Distributed Testing with Multi-terminal Data Compression

By

WENWEN ZHAO

B.E. (University of Science and Technology of China, Hefei, China) 2013

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Assoc. Lifeng Lai, Chair

_____

Prof. Shuguang Cui

_____

Prof. Bernard C. Levy

Committee in Charge

2018

Wenwen Zhao
September 2018
Electrical and Computer Engineering

# Abstract

Due to the growth of size and scale of the dataset, data are naturally collected or stored in multiple terminals in many scenarios. In these cases, each terminal may not have full information about all the variables involved in learning or inference problems and it may need to send compressed data due to the limitation of communication channel capacity. Different from the centralized case where all information is stored in one terminal, we have to utilize compressed data directly for statistical inference and learning in the distributed case.

In this dissertation, we investigate the distributed statistical inference problems using information theoretic tools. In particular, we consider problems consisting of $L$ encoders $\{\mathcal{X}_l\}_{l=1}^{L}$, and one decision maker $\mathcal{Y}$, in which terminal $\mathcal{X}_l$ has local data related to random variable $X_l$ only and terminal $\mathcal{Y}$ has data related to random variable $Y$ only. In these problems, terminals $\{\mathcal{X}_l\}_{l=1}^{L}$ send compressed data to terminal $\mathcal{Y}$. Based on the received data along with its local information, terminal $\mathcal{Y}$ makes a decision about the joint probability mass function (PMF) of $(X_1, \cdots, X_L, Y)$ from given hypotheses $H_0 : P_{X_1 \cdots X_L Y}$ and $H_1 : Q_{X_1 \cdots X_L Y}$. Using Neyman-Pearson criterion, our goal is to maximize the type 2 error exponent under the constraints on the type 1 error probability and communication rates.

Three increasingly sophisticated scenarios are considered in this dissertation: $1)$ Basic model with non-interactive communications; $2)$ The scenario with cascaded communications; and $3)$ The scenario with model uncertainties. We first characterize the fundamental limits of using compressed data in the non-interactive case. We provide a lower bound on the type 2 error exponent for the general PMF case and establish both a lower bound and an upper bound for the special case of testing against independence. Moreover, we study the optimal inference performance with a diminishing communication rate (*zero rate*) and show

that for certain distributions, the performance is as good as that in the centralized case.

We then extend our study to a more complicated case with cascaded communication among terminals $\{\mathcal{X}_l\}_{l=1}^{L}$. To be specific, we assume that these terminals broadcast their messages in a sequential order from terminal $\mathcal{X}_1$ to terminal $\mathcal{X}_L$, and each terminal uses all previously received messages along with its own observations for encoding. We investigate both the case with a general PMF and a special case of testing against independence. We show that for certain PMFs, cascaded communication will help in improving the inference performance. However, we also prove that cascaded communication does not help in the case with zero-rate data compression.

Finally, we consider the case with uncertainties in the model. In this case, instead of knowing exact forms of the PMF in both hypotheses, we only have partial information about them. One of these problems is called distributed identity testing and it can be transformed into two composite hypothesis testing problems. We focus on the more complex one and establish bounds on the type 2 error exponent for both the case with a general PMF and the special case of testing against independence under constraints on the type 1 error probability and communication rates. Furthermore, we provide a matching upper and lower bound on the type 2 error exponent for the zero-rate data compression case.

# Acknowledgement

I would like to thank and acknowledge all those who supported me, helped me and encouraged me in the amazing experience of pursuing my Ph.D degree.

First and foremost, I would like to show my greatest gratitude to my advisor, Dr. Lifeng Lai. Ever since the first day I joined his group, Dr. Lai believed in me and gave me endless support. On the academic level, Dr. Lai taught me the fundamentals of conducting the scientific research in the field of machine learning and information theory. Under his supervision, I learned how to define a research problem, find a solution to it, publish the results and present our results to general and academic audience. On a personal level, I am inspired by Dr. Lai's hardworking and passionate attitude toward research. Moreover, I am very grateful for his patience and encouragement, which gives me time to grow. Without him, I won't become the kind of researcher I am today.

Besides my advisor, I would like to thank the rest of my dissertation committee members, Prof. Bernard Levy and Prof. Shuguang "Robert" Cui for their great support and invaluable advice. I am thankful to Prof. Levy for his remarkable comments in the qualifying exam, which helped shape my final dissertation. I am also very thankful to Prof. Cui for his insightful suggestions in the area of submodular maximization. I also would like to say thanks to Prof. Khaled Abdel-Ghaffar for the time he spent in the final exit seminar and say thanks to Prof. Zhi Ding and Prof. Weiyu Xu for serving on my Ph.D qualifying exam committee. It is both my honor and my pleasure to share my research work with them all.

I also would like to thank my labmates for their continuous support and accompany. I would like to thank Dr. Jun Geng, who shared with me lots of experience in conducting scientific research and finding an academic position. I am very thankful for the opportunity he offered to visit Harbin Institute of Technology and to discuss the trends in the field of machine learning with excellent peers. I am also very thankful to Dr. Bingwen Zhang and Dr. Ain-ul-Aisha, who helped me a lot in the early stage of my Ph.D life.

Last but not least, I would like to thank my family: my parents, my brother and my husband. It is my family who support me, always believe in me and give me endless encouragement. Especially, I would like to thank my husband, Wenwen Tu. The days that we spent together in living, studying, arguing, and working hard for our common future, I will cherish forever.

The work involved in this dissertation was supported in part by the National Science Foundation under Grant DMS-1265663, the Qatar National Research Fund under Grant QNRF-6-1326-2-532, and the National Science Foundation under Grants CNS-1660128 and CCF-1717943.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Introduction to Distributed Hypothesis Testing

Nowadays, there is an explosive growth in the size and scale of modern datasets. On the one hand, a large amount of data bring in opportunities such as that they inspire and facilitate the development of many deep learning/inference algorithms. On the other hand, they also bring in significant challenges as inference algorithms on large dataset are computationally demanding, while data are often redundant. Hence, how to efficiently store, transmit and utilize this large amount of data is under active investigation [22, 24, 33, 35, 38, 50, 51, 61]. One way to overcome the problem of storage and computation complexity is that instead of collecting all data in a centralized location, we leave the data at multiple terminals and infer useful information from these distributed data using the computation power offered by these distributed machines.

There are two basic ways to distribute the dataset: *sample partition* and *feature partition*, which originate in different scenarios in real life and result in totally different approaches to transmit and infer useful information from the distributed data.

### 1.1.1 Sample Partition and Feature Partition

As shown in Figure 1.1, in *sample partition*, each terminal has data samples related to all random variables [35, 50]. In this figure, we use a matrix to represent the available data, and the columns denote samples related to the corresponding random variables. The data matrix is partitioned in a row-wise manner and each terminal observes a subset of the samples, which relates to all random variables $(X_1, X_2, \cdots, X_L)$. This scenario is quite common in real life. For example, there are large quantities of voice and image data stored in personal smart devices but due to the sensitive nature of the data, we cannot ask all users to send their voice or photos to a centralized location. Hence, the data are distributed in multiple locations, and we need to adopt certain learning methods on each device if we want to obtain a certain learning result like speech recognition. Generally, in this scenario, even though each terminal has fewer data than that of the *centralized setting*, in which all data are stored in one terminal, each terminal can still apply learning methods to its local data. Certainly, communicating and combining learning results from distributed terminals may improve the performance.

