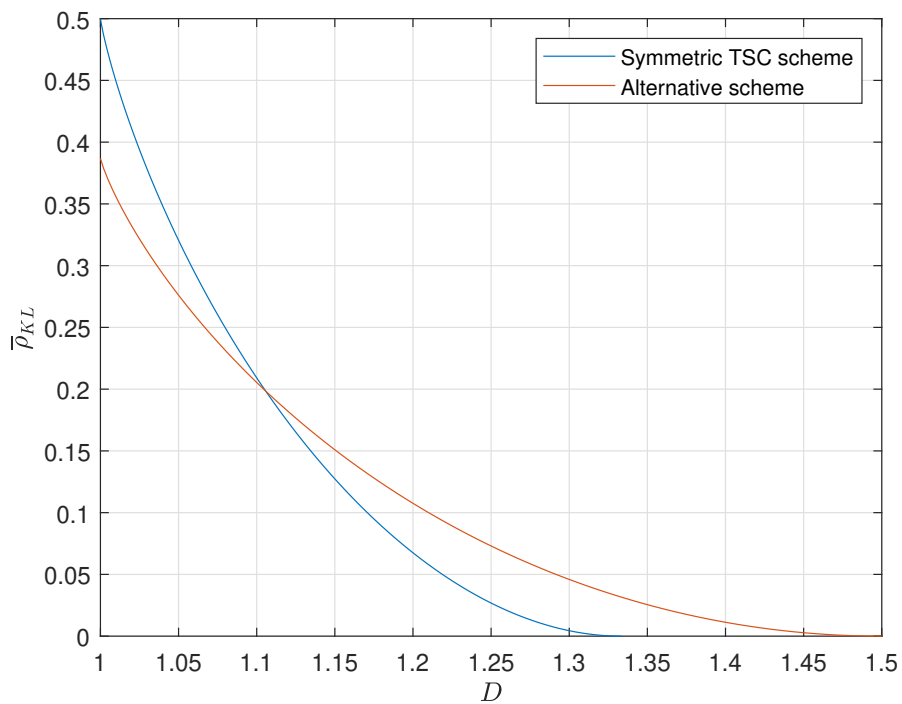


Figure 3.8: Optimal tradeoff between normalized information leakage and the download cost for two PIR schemes with  $N = 3$ ,  $K = 2$ .

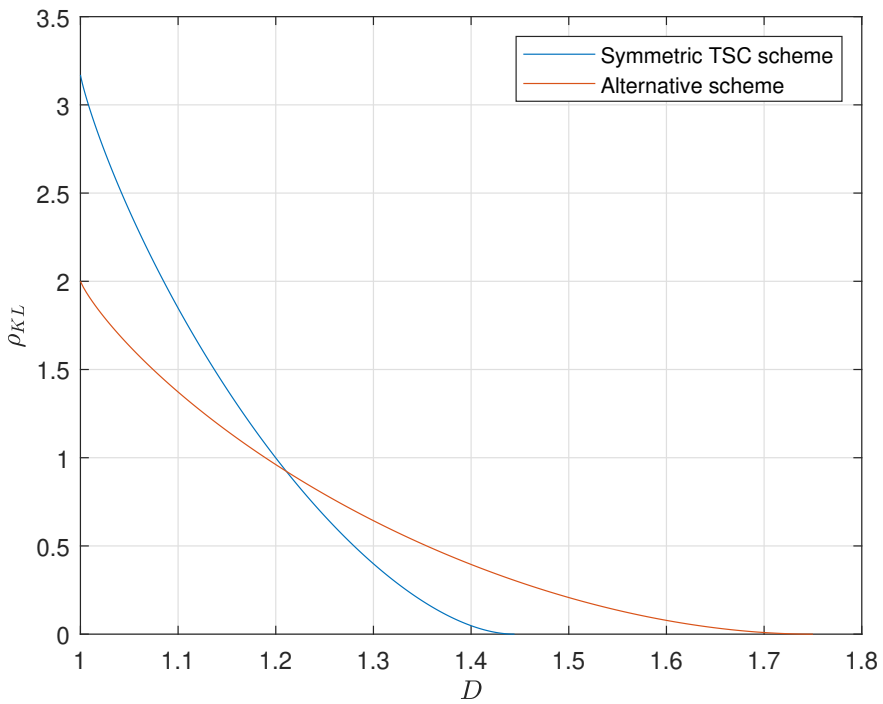


Figure 3.9: Optimal tradeoff between information leakage and the download cost for two PIR schemes with  $N = 3$ ,  $K = 3$ .

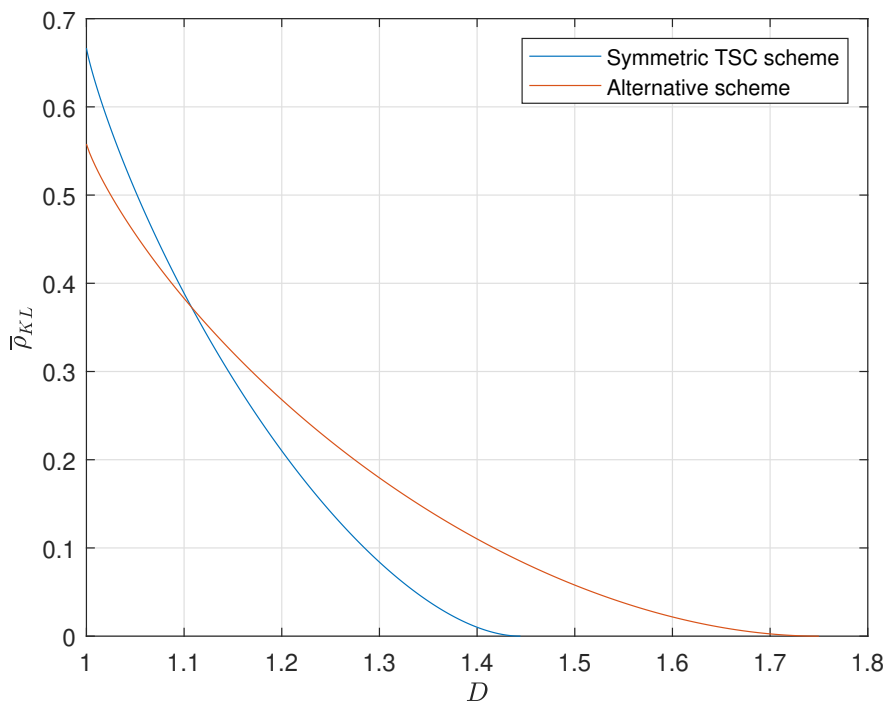


Figure 3.10: Optimal tradeoff between normalized information leakage and the download cost for two PIR schemes with  $N = 3$ ,  $K = 3$ .