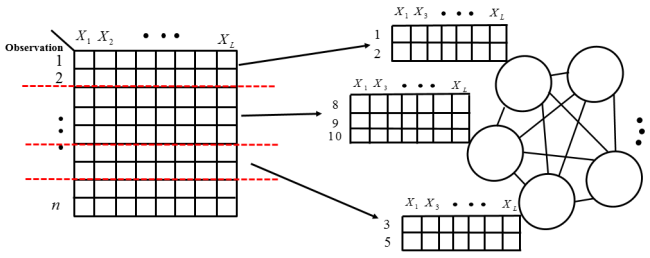



Figure 1.1: Sample partition

In *feature partition*, the data stored in each terminal only relate to a subset but not all of the random variables. Figure 1.2 illustrates the feature partition scenario, in which the data matrix is partitioned in a column-wise manner and each terminal observes the data related to a subset of random variables. For example, terminal $\mathcal{X}_1$ has all observations related to

random variable $X_1$. This scenario is also quite common in practice. For example, the physical information of patients is typically stored in different locations as patients may go to different departments or different hospitals for different physical tests. In general, this scenario is more challenging than the sample partition as each terminal in the feature partition scenario is not able to obtain meaningful information from local data alone. Moreover, due to the limited communication budget, recovering data first and then conducting inference from the recovered data is not optimal or necessary. We need to design approaches that can infer useful information directly from compressed data, which is a much more complicated and challenging problem compared with the problems in the sample partition scenario.

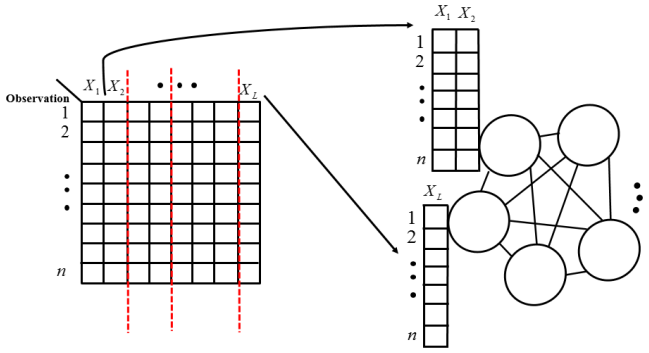



Figure 1.2: Feature partition

The main difference between the sample partition and the feature partition is that in the sample partition scenario, local data at each terminal is related to all random variables, while in the feature partition scenario, local data is only related to a subset of random variables. This difference results in the following effects. The first one is that in the sample partition scenario, all existing learning and inference methods can still be applied directly to the local data, while in the feature partition scenario we need to check each method carefully to see whether we can apply it to the compressed messages and attain reasonable performance. The second one is that in the sample partition scenario, each terminal is able to communicate and exchange local learning results, which typically requires much less communication resources than exchanging the local raw data. Therefore, in the sample partition scenario, researchers

focus on developing optimization approaches to combine the results computed locally for various learning purposes [22–24, 35, 50]. However, in the feature partition scenario, as terminals cannot perform learning methods on local data alone, they have to exchange the raw data (or compressed version of the raw data), and hence the requirement on communication resources is much more demanding. Moreover, research works in the feature partition scenario are less complete due to the complexity of problems.

### 1.1.2 Distributed Inference in the Feature Partition Scenario

Now we take a closer look at the problem of how to infer information from the distributed data in the feature partition scenario. As the general inference problem is quite complex, we first study a basic inference problem to gain some insights, which is illustrated in Figure 1.3. In this simple network, terminal $\mathcal{X}$ observes a sequence $X^n$ and terminal $\mathcal{Y}$ observes a

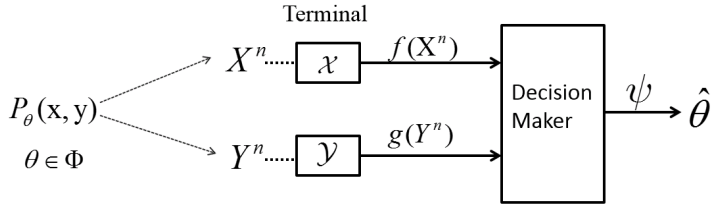



Figure 1.3: A canonical example for distributed inference problems.

sequence $Y^n$. $(X^n, Y^n)$ are generated independently and identically (i.i.d.) according to a certain given joint PMF

$$\{P_{XY,\theta}(x, y)\}_{\theta\in\Phi} \tag{1.1}$$

with an unknown parameter $\theta \in \Phi$ ($\Phi$ is an appropriate prescribed set). At first, terminal $\mathcal{X}$ sends a function of its observations, $f(X^n)$ and terminal $\mathcal{Y}$ sends a function of its observation $g(Y^n)$ to the decision maker. Then the decision maker infers the value of $\theta$ based on the received message $f(X^n)$ and $g(Y^n)$ using a decoding function $\psi$. Due to the limited communication budget, there are rate constraints on the transmission link: $\frac{1}{n}\log||f|| \leq R_1$ and $\frac{1}{n}\log||g|| \leq R_2$, where $||\cdot||$ denotes the cardinality of possible values of the function.

The goal is to design encoding functions $f$, $g$ and a decoding function $\psi$ such that $\hat{\theta}$ is a good estimate of $\theta$ under certain performance metric.

The problem of distributed inference was proposed by Berger [5], and it combines interesting techniques of statistics and information theory. Many existing works on the classic distributed inference problem focus on the following three branches: distributed hypothesis testing or distributed detection, distributed pattern classification and distributed estimation [1, 13, 15, 16, 20, 25, 31, 34, 39, 56, 57]. In *distributed hypothesis testing* or *distributed detection* problem, $\Phi$ consists of only two elements, that is, $\Phi = \{H_0, H_1\}$, where $H_0$ and $H_1$ are called *null hypothesis* and *alternative hypothesis* respectively. When $\Phi$ consists of a finite number $m$ ($m \geq 2$) of pattern classes, this class of problems is called *distributed pattern classification*. When $\Phi$ is an open set, we need to design a statistical inference system such that the intended norm of the covariance of $\hat{\theta}$ is as small as possible. This class of problems is called *distribution parameter estimation* or *distributed estimation* problems. Using powerful information theoretic tools, good upper and lower bounds on the inference performance are derived for the basic model and many special cases. Moreover, some results are extended to more general cases. However, due to the formidable complexity in this problem, it rarely allows us to reach the *single-letter characterization* for all achievable error exponents. This means that this research field is not yet mature enough and it remains to be further developed.

### 1.1.3 Connections and Differences with Distributed Source Coding Problems

The distributed inference problem is different from the classic distributed source coding problems [11]. In the source coding problem, which is illustrated in Figure 1.4, the goal of the decoder is to recover the source sequences after it receives the compressed messages from terminal $\mathcal{X}$ and $\mathcal{Y}$. According to Slepian-Wolf theorem [11], the decoder can recover the original sequences with diminishing error probability when the encoding rates are larger

than certain values. Hence, when the encoding rate constraints are loose enough, we can adopt the source coding method to get the original sequences in our distributed inference problems. However, in general, the rate constraints are typically too strict for the decision maker to fully recover $X^n$ in the inference problem. Moreover, in the inference problem, recovery of source sequences is not its goal and it typically is not necessary. Hence, we will use the compressed messages directly in the distributed inference problems, which requires different methods and is more complex.

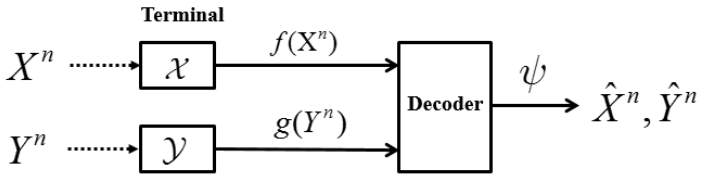



Figure 1.4: A canonical example for source coding problem.