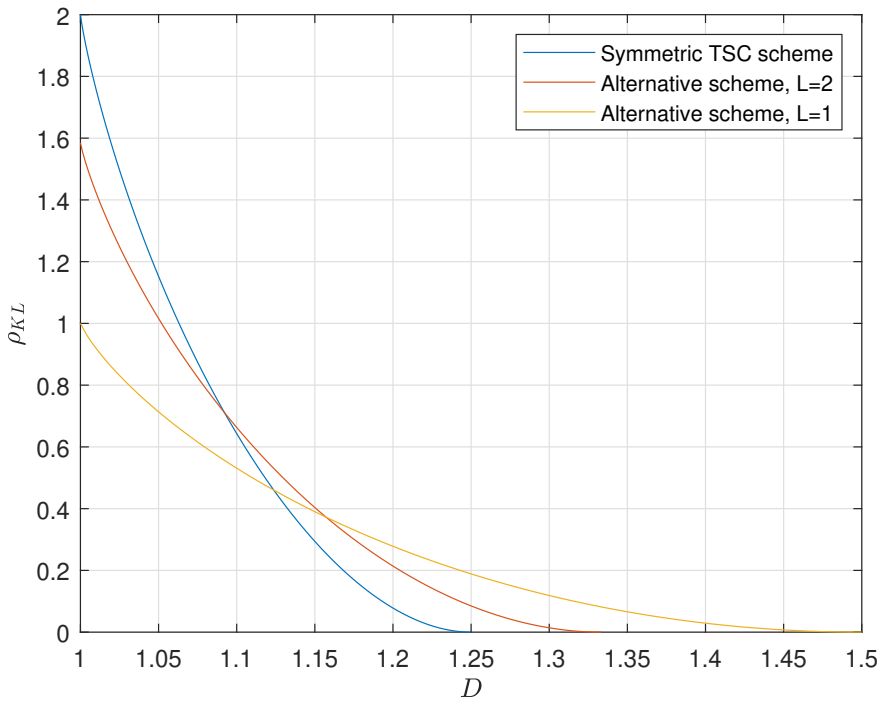


Figure 3.11: Optimal tradeoff between information leakage and the download cost for two PIR schemes with  $N = 4$ ,  $K = 2$ .

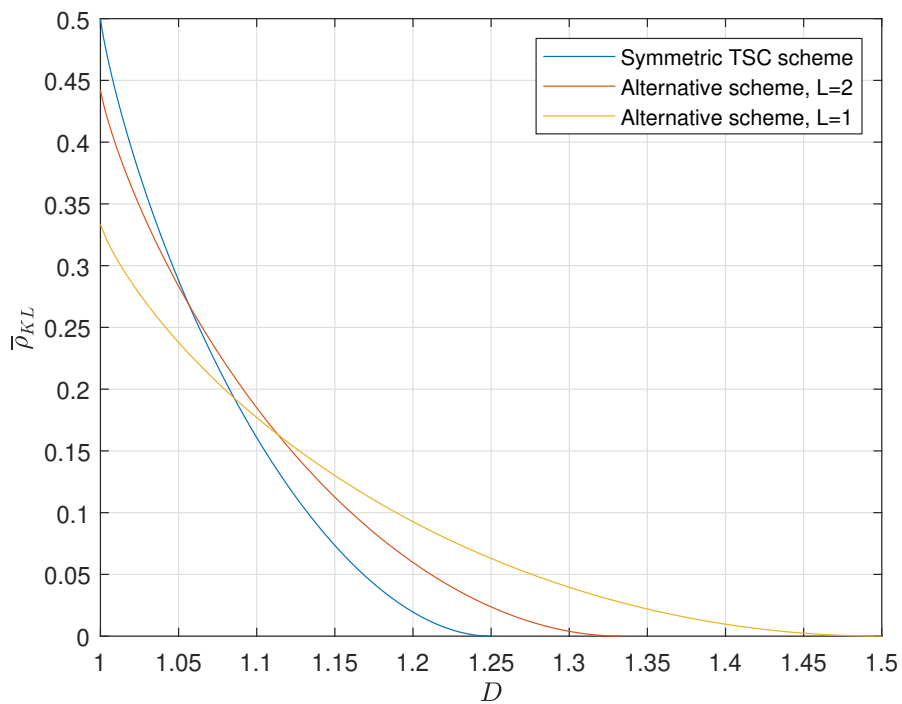


Figure 3.12: Optimal tradeoff between normalized information leakage and the download cost for two PIR schemes with  $N = 4$ ,  $K = 2$ .

## Chapter 4

# PIR with Information Leakage under the Jensen-Shannon Divergence

### 4.1 Introduction

In this chapter, a PIR problem with information leakage with the same purpose as in the previous chapter is introduced but under a different information leakage measure. In the problem setting and the optimization process in the previous chapter, there was a probability distribution  $U$  as a reference uniform distribution indicating no information leakage in the PIR system. The Kullback-Leibler (KL) divergence from the reference probability distribution  $U$  to arbitrary probability distribution  $P$  was measured to show how much the information leakage is occurring when the user chooses  $P$  instead of  $U$ . We have the same probabilistically generated query model, and the information leakage is measured in a more generalized way. As will be explained in more detail later, it is no longer necessary to have the predefined probability distribution  $U$ .

Here we propose the Jensen-Shannon (JS) divergence as a measure of the information leakage, which is based on the KL divergence. The JS divergence has its nomenclature since it is in the form of the Jensen's inequality applied to Shannon's entropy. Before we begin to explore the detailed problem setting, the formal definition of the



JS divergence is given as follows.

**Definition 4.1.** *The Jensen-Shannon (JS) divergence between two probability distributions  $P_1$  and  $P_2$  is defined as*

$$D_{JS,\pi}(P_1, P_2) = H(\pi_1 P_1 + \pi_2 P_2) - \pi_1 H(P_1) - \pi_2 H(P_2) \quad (4.1)$$

where  $\pi = (\pi_1, \pi_2)$  is the weight, or prior of  $P_1$  and  $P_2$ , respectively, such that  $\pi_1, \pi_2 \geq 0, \pi_1 + \pi_2 = 1$ . With little efforts, (4.1) can be expressed equivalently using the KL divergence as

$$D_{JS,\pi}(P_1, P_2) = \pi_1 D_{KL}(P_1 \parallel M) + \pi_2 D_{KL}(P_2 \parallel M), \quad (4.2)$$

where  $M = \pi_1 P_1 + \pi_2 P_2$ .

If uniform  $\pi$  is used, that is,  $\pi = (\pi_1, \pi_2) = (\frac{1}{2}, \frac{1}{2})$ , then (4.1) and (4.2) are simply

$$D_{JS}(P_1, P_2) = H\left(\frac{1}{2}P_1 + \frac{1}{2}P_2\right) - \frac{1}{2}H(P_1) - \frac{1}{2}H(P_2)$$

and

$$D_{JS}(P_1, P_2) = \frac{1}{2}D_{KL}(P_1 \parallel M) + \frac{1}{2}D_{KL}(P_2 \parallel M),$$

respectively, where  $M = \frac{P_1 + P_2}{2}$ .

Note that by definition  $D_{JS,\pi}$  is symmetric with its elements, that is,

$$D_{JS,\pi}(P_1, P_2) = D_{JS,\pi}(P_2, P_1).$$

Also, a well-known property of the Jensen-Shannon divergence is that it has the bounded range [27]. When the logarithm with base 2 is used, then

$$0 \leq D_{JS,\pi}(P_1, P_2) \leq 1$$

and with base  $e$ ,

$$0 \leq D_{JS,\pi}(P_1, P_2) \leq \ln 2.$$

Also, note that  $D_{JS,\pi}(P_1, P_2) = 0$  if and only if  $P_1 = P_2$ .

Unlike KL divergence having two probability distributions as variables, the JS divergence has the generalized version about more than two probability distributions. We will find the generalized version useful for our problem configuration. The definition of the generalized Jensen-Shannon divergence is given below.

**Definition 4.2.** *The generalized Jensen-Shannon (JS) divergence among  $K$  probability distributions  $P_1, \dots, P_K$  is defined as*

$$D_{JS,\pi}(P_1, \dots, P_K) = H\left(\sum_{k=1}^K \pi_k P_k\right) - \sum_{k=1}^K \pi_k H(P_k), \quad (4.3)$$

where  $\pi = (\pi_1, \dots, \pi_K)$  is the weight, or prior of  $P_1, \dots, P_K$ , respectively, such that  $\pi_k \geq 0$ ,  $\sum_{k=1}^K \pi_k = 1$ . Again, (4.3) can be expressed equivalently using KL divergence as

$$D_{JS,\pi}(P_1, \dots, P_K) = \sum_{k=1}^K \pi_k D_{KL}(P_k \parallel M), \quad (4.4)$$

where  $M = \sum_{k=1}^K \pi_k P_k$ .

If uniform  $\pi$  is used, then (4.3) and (4.4) are simply

$$D_{JS}(P_1, \dots, P_K) = H\left(\frac{1}{K} \sum_{k=1}^K P_k\right) - \frac{1}{K} \sum_{k=1}^K H(P_k) \quad (4.5)$$

and

$$D_{JS}(P_1, \dots, P_K) = \frac{1}{K} \sum_{k=1}^K D_{KL}(P_k \parallel M), \quad (4.6)$$

respectively, where  $M = \frac{1}{K} \sum_{k=1}^K P_k$ .

Also, the generalized version of the JS divergence is symmetric with its elements and has the bounded range. With logarithm base 2, we have

$$0 \leq D_{JS,\pi}(P_1, \dots, P_K) \leq \log_2 K$$

and with base  $e$ ,

$$0 \leq D_{JS,\pi}(P_1, \dots, P_K) \leq \ln K.$$

Also, note that  $D_{JS,\pi}(P_1, \dots, P_K) = 0$  if and only if  $P_1 = P_2 = \dots = P_K$ .

## 4.2 Problem Formulation under the Jensen-Shannon Divergence

The problem we want to solve is almost the same as the previous one. There are  $N$  replicated databases that each of them storing all  $K$  messages with equal size  $L$ . Table 4.1 shows a probabilistic query structure with  $M$  options to retrieve the message  $W_\theta$ . Note that the query sent to the  $n$ -th database by using  $m$ -th option to retrieve the message  $W_\theta$  is denoted by  $Q_n^{[\theta]}(m)$  and its corresponding answer is  $A_n^{[\theta]}(m)$ . Any choice of option should give the user the desired message  $W_\theta$ , which can be written as

$$H(W_\theta | Q_{[1:N]}^{[\theta]}(m), A_{[1:N]}^{[\theta]}(m)) = 0, \quad m \in [1 : M],$$

which is called correctness condition.  $M$  options with probabilities of  $p_1^{[\theta]}, \dots, p_M^{[\theta]}$  to be chosen by the user have their corresponding download costs  $d_1^{[\theta]}, \dots, d_M^{[\theta]}$ , respectively. The download cost  $d_m^{[\theta]}$  of  $m$ -th option is summation of the answer sizes across the databases given as

$$d_m^{[\theta]} = \sum_{n=1}^N H(A_n^{[\theta]}(m)).$$

Again the performance measure of the probabilistic PIR model is the expectation of the normalized download cost and computed as

$$D = \frac{1}{L} \sum_{m=1}^M p_m^{[\theta]} d_m^{[\theta]}.$$

Table 4.1: A probabilistic PIR query structure to retrieve  $W_\theta$

Option	Database 1	...	Database $N$	Probability	Download cost
1	$Q_1^{[\theta]}(1)$	...	$Q_N^{[\theta]}(1)$	$p_1^{[\theta]}$	$d_1^{[\theta]}$
2	$Q_1^{[\theta]}(2)$	...	$Q_N^{[\theta]}(2)$	$p_2^{[\theta]}$	$d_2^{[\theta]}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$M$	$Q_1^{[\theta]}(M)$	...	$Q_N^{[\theta]}(M)$	$p_M^{[\theta]}$	$d_M^{[\theta]}$

As mentioned before, the range of  $D$  is as in (3.2), that is,  $1 \leq D \leq 1/C$ , where  $C$  is the capacity of PIR without information leakage.

Note that we denoted the probability  $p_m^{[\theta]}$  and the download cost  $d_m^{[\theta]}$  including its concerning message index  $\theta$  unlike the previous ones in Table 2.2. Now define the probability distributions for query sent to the database  $n$  to retrieve the message  $W_\theta$  as

$$P_{Q_n^{[\theta]}}(q) = P(Q_n^{[\theta]} = q), \quad n \in [1 : N], \theta \in [1 : K], q \in \mathcal{Q},$$

where  $\mathcal{Q}$  with the cardinality  $|\mathcal{Q}| = M$  is the set of all possible query realizations that the user can ask.