On the other hand, this inference problem is closely connected to distributed source coding problems. In particular, the general idea of the existing schemes in distributed inference problems is to mimic the schemes used in distributed source coding problems. In the existing studies [1, 3, 13, 15, 16], each terminal $\mathcal{X}_l$ compresses its sequence $X_l^n$ into $U_l^n$. Then these terminals send the auxiliary sequences $\{U_l^n\}_{l=1}^L$ to the decision maker using source coding ideas so that the decision maker can obtain $\{\hat{U}_l^n\}_{l=1}^L$, which has a high probability to be the same as $\{U_l^n\}_{l=1}^L$. The compression step is to make sure each terminal sends enough information needed for recover $U_l^n$ but does not exceed the rate constraint. Finally, the decision maker will estimate $\theta$ using $\{\hat{U}_l^n\}_{l=1}^L$. Hence, we can see that even though the decision maker does not need to recover the sequences $\{X_l^n\}_{l=1}^L$, it does need to recover $\{U_l^n\}_{l=1}^L$ from the compresses messages.

### 1.1.4 Differences with Distributed Detection Problems using Scalar Quantizer

There have been a large number of existing works on distributed detection problems, see [6–8, 18, 28–30, 32, 40, 42, 44–46, 49, 60] and references therein. Most of the existing works consider the *scalar quantizer*, in which the quantizer at terminal $\mathcal{X}_l$ quantizes each component of $X_l^n$ one by one. This setup fits certain sensor network applications, as the complexity of the scalar quantizer is low and it incurs minimal decision delay. Under this scalar quantization setup, it is typically assumed that the observations at different terminals are conditionally (conditioned on the hypothesis) independent. The problem becomes very challenging once the assumption of conditional independence is relaxed [40, 46]. Some recent interesting works have made an important progress for the case with correlated observations [8, 42, 49].

In this dissertation, we focus on distributed detection problems with block encoding, in other words, *vector quantizer*, in which the observations are processed in blocks. The use of vector quantizer allows us to borrow powerful tools from information theory to distributed detection. As we will show in the dissertation, these tools enable us to make progress in understanding the general problems without the conditional independence assumptions.

### 1.1.5 Distributed Hypothesis Testing

In this dissertation, we study a class of distributed hypothesis testing problems from information theoretic perspective. In particular, the two hypotheses $H_0$ and $H_1$ are:

$$H_0 : P_{XY} \qquad \text{vs} \qquad H_1 : Q_{XY}. \tag{1.2}$$

The decision maker should decide between the two hypotheses after receiving the encoded massages from terminals $\mathcal{X}$ and $\mathcal{Y}$. Typically, the decision maker defines a region as

$$\mathcal{A}_n = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \psi(f(x^n)g(y^n)) = H_0\}, \tag{1.3}$$

and it decides $H_0$ is true if the $(x^n, y^n) \in \mathcal{A}_n$; otherwise it decides $H_1$ is true. $\mathcal{A}_n$ is called the *acceptance region*. Two types of error probability are considered for the decision maker's behavior:

$$\alpha_n = P_{XY}^n(\mathcal{A}_n^c), \tag{1.4}$$

$$\beta_n = Q_{XY}^n(\mathcal{A}_n), \tag{1.5}$$

where $\alpha_n$ means the probability that the decision maker decides $H_1$ to be true while the truth is the $H_0$ and $\beta_n$ means vice verse. In this dissertation, $\alpha_n$ is called *type 1 error probability* and $\beta_n$ is called *type 2 error probability*. The goal is to make the type 2 error probability as small as possible while the type 1 error probability is constrained. To better understand the performance of the type 2 error probability, we define the error exponent of type 2 error probability to be $\liminf_{n\to\infty} \left(-\frac{1}{n}\log\left(\beta_n\right)\right)$, which is the asymptotic behaviour of the type 2 error probability. We would like to maximize the type 2 error exponent over all possible encoding functions $f$ and $g$ and the decoding function $\psi$.

[13] establishes a good lower bound the type 2 error exponent in 1987, while providing a matching upper bound is still an open problem. Later, for a special case of *zero-rate data compression*, which means the compression rate should decay with a certain rate, [15,16,34] establish a matching upper and lower bound on the type 2 error exponent under different kinds of constraints on the type 1 error probability.

Another special case, which is illustrated in Figure 1.5, is studied in [3]. In this case, one can view terminal $\mathcal{Y}$ as the decision maker and $Y^n$ as any side information available at this terminal. Furthermore, the decision maker only cares about whether $X$ and $Y$ are independent or not, which is called *testing against independence*. [3] provides a matching upper and lower bound in this case by connecting the distributed hypothesis testing problem with the distributed source coding with one helper problem. However, when there is more than one terminal $\mathcal{X}$ that sends encoded messages to terminal $\mathcal{Y}$, the problem is still open.

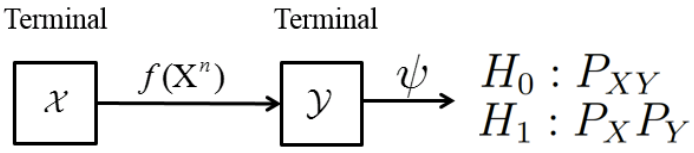



Figure 1.5: A special case of distributed hypothesis testing.

[21,47,48] extends the testing against independence case in Figure 1.5 by allowing *interactive communication* between terminal $\mathcal{X}$ and $\mathcal{Y}$. Here, *interactive communication* means that there are multiple rounds of communication between terminal $\mathcal{X}$ and $\mathcal{Y}$. When there is only one round of communication, e.g., the case shown in Figure 1.5, the case is called *non-interactive communication*. The distributed interactive hypothesis testing is shown in Figure 1.6. Intuitively, after multiple rounds of communication, the decision maker will get more information than that in non-interactive case and hence a better result is expected. [21] proves that this intuitive idea is true by providing a single-letter characterization of the type 2 error exponent. When there are more than two terminals, the case becomes extremely complex and thus it is still open.

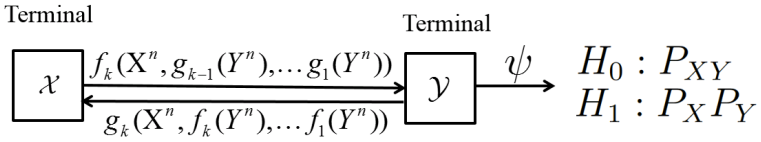



Figure 1.6: Distributed interactive hypothesis testing.

In this dissertation, we focus on answering the following aspects to further explore the distributed hypothesis testing problem:

- When there are multiple terminals with non-interactive encoders, i.e. there is more than one terminal $\mathcal{X}$ sending encoded messages to the decision maker $\mathcal{Y}$, this dissertation analyzes the fundamental limits of compressing data, evaluates the performance and compares the performance with that in centralized case.

- Then this dissertation extends the study to a scenario where cascaded communication is allowed among multiple terminals. Comparisons are made with the non-interactive

scenario case by case.

- Finally, if there are uncertainties in the model, this dissertation designs efficient universal coding schemes and evaluates the performance.

## 1.2 Distributed Hypothesis Testing with Non-Interactive Encoders

We first explore a model with non-interactive encoders to obtain the fundamental limits of compressing data in hypothesis testing problems. In this model, we consider a setup with $L$ terminals (encoders) $\{\mathcal{X}_l\}_{l=1}^{L}$, and a decision making terminal $\mathcal{Y}$, where terminal $\mathcal{X}_l$ has observations related to random variable $X_l$ and terminal $\mathcal{Y}$ has data only related to random variable $Y$. Terminals $\{\mathcal{X}_l\}_{l=1}^{L}$ can send information related to their own data with limited rates to the decision maker $\mathcal{Y}$. Based on the messages received from these terminals and its own data, the decision maker $\mathcal{Y}$ tries to determine the joint PMF from the following two given hypotheses:

$$H_0 : P_{X_1 \cdots X_L Y} \quad \text{vs} \quad H_1 : Q_{X_1 \cdots X_L Y}. \tag{1.6}$$

This process is as shown in Figure 1.7. As the communication rates between the terminals and the decision maker are limited, terminal $\mathcal{X}_l$ has to compress its observations $X_l^n$ with $n$ being the number of samples available at terminal $\mathcal{X}_l$. Recall that the goal of distributed hypothesis testing is to design encoding and decision functions under various resource (e.g., communication cost) and performance (e.g., error probabilities) constraints. This model is related to but different from the basic model introduced in Section 1.1.2. In the basic model, the random variables $(X_1, \cdots, X_L)$ are all at one terminal $\mathcal{X}$ and there is no side information available at the decision maker. We can view the side information $Y^n$ as the fully recovered sequence by the decision maker when terminal $\mathcal{Y}$ in the basic model compresses

its information with a very large rate.

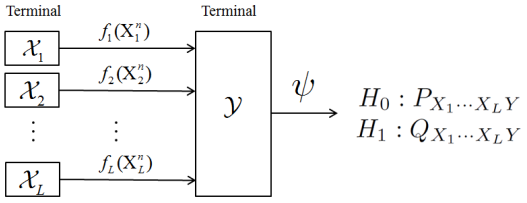



Figure 1.7: Model with non-interactive encoders

As discussed in Section 1.1.5, the basic model is well-studied in [3, 13, 15, 16, 34]. However, the increase of the number of terminals brings in more new challenges, which will be discussed in details in Chapter 2.

## 1.3 Distributed Hypothesis Testing with Cascaded Encoders

Building on the fundamental limits gained in non-interactive communication case, we move on to the interactive communication case. As discussed in Section 1.1.5, in the interactive case, multiple rounds of communication are allowed among all terminals. However, the problem becomes very complex if an arbitrary form of interaction among encoders are allowed, especially when there are more than two terminals. In this dissertation, we study a special form of interaction among terminals: cascaded communication, which is illustrated in Figure 1.8. In particular, we assume that terminals broadcast their messages in a sequential

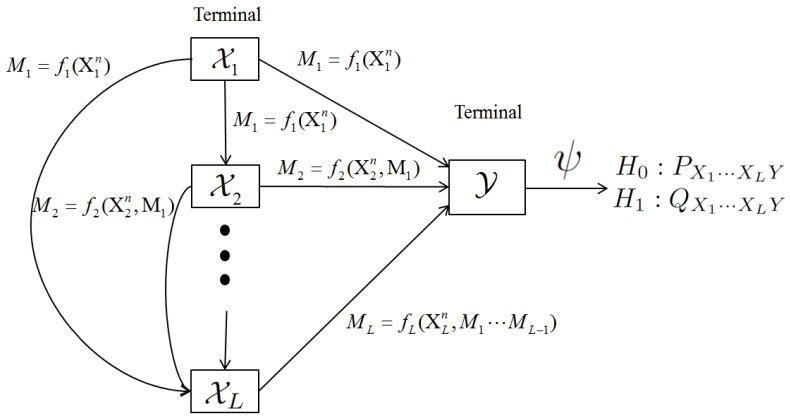



Figure 1.8: Model with interactive encoders

order from terminal $\mathcal{X}_1$ to terminal $\mathcal{X}_L$, and each terminal uses all messages received so far along with its own observations for encoding. More specifically, terminal $\mathcal{X}_1$ first broadcasts the encoded message based on its own observations $X_1^n$, and then terminal $\mathcal{X}_2$ broadcasts its encoded message based on both its own observations $X_2^n$ and the message received from terminal $\mathcal{X}_1$. This process continues until terminal $\mathcal{X}_L$ broadcasts its message based on its own observations $X_L^n$ and all previously received messages. Finally, terminal $\mathcal{Y}$ performs a statistical inference based on the messages received from all terminals $\{\mathcal{X}_l\}_{l=1}^L$ and its own data related to $Y$. In this dissertation, we focus on the same inference problem as in the non-interactive case, in which terminal $\mathcal{Y}$ tries to decide the joint PMF of the data from the two hypotheses in (1.6).

Our goal is to maximize the type 2 error exponent under constraints on the type 1 error probability and the communication rates. Note that one may consider other forms of interaction among encoders. However, the problem becomes very complicated if an arbitrary form of interaction among encoders are allowed, and these cases are left for future study.

The problem studied in this dissertation is related to but different from several existing interesting works on inference with interactive communication [4, 48]. In particular, in [48], the authors discussed a case in which $\{X_l\}_{l=1}^L$ are all at terminal $\mathcal{X}$ (and hence $(X_1, \cdots, X_L)$ can be denoted as one random variable $X$) and terminal $\mathcal{X}$ and terminal $\mathcal{Y}$ can communicate with each other in multiple rounds. [4] considers the same setup with [48] but uses sample-by-sample processing, i.e. scaler quantization at each stage. Different from these interesting studies, in our problem, we consider a case in which $\{\mathcal{X}_l\}_{l=1}^L$ are at different terminals. Furthermore, cascaded communication among the encoders is allowed.

## 1.4 Distributed Identity Testing

In Section 1.2 and Section 1.3, we discussed the cases with basic hypotheses as shown in (1.6). However, in certain scenarios, we do not have complete information about underlying

distributions. One of these problems is identity testing problem.

We study the identity testing problem in the feature partition scenario with non-interactive communication of the encoders. Similar to Section 1.2, we consider a setup with $L$ terminals (encoders) $\{\mathcal{X}_l\}_{l=1}^L$ and a decision making terminal $\mathcal{Y}$. $(X_1^n, \cdots, X_L^n, Y^n)$ are generated according to some unknown PMF $P_{X_1 \cdots X_L Y}$. Terminals $\{\mathcal{X}_l\}_{l=1}^L$ can send compressed messages related to their own data with limited rates to the decision maker, then the decision maker performs statistical inference based on the messages received from terminals $\{\mathcal{X}_l\}_{l=1}^L$ and its local data related to $Y$. In particular, we focus on the problem that the decision maker tries to decide whether $P_{X_1 \cdots X_L Y}$ is the same as a given distribution $Q_{X_1 \cdots X_L Y}$, i.e. $P_{X_1 \cdots X_L Y} = Q_{X_1 \cdots X_L Y}$ or they are $\lambda$-far away, i.e., $||P_{X_1 \cdots X_L Y} - Q_{X_1 \cdots X_L Y}||_1 \geq \lambda \, (\lambda > 0)$, and $|| \cdot ||_1$ denotes the $\ell_1$ norm of its argument. This problem can be interpreted as two hypothesis testing problems:

- *Problem 1*:

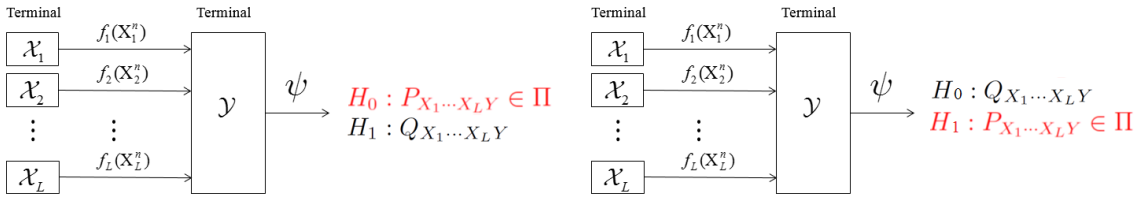
$$H_0 : ||P_{X_1 \cdots X_L Y} - Q_{X_1 \cdots X_L Y}||_1 \geq \lambda \quad \text{vs} \quad H_1 : P_{X_1 \cdots X_L Y} = Q_{X_1 \cdots X_L Y}. \tag{1.7}$$

- *Problem 2*:

$$H_0 : P_{X_1 \cdots X_L Y} = Q_{X_1 \cdots X_L Y} \quad \text{vs} \quad H_1 : ||P_{X_1 \cdots X_L Y} - Q_{X_1 \cdots X_L Y}||_1 \geq \lambda. \tag{1.8}$$

These problems are as shown in Figure 1.11. In both problems, our goal is to characterize the type 2 error exponents under the constraints on the communication rates and type 1 error probabilities.