Now we define the information leakage denoted as  $\rho_{JS}$  measured by JS divergence at the  $n$ -th database as

$$\rho_{JS} = D_{JS} \left( P_{Q_n^{[1]}}, \dots, P_{Q_n^{[K]}} \right). \quad (4.7)$$

Note that (4.5) or (4.6) is used instead of (4.3) or (4.4) since we assume that messages are requested uniformly by the user, that is, there is no bias in requesting messages. From a system-wide perspective, when an option is selected, then the corresponding query is automatically determined across databases. If the query structure is symmetric across databases, the information leakage defined in (4.7) is identical across databases. Therefore, if it is the case, it is sufficient to consider the information leakage only at one arbitrary database.

The goal of the problem setting for the probabilistic PIR with JS divergence is to find the most efficient probability distributions  $P_{Q_n^{[k]}}$  for all  $k \in [1 : K]$  that minimizes the information leakage measured by  $\rho_{JS}$  given a certain amount of target download cost. Obviously, the download cost must be the same no matter what message the user wants for fairness and privacy. As in the previous chapter, we want to find the optimal tradeoff between the download cost of the PIR system and the information leakage measured in the JS divergence. The problem is written as a convex optimization problem as follows:

$$\begin{aligned}
& \text{minimize} && \rho_{JS} = D_{JS}(P_{Q_n^{[1]}}, \dots, P_{Q_n^{[K]}}) \\
& \text{subject to} && \frac{1}{L} \sum_{m=1}^M p_m^{[k]} d_m^{[k]} = D, \quad k \in [1 : K] \\
& && \sum_{m=1}^M p_m^{[k]} = 1, \quad k \in [1 : K].
\end{aligned}$$

## 4.3 Achievable Scheme under the Jensen-Shannon Divergence

### 4.3.1 Probabilistic Query Generation

We use the same probabilistic PIR query structure of Section 3.3. Its queries are symmetric across all the databases, and all possible queries appear in each database with probability, which is desirable in our case. Again, only the query structure is adopted without different semantics between messages. If the different popularity between messages is of interest, we can use non-uniform  $\pi$  in the definition of JS divergence. However, it is not the scope of this dissertation. Refer to Section 3.3 for the detailed query structure. We start with the simplest example to capture the main idea.

### 4.3.2 Example of Symmetric TSC Scheme with $N = 2, K = 2$

Consider the simplest PIR example with  $N = 2$  databases and  $K = 2$  messages. The message size is  $L = N - 1 = 1$  and two messages are denoted as  $W_1 = W_1(1)$  and  $W_2 = W_2(1)$ , respectively. The number of query options is  $M = N^K = 4$ . The probabilistic query structures to retrieve  $W_1$  and  $W_2$  are given in Table 4.2. Any choice of options in the query structure will give the user arbitrary desired messages for sure, but possibly with different download cost 1 or 2. The expected normalized download costs to retrieve  $W_1$  and  $W_2$  can be computed from the tables. No matter what message is wanted, the expectation of the cost must be the same. Denote the download cost as  $D$ , and then we have

$$D = \frac{1}{L} \sum_{m=1}^4 p_m^{[1]} d_m^{[1]} = p_1^{[1]} \cdot 1 + p_2^{[1]} \cdot 1 + p_3^{[1]} \cdot 2 + p_4^{[1]} \cdot 2, \quad (4.8)$$

and

$$D = \frac{1}{L} \sum_{m=1}^4 p_m^{[2]} d_m^{[2]} = p_1^{[2]} \cdot 1 + p_2^{[2]} \cdot 1 + p_3^{[2]} \cdot 2 + p_4^{[2]} \cdot 2, \quad (4.9)$$

respectively. Adjusting the probability distributions  $(p_1^{[k]}, \dots, p_M^{[k]})$ ,  $k = 1, 2$  will determine  $D \in [1, 1/C]$ , where  $C$  is the capacity of the PIR without information leakage.

Tables 4.2 (a) and 4.2 (b) are analyzed together to investigate the information leakage generated in the database 1. Due to the symmetry of the query structure, that is, every possible query appears once in each of the databases, minimizing the information leakage in the database 1 will suffice to optimize the entire system. For ease of handling the dissimilarity seen at the database 1, Tables 4.2 (a) and 4.2 (b) are recasted in Table 4.3. It shows the information leakage occurring in the database 1. Four possible queries are requested with two different probability distributions according to the index of the message being retrieved. This difference or dissimilarity is measured by

Table 4.2: The probabilistic query structure of symmetric TSC scheme with  $N = 2, K = 2$

(a) To retrieve  $W_1$

Option	Database 1	Database 2	Probability	Download cost, $d_m^{[1]}$
1	$W_1(1)$	$\phi$	$p_1^{[1]}$	1
2	$\phi$	$W_1(1)$	$p_2^{[1]}$	1
3	$W_2(1)$	$W_1(1) + W_2(1)$	$p_3^{[1]}$	2
4	$W_1(1) + W_2(1)$	$W_2(1)$	$p_4^{[1]}$	2

(b) To retrieve  $W_2$

Option	Database 1	Database 2	Probability	Download cost, $d_m^{[2]}$
1	$W_2(1)$	$\phi$	$p_1^{[2]}$	1
2	$\phi$	$W_2(1)$	$p_2^{[2]}$	1
3	$W_1(1)$	$W_1(1) + W_2(1)$	$p_3^{[2]}$	2
4	$W_1(1) + W_2(1)$	$W_1(1)$	$p_4^{[2]}$	2

Table 4.3: The dissimilarity of probability distribution of queries according to the desired message index seen at database 1 for  $N = 2, K = 2$