Both problems are well-studied in the centralized case [12, 41, 43, 62], however, the understanding in the distributed setting is limited. There are several existing interesting works [1, 13, 15, 16, 34] that are related to our work. In particular, [1, 13, 15, 16, 34] discuss a distributed testing problem with simple hypotheses as shown in (1.6). Our distributed iden-




Problem 1 Problem 2

Figure 1.11: Model for Problem 1 and Problem 2 with $\Pi := \{P_{X_1\cdots X_L Y} : ||P_{X_1\cdots X_L Y} - Q_{X_1\cdots X_L Y}||_1 \geq \lambda\}$.

tity testing problem with composite hypotheses, can be viewed as a generalization of those problems considered in [1, 13, 15, 16]. Between the two possible problems defined in (1.7) and (1.8), *Problem 2* is relatively simple and it can be solved using similar schemes proposed in [1, 13, 15, 16]. In particular, the encoding schemes and the definition of the acceptance regions at the decision maker in [1,13,15,16] depend only on the form of PMF under $H_0$. Since the form of PMF under $H_0$ in *Problem 2* is known, we can apply the existing coding/decoding schemes as that in [1, 13, 15, 16] and take the type 2 error probability as the supreme of the type 2 error probabilities under each $P_{X_1\cdots X_L Y}$ that satisfies $||P_{X_1\cdots X_L Y} - Q_{X_1\cdots X_L Y}||_1 \geq \lambda$. Furthermore, it can be shown that these schemes are optimal for *Problem 2*. However, in *Problem 1*, as $H_0$ is composite, we need to design universal encoding/decoding schemes so that our schemes can provide performance guarantee regardless of what the true PMF under $H_0$ is. In this dissertation, we will only focus on the more challenging *Problem 1*.

## 1.5 Summary of Contribution and Organization

Here, we summarize the main contributions of this dissertation.

- First, we study three cases in the distributed hypothesis testing with multiple non-interactive encoders. In the first case, the testing against independence problem is studied. Instead of connecting it to the open distributed source coding with multiple helpers problem, we propose encoding/decoding schemes to provide a lower bound on

the type 2 error exponent. Furthermore, we establish an upper bound on the type 2 error exponent using any scheme that satisfies the communication rate and type 1 error probability constraints. Then we study the zero-rate data compression case. We fully characterize the best achievable type 2 error exponent under a special exponential-type type 1 error probability constraint. Finally, we extend the work to the general hypothesis testing case, and compare the result with that in the centralized cases. The results are published in [54, 56, 57].

- Second, we propose a novel distributed testing with cascaded encoders setting. We first study the special case of testing against independence case. We provide a scheme to give a lower bound on the type 2 error exponent, and then show that this scheme is optimal by properly choosing auxiliary random variables and rigorously justifying that they satisfy various Markov chain conditions. We then generalize the results to general cases and compare them with cases with non-interactive encoders. We provide examples to show that when the decision maker receives more information using cascaded encoder, it outperforms that using non-interactive encoders. However, we also find that there is no performance gain in zero-rate data compression case. These results are published in [53, 55].

- Finally, the novel distributed identity testing model is proposed in this dissertation. A special case under zero-rate compression is studied first. As $H_0$ is composite in our case, the decoding scheme in [16,56], which depends on the knowledge of distribution in $H_0$, is not applicable anymore. By devising a new universal decoding scheme and providing a matching upper bound, we fully characterize the type 2 error exponent under the zero-rate compression and the exponential-type constraint on the type 1 error probability. Then the results are generalized to different cases in the distributed identity testing model: testing against independence and general hypothesis testing under different kinds of constraints on the type 1 error probability. Part of this work has been

published in [58] and a journal summarizing these results was submitted to the *IEEE Transactions on Information Theory* in Aug. 2017 [52].

The reminder of the dissertation is organized as follows. In Chapter 2 , we introduce the distributed hypothesis testing problem with non-interactive encoders. Then, we study the distributed hypothesis testing with cascaded encoders in Chapter 3. In Chapter 4, we introduce the distributed identity testing model. Finally, we offer concluding remarks and provide certain potential directions for the future work in Chapter 5.

# Chapter 2

# Distributed Hypothesis Testing with Non-Interactive Encoders

## 2.1 Introduction

In this chapter, we study the fundamental performance limits of problem in the feature partition scenario with non-interactive encoders.

We first focus on the zero-rate compression case in which each terminal is only allowed to send messages to the decision maker with zero-rate compression. If the decision maker were required to fully recover the data of terminals $\{\mathcal{X}_l\}_{l=1}^{L}$ as in the distributed source coding problems [11, 37], this zero-rate compression is not enough. However, in our setup, this zero-rate compression will still be valuable for the decision maker for statistical inference. In addition, we impose an exponential-type constraint on the type 1 error probability (i.e., the type 1 error probability is required to decrease exponentially fast with a certain error exponent). We fully characterize the best achievable error exponent of the type 2 error probability under these zero-rate compression and exponential-type type 1 error probability constraints by providing matching lower and upper bounds. A clear benefit of this zero-rate compression approach is that the terminals only need to consume a limited amount of communication

resources. In addition, we show that a very simple scheme in which each terminal only sends the empirical distribution (or an approximation of it) is optimal. This implies that the complexity of the optimal scheme employed by sensors in practical detection problem can be very low. Furthermore, we provide an example in which the performance of the scheme with zero-rate compression is very close to that of the centralized case.

We then extend the study to the positive rate constraints case. Compared with the zero-rate compression case, in this scenario, each terminal can convey more information to the decision maker. As the general problem is very complicated, we focus on the special case of testing against independence. The case with $(X_1, \cdots, X_L)$ all at one terminal (with $Y$ being at another terminal) was first considered by Ahlswede and Csiszár [1]. In [1], the problem was converted to a source coding with a helper problem. However, this approach may not work for our case, as the corresponding problem will be a source coding with multiple helpers problem, which is still open. In this chapter, we use a different approach to exploit the more flexible rate constraints and characterize the corresponding type 2 error exponents. Furthermore, we provide an upper-bound on the best achievable type 2 error exponent using any scheme that satisfies the communication rate and type 1 error exponents.

The remainder of the chapter is organized as follows. In Section 2.2, we introduce the model studied in this chapter. In Section 2.4, we present out results for the zero-rate compression case. In Section 2.5, we focus on the scenario with positive rate constraints. In Section 2.6, we use several numerical examples to illustrate analytical results obtained in this chapter. Finally, we offer some concluding remarks in Section 2.7.

## 2.2 Model

Consider a system with $L$ terminals: $\mathcal{X}_l$, $l = 1, \cdots, L$ and a decision maker $\mathcal{Y}$. Each terminal and the decision maker observe a component of the random vector $(X_1, \cdots, X_L, Y)$ that take

values in a finite set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L \times \mathcal{Y}$ and admit a joint PMF with two possible forms:

$$H_0 : P_{X_1 \cdots X_L Y}, \qquad H_1 : Q_{X_1 \cdots X_L Y}. \tag{2.1}$$

With a slight abuse of notation, we use $\mathcal{X}_l$ to denote both the terminal and the alphabet set from which the random variable $X_l$ takes values. $(X_1^n, \cdots, X_L^n, Y^n)$ are independently and identically generated according to one of the above joint PMFs. In other words, $(X_1^n, \cdots, X_L^n, Y^n)$ is generated by either $P^n_{X_1 \cdots X_L Y}$ or $Q^n_{X_1 \cdots X_L Y}$. In a typical hypothesis testing problem, one determines which hypothesis is true under the assumption that $(X_1^n, \cdots, X_L^n, Y^n)$ are fully available at the decision maker. In this chapter, we consider a distributed setting in which $X_l^n$, $l = 1, \cdots, L$ and $Y^n$ are at different locations. In particular, terminal $\mathcal{X}_l$ observes only $X_l^n$ and terminal $\mathcal{Y}$ observes only $Y^n$. Terminals $\mathcal{X}_l$s are allowed to send messages to the decision maker $\mathcal{Y}$. Using $Y^n$ and the received messages, $\mathcal{Y}$ determines which hypothesis is true. We denote this system as $S_{X_1 \cdots X_L|Y}$. Figure 2.1 illustrates the system model. In the following, we will use the term "decision maker" and terminal $\mathcal{Y}$ interchangeably. Here, $Y^n$ is used to model any side information available at the decision maker. If $\mathcal{Y}$ is set to be an empty set, then the decision maker does not have side information.