Query	Probability	
	$P_{Q_1^{[1]}}(q)$	$P_{Q_1^{[2]}}(q)$
$q$		
$\phi$	$p_2^{[1]}$	$p_2^{[2]}$
$W_1(1)$	$p_1^{[1]}$	$p_3^{[2]}$
$W_2(1)$	$p_3^{[1]}$	$p_1^{[2]}$
$W_1(1) + W_2(1)$	$p_4^{[1]}$	$p_4^{[2]}$

the JS divergence given as

$$\begin{aligned}
 \rho_{JS} & \tag{4.10} \\
 &= D_{JS}(P_{Q_1^{[1]}}, P_{Q_1^{[2]}}) \\
 &= \frac{1}{2} D_{KL} \left( P_{Q_1^{[1]}} \left\| \frac{P_{Q_1^{[1]}} + P_{Q_1^{[2]}}}{2} \right\| \right) + \frac{1}{2} D_{KL} \left( P_{Q_1^{[2]}} \left\| \frac{P_{Q_1^{[1]}} + P_{Q_1^{[2]}}}{2} \right\| \right) \\
 &= \frac{1}{2} \left( p_2^{[1]} \log_2 \frac{p_2^{[1]}}{\frac{p_2^{[1]} + p_2^{[2]}}{2}} + p_1^{[1]} \log_2 \frac{p_1^{[1]}}{\frac{p_1^{[1]} + p_3^{[2]}}{2}} + p_3^{[1]} \log_2 \frac{p_3^{[1]}}{\frac{p_3^{[1]} + p_1^{[2]}}{2}} + p_4^{[1]} \log_2 \frac{p_4^{[1]}}{\frac{p_4^{[1]} + p_4^{[2]}}{2}} \right) \\
 &+ \frac{1}{2} \left( p_2^{[2]} \log_2 \frac{p_2^{[2]}}{\frac{p_2^{[1]} + p_2^{[2]}}{2}} + p_3^{[2]} \log_2 \frac{p_3^{[2]}}{\frac{p_1^{[1]} + p_3^{[2]}}{2}} + p_1^{[2]} \log_2 \frac{p_1^{[2]}}{\frac{p_3^{[1]} + p_1^{[2]}}{2}} + p_4^{[2]} \log_2 \frac{p_4^{[2]}}{\frac{p_4^{[1]} + p_4^{[2]}}{2}} \right).
 \end{aligned}$$

Note that if and only if  $P_{Q_1^{[1]}} = P_{Q_1^{[2]}}$ , then the JS divergence  $\rho_{JS}$  equals zero, which means no information leakage. Assume the extreme case with uniform distribution,  $P_{Q_1^{[1]}}(q) = P_{Q_1^{[2]}}(q) = 1/4, q \in \mathcal{Q}$ , and then it is easy to see that  $\rho_{JS} = 0$  from (4.10). The expected download cost in this case is  $3/2$  from (4.8) and (4.9), which is the reciprocal of the PIR capacity without information leakage. Another solution to the



extreme case achieving the download cost 1 to retrieve  $W_1$  or  $W_2$  is given as

$$\begin{aligned}(p_1^{[1]}, p_2^{[1]}, p_3^{[1]}, p_4^{[1]}) &= \left(\frac{1}{2}, \frac{1}{2}, 0, 0\right), \\(p_1^{[2]}, p_2^{[2]}, p_3^{[2]}, p_4^{[2]}) &= \left(\frac{1}{2}, \frac{1}{2}, 0, 0\right),\end{aligned}$$

respectively, which is the case using only the queries with download cost of 1. The JS divergence for this solution is calculated as  $\rho_{JS} = 0.5$  from (4.10). Lastly, an example of intermediate solution with the download cost between the capacity and 1 is given as

$$\begin{aligned}(p_1^{[1]}, p_2^{[1]}, p_3^{[1]}, p_4^{[1]}) &= \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right), \\(p_1^{[2]}, p_2^{[2]}, p_3^{[2]}, p_4^{[2]}) &= \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right),\end{aligned}$$

respectively, which can achieve the download cost  $4/3$ .  $\rho_{JS} \approx 0.0409$  is obtained for this case.

The goal of this chapter is to find the optimal tradeoff between the information leakage  $\rho_{JS}$  and the download cost  $D$ . We want to minimize the information leakage (4.10) satisfying the download cost as in (4.8) and (4.9). We have three download cost-information leakage pairs for now, and they are achievable in the regime of information leakage measured by the JS divergence. We want to find the optimal pairs and connect them into the optimal tradeoff curve. Then the optimization problem we have in this example is given as

$$\begin{aligned}\text{minimize} \quad & \rho_{JS} = D_{JS}(P_{Q_1^{[1]}}, P_{Q_1^{[2]}}) \\ \text{subject to} \quad & \sum_{m=1}^2 p_m^{[k]} + \sum_{m=3}^4 2p_m^{[k]} = D, \quad k = 1, 2 \\ & \sum_{m=1}^4 p_m^{[k]} = 1, \quad k = 1, 2,\end{aligned} \tag{4.11}$$

where  $1 \leq D \leq 1/C = 1 + 1/2 = 3/2$ .

### 4.3.3 Example of Symmetric TSC Scheme with $N = 3, K = 2$

In this subsection, one more example with small  $N$  and  $K$  is presented without much detail. A table similar to Table 4.3 is given to explore the dissimilarity between distributions. We use query structures to retrieve  $W_1$  and  $W_2$  as presented in Tables 3.2 (a) and 3.2 (b) from Subsection 3.3.3 but with desired message index added in the superscript. The probabilistic queries of symmetric TSC scheme are in Tables 4.4 (a) and 4.4 (b). Then the dissimilarity of probability distribution according to the desired message index seen at the database 1 can be recasted in Table 4.5.

Then the optimization problem we want to solve is given as

$$\begin{aligned}
 & \text{minimize} \quad \rho_{JS} = D_{JS}(P_{Q_1^{[1]}}, P_{Q_1^{[2]}}) \\
 & \text{subject to} \quad \frac{1}{2} \left( \sum_{m=1}^3 2p_m^{[k]} + \sum_{m=4}^9 3p_m^{[k]} \right) = D, \quad k = 1, 2 \\
 & \quad \quad \quad \sum_{m=1}^9 p_m^{[k]} = 1, \quad k = 1, 2,
 \end{aligned} \tag{4.12}$$

where  $1 \leq D \leq 1/C = 1 + 1/3 = 4/3$ . The problem (4.12) will be solved in the next section.

## 4.4 Optimal Tradeoff Between Information Leakage and Download Cost under the Jensen-Shannon Divergence

### 4.4.1 Optimization Problem with General $N, K$

In this subsection, we will formulate the optimization problems (4.11) and (4.12) with general  $N$  databases and  $K$  messages. From the general PIR query structure in Table 3.4 with general  $N$  and  $K$ , the optimization problem to minimize the JS divergence seen at the database  $n$  with  $K$  probability distributions  $P_{Q_n^{[1]}}, \dots, P_{Q_n^{[K]}}$  as optimiza-

Table 4.4: The probabilistic query structure of PIR scheme with  $N = 3, K = 2$

(a) To retrieve  $W_1$

Option	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_1(1)$	$W_1(2)$	$\phi$	$p_1^{[1]}$	2
2	$\phi$	$W_1(1)$	$W_1(2)$	$p_2^{[1]}$	2
3	$W_1(2)$	$\phi$	$W_1(1)$	$p_3^{[1]}$	2
4	$W_2(1)$	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$p_4^{[1]}$	3
5	$W_1(2) + W_2(1)$	$W_2(1)$	$W_1(1) + W_2(1)$	$p_5^{[1]}$	3
6	$W_1(1) + W_2(1)$	$W_1(2) + W_2(1)$	$W_2(1)$	$p_6^{[1]}$	3
7	$W_2(2)$	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$p_7^{[1]}$	3
8	$W_1(2) + W_2(2)$	$W_2(2)$	$W_1(1) + W_2(2)$	$p_8^{[1]}$	3
9	$W_1(1) + W_2(2)$	$W_1(2) + W_2(2)$	$W_2(2)$	$p_9^{[1]}$	3

(b) To retrieve  $W_2$

Option	Database 1	Database 2	Database 3	Prob.	Cost
1	$W_2(1)$	$W_2(2)$	$\phi$	$p_1^{[2]}$	2
2	$\phi$	$W_2(1)$	$W_2(2)$	$p_2^{[2]}$	2
3	$W_2(2)$	$\phi$	$W_2(1)$	$p_3^{[2]}$	2
4	$W_1(1)$	$W_1(1) + W_2(1)$	$W_1(1) + W_2(2)$	$p_4^{[2]}$	3
5	$W_1(1) + W_2(2)$	$W_1(1)$	$W_1(1) + W_2(1)$	$p_5^{[2]}$	3
6	$W_1(1) + W_2(1)$	$W_1(1) + W_2(2)$	$W_1(1)$	$p_6^{[2]}$	3
7	$W_1(2)$	$W_1(2) + W_2(1)$	$W_1(2) + W_2(2)$	$p_7^{[2]}$	3
8	$W_1(2) + W_2(2)$	$W_1(2)$	$W_1(2) + W_2(1)$	$p_8^{[2]}$	3
9	$W_1(2) + W_2(1)$	$W_1(2) + W_2(2)$	$W_1(2)$	$p_9^{[2]}$	3

Table 4.5: The dissimilarity of probability distribution of queries according to the desired message index seen at database 1 for  $N = 3, K = 2$

Query	Probability	
	$P_{Q_1^{[1]}}(q)$	$P_{Q_1^{[2]}}(q)$
$\phi$	$p_2^{[1]}$	$p_2^{[2]}$
$W_1(1)$	$p_1^{[1]}$	$p_4^{[2]}$
$W_1(2)$	$p_3^{[1]}$	$p_7^{[2]}$
$W_2(1)$	$p_4^{[1]}$	$p_1^{[2]}$
$W_2(2)$	$p_7^{[1]}$	$p_3^{[2]}$
$W_1(1) + W_2(1)$	$p_6^{[1]}$	$p_6^{[2]}$
$W_1(1) + W_2(2)$	$p_9^{[1]}$	$p_5^{[2]}$
$W_1(2) + W_2(1)$	$p_5^{[1]}$	$p_9^{[2]}$
$W_1(2) + W_2(2)$	$p_8^{[1]}$	$p_8^{[2]}$

tion variables is formulated as follows:

$$\begin{aligned}
& \text{minimize} && \rho_{JS} = D_{JS}(P_{Q_n^{[1]}}, \dots, P_{Q_n^{[K]}}) \\
& \text{subject to} && \frac{1}{N-1} \left( \sum_{m=1}^N (N-1)p_m^{[k]} + \sum_{m=N+1}^{N^K} Np_m^{[k]} \right) = D, \quad k \in [1 : K] \\
& && \sum_{m=1}^N p_m^{[k]} = 1, \quad k \in [1 : K].
\end{aligned} \tag{4.13}$$

Note that there are  $N^K$  query options. Among them,  $N$  query options have the download cost of  $(N-1)$ , and the rest of  $N^K - N$  options have the download cost of  $N$ . Again, queries are symmetric across databases and all possible queries appear in each database with probability. Therefore considering only one arbitrary database  $n \in [1 : N]$  is enough to solve the optimization problem.

**Remark 6.** *Note that the JS divergence suits well to compare more than two multiple probability distributions. At a glance,  $\rho_{JS} = D_{JS}(P_{Q_n^{[1]}}, \dots, P_{Q_n^{[K]}})$  captures how the  $K$  distributions vary depending on what the desired message is. Furthermore, there is no need to introducing the reference probability distribution  $U$  indicating no information leakage in the PIR system as in Chapter 3. The amount of information leakage is decided only from the user's choice of retrieval probability.*

An analytic or explicit solution to the optimization problem (4.13) is hard to obtain because of the complexity. Each of the  $K$  probability distributions  $P_{Q_n^{[k]}}$ ,  $k \in [1 : K]$  have  $N^K$  optimization variables. Therefore there are  $KN^K$  optimization variables in (4.13). However, fortunately, the JS divergence is known to be convex [29] in its domain. By intuition, the JS divergence is a weighted sum of KL divergences which is convex in its domain. Therefore with the affine constraints in (4.13), there exists a global optimum solution.

#### 4.4.2 Numerical Analysis with Examples

In this subsection, we present the numerical solution to two examples previously considered,  $N = 2, K = 2$  case and  $N = 3, K = 2$  case. Curves representing the trade-offs between information leakage and the download cost are shown in Figures 4.1 and 4.2. For both cases, the download costs shown in the figures are normalized with their desired message size  $L$ . We can find two extreme points located at the bottom right corner and the top left corner of each graph. The bottom right corner corresponds to the case where there is no information leakage. The download cost for this point is the reciprocal of PIR capacity (2.1),

$$\frac{1}{C} = 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}.$$

In the two examples,  $1/C$  are computed as

$$\begin{aligned} 1 + \frac{1}{2} &= \frac{3}{2} = 1.5, \\ 1 + \frac{1}{3} &= \frac{4}{3} \approx 1.3333, \end{aligned}$$

respectively, which agree with the graphical results. This point can be achieved by using uniform distribution to  $P_{Q_n^{[k]}}$ ,  $k \in [1 : K]$  for all  $N^K$  options.

The top left corner corresponds to the case where only direct downloading is used without keeping any privacy. The download cost for this point is 1 and information leakage reaches its maximum. This point can be achieved by using uniform distribution to  $P_{Q_n^{[k]}}$ ,  $k \in [1 : K]$  for the first  $N$  options, that is,

$$p_m^{[k]} = \begin{cases} \frac{1}{N}, & m = 1, \dots, N \\ 0, & \text{otherwise.} \end{cases}$$

One can think of only using the first option among the  $N$  direct downloading options.