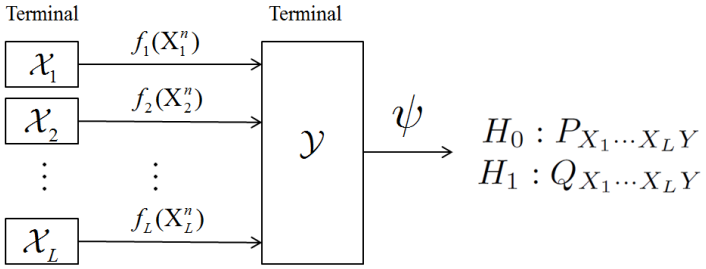



Figure 2.1: Model

After observing the data sequence $x_l^n \in \mathcal{X}_l^n$, terminal $\mathcal{X}_l$ will use a encoder $f_l$ to transform the sequence $x_l^n$ into a message $f_l(x_l^n)$, which takes values from the message set $\mathcal{M}_l$

$$f_l : \mathcal{X}_l^n \to \mathcal{M}_l = \{1, 2, \ldots, M_l\}, \tag{2.2}$$

with rate constraint:

$$\limsup_{n\to\infty} \frac{1}{n} \log M_l \leq R_l, \quad l = 1, \cdots, L. \tag{2.3}$$

We also use the notation $||f_l||$ to denote the cardinality of $f_l$, $l = 1, \cdots, L$. Hence, we have $||f_l|| = M_l$, $l = 1, \cdots, L$

Using messages $M_l$, $l = 1, \cdots, L$ and its side information $Y^n$, the decision maker will employ a decision function $\psi$ to determine which hypothesis is true:

$$\psi : \mathcal{M}_1 \times \cdots \times \mathcal{M}_L \times \mathcal{Y}^n \to \{H_0, H_1\}. \tag{2.4}$$

For any given encoders $f_l,\ l = 1, \cdots, L$ and decision function $\psi$, one can define the acceptance region as

$$\begin{aligned} \mathcal{A}_n = \big\{ &(x_1^n, \cdots, x_L^n, y^n) \in \mathcal{X}_1^n \times \cdots \times \mathcal{X}_L^n \times \mathcal{Y}^n : \\ &\psi(f_1(x_1^n) \cdots f_L(x_L^n) y^n) = H_0 \big\}. \end{aligned} \tag{2.5}$$

Correspondingly, the type 1 error probability is defined as

$$\alpha_n = P^n_{X_1 \cdots X_L Y}(\mathcal{A}_n^c), \tag{2.6}$$

in which $\mathcal{A}_n^c$ denotes the complement of $\mathcal{A}_n$, and the type 2 error probability is defined as

$$\beta_n = Q^n_{X_1 \cdots X_L Y}(\mathcal{A}_n). \tag{2.7}$$

Our goal is to design the quantization functions $f_l$, $l = 1, \cdots, L$ and the decision function $\psi$ to maximize the type 2 error exponent under certain type 1 error and communication rate constraints (2.3).

More specifically, we consider two kinds of type 1 error constraint, namely:

- Constant-type constraint

$$\alpha_n \leq \epsilon \tag{2.8}$$

for a prefixed $\epsilon > 0$, which implies that the type 1 error probability must be smaller than a given threshold; and

- Exponential-type constraint

$$\alpha_n \leq \exp(-nr) \tag{2.9}$$

for a given $r > 0$, which implies that the type 1 error probability must decrease exponentially fast with an exponent no less than $r$. Hence the exponential-type constraint is stricter than the constant-type constraint.

To distinguish these two different type 1 error constraints, we use different notations to denote the corresponding type 2 error exponent.

- Under the constant-type constraint, we define the type 2 error exponent as

$$\theta(R_1, \cdots, R_L, \epsilon) = \liminf_{n\to\infty} \left( -\frac{1}{n} \log \left( \min_{f_1,\cdots,f_L,\psi} \beta_n \right) \right),$$

in which the minimization is over all $f_l$s and $\psi$ satisfying condition (2.3) and (2.8).

- Under the exponential-type constraint, we define the type 2 error exponent as

$$\sigma(R_1, \cdots, R_L, r) = \liminf_{n\to\infty} \left( -\frac{1}{n} \log \left( \min_{f_1,\cdots,f_L,\psi} \beta_n \right) \right),$$

in which the minimization is over all $f_l$s and $\psi$ satisfying condition (2.3) and (2.9).

## 2.3 Preliminary

Following [11], for any sequence $x^n = (x(1), \cdots, x(n)) \in \mathcal{X}^n$, we use $n(a|x^n)$ to denote the total number of indices $t$ at which $x(t) = a$. Then, the relative frequencies or empirical PMF- $\pi(a|x^n) \triangleq n(a|x^n)/n, \forall a \in \mathcal{X}$ of the components of $x^n$, is called the type of $x^n$ and is denoted by $tp(x^n)$. The set of all types of sequences in $\mathcal{X}^n$ is denoted by $\mathcal{P}^n(\mathcal{X})$. Furthermore, we call a random variable $X^{(n)}$ that has the same distribution as $tp(x^n)$ as the type variable of $x^n$.

For any given sequence $x^n$, we use typical sequence [11] and $r$-divergent sequence [16] to measure how likely this sequence is generated from a PMF $P_X$.

### 2.3.1 Typical Sequence

For a given a type $P_X \in \mathcal{P}^n(\mathcal{X})$ and a constant $\eta$, we denote by $T_\eta^n(X)$ the set of $(P_X, \eta)$-typical sequences in $\mathcal{X}^n$:

$$T_\eta^n(X) \triangleq \{x^n \in \mathcal{X}^n : |\pi(a|x^n) - P_X(a)| \leq \eta P_X(a), \forall a \in \mathcal{X}\}.$$

In the same manner, we use $\tilde{T}_\eta^n(X)$ to denote the set of $(\tilde{P}_X, \eta)$-typical sequences. Note that when $\eta = 0$, $T_0^n(X)$ denote the set of sequences $x^n \in \mathcal{X}^n$ of type $P_X$, and we use $T^n(X)$ for simplicity.

Furthermore, for $y^n \in \mathcal{Y}^n$, we define $T_\eta^n(X|y^n)$ as the set of all $x^n$s that are jointly typical with $y^n$:

$$T_\eta^n(X|y^n) = \{x^n \in \mathcal{X}^n : (x^n, y^n) \in T_\eta^n(XY)\}. \tag{2.10}$$

We use the following lemma to summarize key properties of typical sequences. More details can be found in [11].

**Lemma 2.1.** ([16]) Let $\lambda > 0$ be arbitrary.

(1) $P_X^n(T_\eta^n(X)) \geq 1 - \lambda$.

(2) Let $X^{(n)}$ be a type variable for a sequence in $\mathcal{X}^n$, then

$$(n+1)^{-|\mathcal{X}|}\exp[nH(X^{(n)})] \leq |T_0^n(X^{(n)})| \leq \exp[n(H(X^{(n)}))]. \tag{2.11}$$

(3) Let $x^n \in \mathcal{X}^n$ and $X$ be a random variable in $\mathcal{X}$, then

$$\Pr(X^n = x^n) = \exp\left[-n(H(X^{(n)}) + D(X^{(n)}||X))\right]. \tag{2.12}$$

### 2.3.2 $r$-divergent Sequence

The concept of $r$-divergent sequences also plays an important role in the following development. Here, we review the definition and some important properties of $r$-divergent sequences. More details and properties of $r$-divergent sequences can be found in [16].

**Definition 2.1.** ([16]) Let $X$ be a random variable taking values in a finite set $\mathcal{X}$ with PMF $P_X$, and $r \geq 0$. An $n$-sequence $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is called $r$-divergent sequence for $X$ if

$$D(X^{(n)}||X) \leq r, \tag{2.13}$$

where $X^{(n)}$ is the type variable of $x^n$ and $D(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence of the two random variables involved. The set of all $r$-divergent sequences is denoted by $S_r^n(X)$.

In particular, $S_0^n(X)$ (i.e., $r = 0$) represents the set of all $x^n$ sequences such that $tp(x^n) = P_X$, i.e. $S_0^n(X) = T_0^n(X)$. The following lemma from[16] summarizes key properties of $r$-divergent sequences.

**Lemma 2.2.** ([16]) Let $r > 0$ be fixed.

(1) $P_X^n(S_r^n(X)) \geq 1 - (n+1)^{|\mathcal{X}|}\exp(-nr)$.

(2) Let $X^{(n)}$ be a type variable for a sequence in $\mathcal{X}^n$, then

$$(n+1)^{-|\mathcal{X}|}\exp[nH(X^{(n)})] \leq |S_0^n(X^{(n)})| \leq \exp[n(H(X^{(n)}))]. \tag{2.14}$$

(3) Let $\mathcal{A}_n$ be a subset of $\mathcal{X}^n$ and

$$P_X^n(\mathcal{A}_n) \geq 1 - \exp(-nr) \tag{2.15}$$

holds. Let $\mathcal{A}_n(X^{(n)}) \triangleq \mathcal{A}_n \cap S_0^n(X^{(n)})$, we have

$$|\mathcal{A}_n(X^{(n)})| \geq \big(1 - (n+1)^{|\mathcal{X}|}\exp[-n(r-c_n)]\big)|S_0^n(X^{(n)})| \tag{2.16}$$

with $c_n = D(X^{(n)}||X)$.

## 2.4 Testing under Zero-rate Compression with Exponential-type Constraints

In this section, we focus on the "zero-rate" compression, i.e., $R_1 = \cdots = R_L = 0$ under the exponential-type constraint. More specifically, we assume

$$\text{as } n \to \infty,\ M_l \to \infty, \tag{2.17}$$

but

$$R_l = \frac{1}{n}\log M_l \downarrow 0, \quad l = 1, \cdots, L. \tag{2.18}$$

In this case, $\sigma(R_1, \cdots, R_L, r)$ will be denoted as $\sigma(0, \cdots, 0, r)$. This zero-rate compression is of practical interest, as the normalized (normalized by the length of the data) communication cost is minimal. It is well-known that in the traditional distributed source coding with side information problems [11, 37], whose goal is to recover $(X_1^n, \cdots, X_L^n)$ at terminal $\mathcal{Y}$,

this zero-rate information is not useful. However, in our setup, the goal is only to determine which hypothesis is true. This zero-rate information will be very useful.

The scenario with zero-rate compression under the constant-type constraint has been considered in [34]. We will discuss the scenario with zero-rate compression under the exponential-type constraint (2.9).

In the following subsections, we first review several concepts that are useful for our development. We then characterize the type 2 error exponent with $L = 2$ before extending the result to the general case.

### 2.4.1 The Case with $L = 2$

In this subsection, to assist the presentation, we first focus on the case with $L = 2$ and provide details on how to characterize $\sigma(0, 0, r)$. We will then discuss the general case in Section 2.4.2.

We first establish an upper bound on the error exponent that any scheme can achieve. We will follow the similar strategy as in [15]. In particular, we will first convert a problem with the exponential-type constraint to a corresponding problem with the constant-type constraint. We then obtain an upper bound on the error exponent using the results in [34] for the constant-type constraint.

**Theorem 2.1.** Let $P_{X_1X_2Y}$ be arbitrary and $Q_{X_1X_2Y} > 0$. For zero-rate compression in $S_{X_1X_2|Y}$ with $R_1 = R_2 = 0$, the error exponent satisfies

$$\sigma(0, 0, r) \leq \sigma_{opt}, \tag{2.19}$$

in which

$$\sigma_{opt} \triangleq \min_{\tilde{P}_{X_1X_2Y} \in \mathcal{H}_r} D\left(\tilde{P}_{X_1X_2Y} || Q_{X_1X_2Y}\right) \tag{2.20}$$

with

$$\mathcal{H}_r = \left\{ \tilde{P}_{X_1X_2Y} : \tilde{P}_{X_1} = \hat{P}_{X_1}, \tilde{P}_{X_2} = \hat{P}_{X_2}, \tilde{P}_Y = \hat{P}_Y \text{ for some } \hat{P}_{X_1X_2Y} \in \varphi_r \right\}, \quad (2.21)$$

$$\varphi_r = \left\{ \hat{P}_{X_1X_2Y} : D(\hat{P}_{X_1X_2Y} || P_{X_1X_2Y}) \leq r \right\}. \quad (2.22)$$

*Proof.* Please refer to Appendix A.1. □

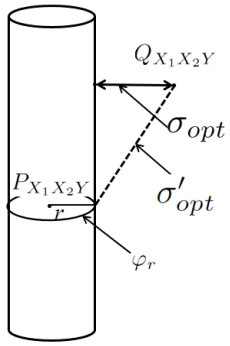



Figure 2.2: $\sigma_{opt}$ for zero-rate hypothesis testing

Figure 2.2 illustrates a geometric interpretation of $\sigma_{opt}$. In a centralized detection problem, $X_1^n,\ X_2^n$ and $Y^n$ are all available to the decision maker, so the decision maker knows the joint distribution of the observations. Setting the acceptance region as all observations whose empirical joint PMF having a KL-divergence to $P_{X_1X_2Y}$ less than or equal to $r$, expressed by $\varphi_r$ in Figure 2.2, then the best type 2 error exponent is the dashed line from $Q_{X_1X_2Y}$ to $\varphi_r$ in Figure 2.2, denoted as $\sigma'_{opt}$. In our distributed setting, different sequences are observed at different terminals and sent to the decision maker using zero-rate compression. Hence, the decision maker only gets the information about the marginal empirical PMF of the observations. Consequently, we should search over all joint distributions that have the same marginal distributions with the ones in $\varphi_r$, which is the region $\mathcal{H}_r$. Therefore, the best type 2 error exponent is the solid line from $Q_{X_1X_2Y}$ to $\mathcal{H}_r$.