In the example of  $N = 2, K = 2$ , this will lead to

$$\begin{aligned} (p_1^{[1]}, p_2^{[1]}, p_3^{[1]}, p_4^{[1]}) &= (1, 0, 0, 0), \\ (p_1^{[2]}, p_2^{[2]}, p_3^{[2]}, p_4^{[2]}) &= (1, 0, 0, 0). \end{aligned}$$

However, the JS divergence derived from these probability distributions is somewhat different. Specifically, in this case,  $\rho_{JS}$  seen at the database 1 becomes 1 and  $\rho_{JS}$  seen at the database 2 becomes 0. Therefore, we can bias information leakage between the databases, which can be advantageous in some practical scenarios. Note that the average information leakage is still 0.5, which agrees with the graphical result.

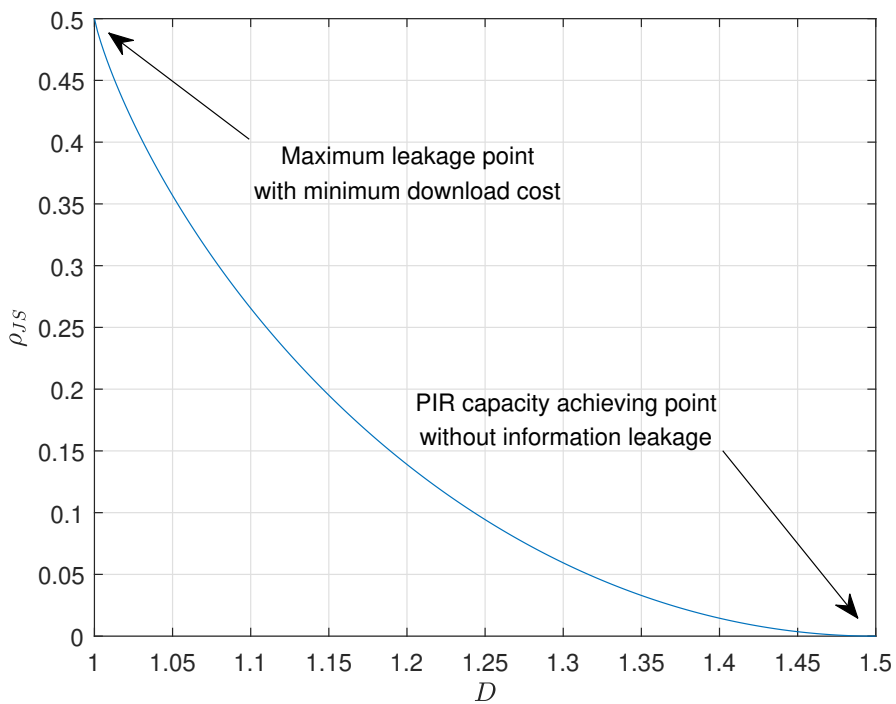


Figure 4.1: Optimal tradeoff between information leakage by the JS divergence and the normalized download cost of PIR scheme with  $N = 2$ ,  $K = 2$ .



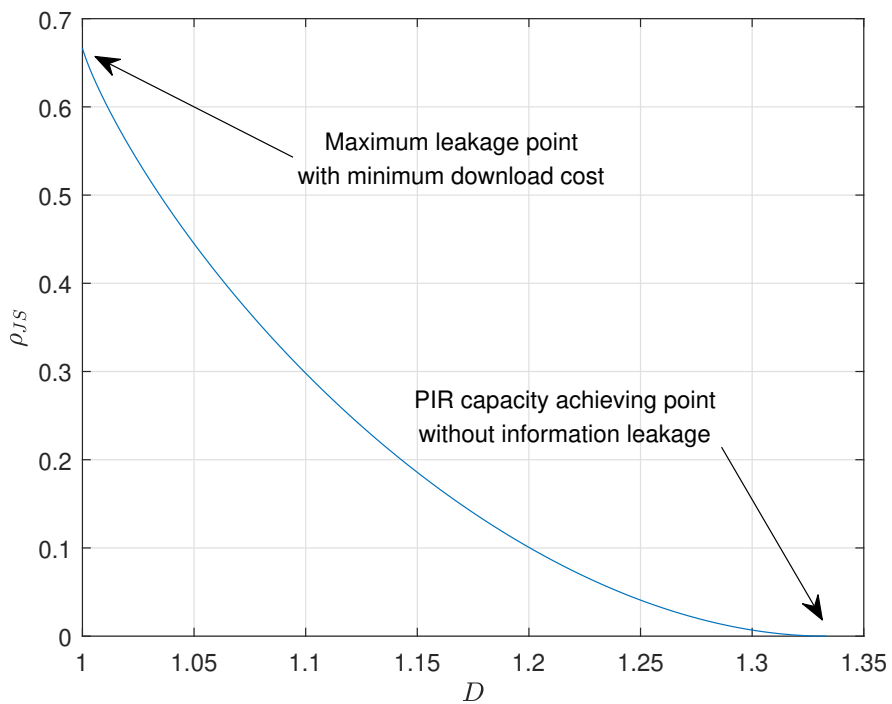


Figure 4.2: Optimal tradeoff between information leakage by the JS divergence and the normalized download cost of PIR scheme with  $N = 3$ ,  $K = 2$ .

## Chapter 5

### Conclusions

In this dissertation, research on the private information retrieval with information leakage was presented.

In Chapter 2, some preliminaries of PIR were briefly overviewed. Basic concepts of PIR and related researches were introduced, especially for the problem with information leakage. The convex optimization was also introduced.

In Chapter 3, the PIR problem with information leakage under the Kullback-Liebler divergence was proposed. The symmetric TSC PIR scheme with probabilistic query structure is adopted as the target of optimization. Given a probability distribution with no leakage as a reference, the KL divergence measures how much a probability distribution diverges from the perfect privacy. Information leakage establishes the tradeoff relationship with the performance measure of the PIR system, the download cost. The information leakage measured by the KL divergence is minimized with the entropy minimization problem. By using the given probabilistic PIR query structure, an analytic solution to the optimal tradeoff is found. We also considered an alternative PIR scheme that has different tradeoff curves. At some range with increased information leakage, the alternative PIR scheme shows a better tradeoff than that of the symmetric TSC scheme.

In Chapter 4, another PIR problem with information leakage under the Jensen-

Shannon divergence is proposed. For the same problem settings with probabilistic PIR query structure, the divergence between the probability distributions of queries that depend on the identity of the desired message was measured with the JS divergence. The JS divergence is advantageous since unlike other commonly used dissimilarity measures, it captures more than two distributions, which is desirable in our scenario of the PIR system. It was no longer necessary to have the predefined probability distribution indicating no leakage in the PIR system. The tradeoff between the information leakage taken by JS divergence and the download cost is solved with convex optimization formulation.

# Bibliography

- [1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, “Private information retrieval,” *J. ACM*, vol. 45, no. 6, pp. 965–981, 1998.
- [2] H. Sun and S. A. Jafar, “The capacity of private information retrieval,” *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.
- [3] N. B. Shah, K. V. Rashmi, and K. Ramchandran, “One extra bit of download ensures perfectly private information retrieval,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, 2014, pp. 856–860.
- [4] H. Sun and S. A. Jafar, “The capacity of robust private information retrieval with colluding databases,” *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.
- [5] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, “Multi-message private information retrieval with private side information,” *2018 IEEE Inf. Theory Workshop (ITW)*, pp. 1–5, 2018.
- [6] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, “Private information retrieval with side information,” *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.
- [7] Z. Chen, Z. Wang, and S. A. Jafar, “The capacity of T-private information retrieval with private side information,” *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4761–4773, Aug. 2020.

- [8] H. Sun and S. A. Jafar, “The capacity of symmetric private information retrieval,” *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.
- [9] Q. Wang, H. Sun, and M. Skoglund, “The capacity of private information retrieval with eavesdroppers,” *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.
- [10] H. Sun and S. A. Jafar, “The capacity of private computation,” *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3880–3897, Jun. 2019.
- [11] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson, “On the capacity of single-server multi-message private information retrieval with side information,” in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, 2018, pp. 180–187.
- [12] H. Sun and S. A. Jafar, “Multiround private information retrieval: capacity and storage overhead,” *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.
- [13] K. Banawan and S. Ulukus, “The capacity of private information retrieval from coded databases,” *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.
- [14] Y. Wei, K. Banawan, and S. Ulukus, “Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching,” *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3215–3232, May 2019.
- [15] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, “Private information retrieval from MDS coded data in distributed storage systems,” *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

- [16] K. Banawan and S. Ulukus, “The capacity of private information retrieval from byzantine and colluding databases,” *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.
- [17] H. Sun and S. A. Jafar, “Optimal download cost of private information retrieval for arbitrary message length,” *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 12, pp. 2920–2932, Dec. 2017.
- [18] H. Sun and S. A. Jafar, “The capacity of robust private information retrieval with colluding databases,” *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.
- [19] I. Samy, R. Tandon, and L. Lazos, “On the capacity of leaky private information retrieval,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, 2019, pp. 1262–1266.
- [20] H. Lin, S. Kumar, E. Rosnes, A. G. i. Amat, and E. Yaakobi, “Weakly-private information retrieval,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, 2019, pp. 1257–1261.
- [21] R. Zhou, T. Guo, and C. Tian, “Weakly private information retrieval under the maximal leakage metric,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, USA, 2020, pp. 1089–1094.
- [22] C. Dwork, “Differential privacy,” in *Proc. 33rd International Colloquium on Automata, Languages and Programming (ICALP)(2)*, 2006, pp. 1–12.
- [23] C. Tian, H. Sun, and J. Chen, “Capacity-achieving private information retrieval codes with optimal message size and upload cost,” *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7613–7627, Nov. 2019.
- [24] S. Vithana, K. Banawan, and S. Ulukus. (2020). “Semantic private information retrieval.” [Online]. Available: <https://arxiv.org/abs/2003.13667>

- [25] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.
- [26] E. T. Jaynes, “Information theory and statistical mechanics,” *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.
- [27] J. Lin, “Divergence measures based on the Shannon entropy,” *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, 1991.
- [28] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Wiley, 1959.
- [29] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968
- [30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

# 초 록

이 논문에서의 두 가지 주요 공헌은 다음과 같다.

- Kullback-Leibler 발산으로 측정된 정보 누출이 존재하는 개인 정보 검색 문제를 만들고 해결하였다.
- Jensen-Shannon 발산으로 측정된 정보 누출이 존재하는 개인 정보 검색 문제를 만들고 해결하였다.

첫째로, Kullback-Leibler 발산을 사용하여 정보 누출이 존재하는 개인 정보 검색 문제를 제안한다. 정보 누출량은 Kullback-Leibler 발산으로 측정된다. 발산이 가지는 의미는 개인 정보 검색 시스템에 누출이 없게 되는 기준이 되는 특정 분포로부터 사용자가 선택할 수 있는 임의 분포와의 차이를 측정한 것이다. 정보 누출은 개인 정보 검색 시스템의 성능인 다운로드 비용 측면에서 도움을 줄 수 있다. 가능한 한 효율적으로 정보 누출을 이용하는 방법을 찾고자 하였으며 정보 누출과 다운로드 비용 간의 최적의 균형 지점을 찾는 문제를 제시하였다. 이 문제는 컨벡스 최적화 문제로 만들어 해결하였다. 또한, 일부 트레이드 오프 구간에서 기존의 개인 정보 검색 방식보다 더 나은 성능을 보여주는 메시지 길이가 더 짧은 개인 정보 검색 방식을 제안하였다.

둘째로, Jensen-Shannon 발산을 사용하여 정보 누출이 존재하는 개인 정보 검색 문제를 제안한다. Jensen-Shannon 발산은 Kullback-Leibler 발산을 기반으로 하는 확률 분포 사이의 비유사성을 나타내는 값이다. 사용자가 원하는 메시지가 무엇이냐에 따라 사용자가 선택할 수 있는 확률 분포의 차이가 발생하고 그 확률 분포들



간의 발산을 측정한다. Jensen-Shannon 발산에는 몇 가지 적절한 특징이 있는데 그 중 하나는 3 개 이상의 확률 분포 간의 비유사성을 측정할 수 있다는 것이다. 이를 이용하여 Jensen-Shannon 발산으로 공식화 된 문제에는 개인 정보 검색 시스템에 누출이 없게 되는 기준이 되는 특정 분포가 필요하지 않다. Jensen-Shannon 발산으로 측정된 정보 누출과 다운로드 비용 간의 균형 지점은 컨벡스 최적화 문제로 만들 수 있으며, 시뮬레이션을 통한 솔루션이 제시되었다.

**주요어:** 컨벡스 최적화, 다운로드 비용, 정보 누출, 정보 이론, Jensen-Shannon(JS) 발산, Kullback–Leibler(KL) 발산, 개인 정보 검색.

**학번:** 2015-21002