Now, we present a scheme that can achieve the type 2 error exponent characterized in Theorem 2.1. Instead of showing that $\sigma(0,0,r) \geq \sigma_{opt}$ directly, we show that $\sigma_{opt}$ is achiev-

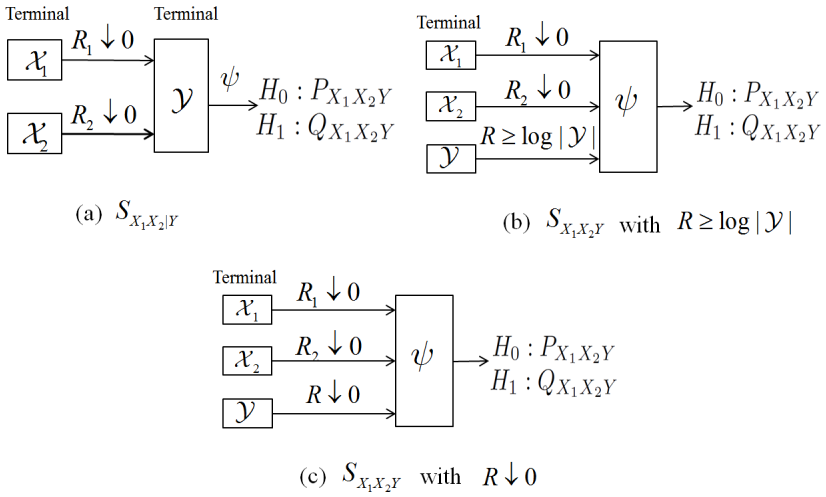



Figure 2.3: Model for achievability in zero-rate compression

able in a transformed model. The original model with $L = 2$ is shown in Figure 2.3 (a), denoted as $S_{X_1X_2|Y}$. In the original model, the decision maker is located in terminal $\mathcal{Y}$, so it has full access to $Y^n$. This model can also be viewed as a scenario in which the decision maker is located in a separate terminal and terminal $\mathcal{Y}$ also sends encoded messages to the decision maker, but its rate $R$ is so large ($R \geq \log|\mathcal{Y}|$) that the decision maker can fully recover $Y^n$. This new view is shown in Figure 2.3 (b). Therefore, the two system shown in Figure 2.3 (a) and (b) are equivalent, resulting in $\sigma(0,0,r) = \sigma(0,0,\log|\mathcal{Y}|,r)$. However, if the rate for terminal $\mathcal{Y}$ is not large enough, such as $R = 0$, which is shown in Figure 2.3 (c), then the decision maker cannot fully recover $Y^n$, thus it has less information than the decision maker in Figure 2.3 (b), and yields a larger error probability. Hence, we have $\sigma(0,0,r) = \sigma(0,0,\log|\mathcal{Y}|,r) \geq \sigma(0,0,0,r)$. We denote the system in Figure 2.3 (b) and (c) as $S_{X_1X_2Y}$. If we can show that $\sigma(0,0,0,r) \geq \sigma_{opt}$ in $S_{X_1X_2Y}$, then we have $\sigma(0,0,r) \geq \sigma_{opt}$ in $S_{X_1X_2|Y}$.

In the following, we will describe a scheme to show that $\sigma(0,0,0,r) \geq \sigma_{opt}$ in $S_{X_1X_2Y}$. Before proceeding to the formal proof, we first describe the high level idea of the scheme. After observing $x_i^n$, terminal $\mathcal{X}_i$ knows the type $tp(x_i^n)$ and sends $tp(x_i^n)$ (or an approximation of it, see below) to the decision maker. Terminal $\mathcal{Y}$ does the same. As there are at most $n^{|\mathcal{X}_i|}$ types [9], the rate required for sending the type from terminal $\mathcal{X}_i$ is $(|\mathcal{X}_i|\log n)/n$, which

goes to zero as $n$ increases. After receiving all type information from the terminals, the decision maker will check whether there is a joint type $\tilde{P}_{X_1X_2Y} \in \mathcal{H}_r$ such that its marginal types are the same as the information received from the terminals. If yes, the decision maker declares $H_0$ to be true, otherwise declares $H_1$ to be true. If the message size $M_i$ is less than $n^{|\mathcal{X}_i|}$, then instead of the exact type information $tp(x_i^n)$, each terminal will send an approximated version. Details on how to approximate the type will be provided in the proof. As long as $M_i \to \infty$, the approximation will be close (to be made precise in the proof) to the true type, and hence the decision maker can still use the above mentioned decision rule. We will show that this scheme can achieve $\sigma_{opt}$ in $S_{X_1X_2Y}$.

The following theorem provides details about the above mentioned idea.

**Theorem 2.2.** For zero-rate compression in $S_{X_1X_2Y}$ with $R_1 = R_2 = R = 0$, the error exponent satisfies

$$\sigma(0,0,0,r) \geq \sigma_{opt} \tag{2.23}$$

where $\sigma_{opt}$ is defined as (2.20).

*Proof.* First, define $g$-distance from any joint distribution to $P_{X_1X_2Y}$ as

$$g\left(\tilde{X}_1, \tilde{X}_2, \tilde{Y}\right) = \min_{\substack{\hat{P}_{X_1X_2Y} \\ \hat{P}_{X_1} = \tilde{P}_{X_1} \\ \hat{P}_{X_2} = \tilde{P}_{X_2} \\ \hat{P}_{Y} = \tilde{P}_{Y}}} D\left(\hat{P}_{X_1X_2Y} || P_{X_1X_2Y}\right) \tag{2.24}$$

which is continuous in $\left((\tilde{P}_{X_1})_{x_1 \in \mathcal{X}_1}, (\tilde{P}_{X_2})_{x_2 \in \mathcal{X}_2}, (\tilde{P}_Y)_{y \in \mathcal{Y}}\right)$.

Next, divide the $(|\mathcal{X}_1| + |\mathcal{X}_2| + |\mathcal{Y}|)$ dimensional unit cube into equal-sized $M_1 \times M_2 \times M$ small cells with each edge of length $\kappa_1$ along the first $|\mathcal{X}_1|$ components, each edge of length $\kappa_2$ along the $|\mathcal{X}_2|$ components and each edge of length $\tau$ along the $|\mathcal{Y}|$ components, where

$$\kappa_1 = M_1^{-1/|\mathcal{X}_1|}, \quad \kappa_2 = M_2^{-1/|\mathcal{X}_2|}, \quad \tau = M^{-1/|\mathcal{Y}|},$$

in which

$$M_1 \to \infty, M_2 \to \infty, M \to \infty, \tag{2.25}$$

but $\log M_i/n \to 0$ for $i = 1, 2$ and $\log M/n \to 0$, as $n \to \infty$ (i.e., zero-rate compression for all three terminals).

Choose and fix a representative point in each cell for every set of variables $(\tilde{X}_1, \tilde{X}_2, \tilde{Y})$. Then in a given cell, we make its representative variable set $(\check{X}_1, \check{X}_2, \check{Y})$ correspond in such a way that$((\check{P}_{X_1})_{x_1\in\mathcal{X}_1}, (\check{P}_{X_2})_{x_2\in\mathcal{X}_2}, (\check{P}_Y)_{y\in\mathcal{Y}})$ is the representative point of $((\tilde{P}_{X_1})_{x_1\in\mathcal{X}_1}$, $(\tilde{P}_{X_2})_{x_2\in\mathcal{X}_2}, (\tilde{P}_Y)_{y\in\mathcal{Y}})$. For each terminal, after observing its sequence, determines its type and then finds the index of the corresponding edge. Each terminal then sends the index to the decision maker. After receiving all the indexes, the decision maker can determine the cell index. Since we have assumed (2.25), we see that with any $\eta > 0$

$$|\tilde{P}_{X_1} - \check{P}_{X_1}| < \eta, \quad x_1 \in \mathcal{X}_1, \tag{2.26}$$

$$|\tilde{P}_{X_2} - \check{P}_{X_2}| < \eta, \quad x_2 \in \mathcal{X}_2, \tag{2.27}$$

$$|\tilde{P}_Y - \check{P}_Y| < \eta, \quad y \in \mathcal{Y}, \tag{2.28}$$

for sufficiently large $n \geq n_0(\eta)$. Furthermore, the continuity of $g(\tilde{X}_1, \tilde{X}_2, \tilde{Y})$ in $(\tilde{X}_1, \tilde{X}_2, \tilde{Y})$ yields

$$|g(\tilde{X}_1, \tilde{X}_2, \tilde{Y}) - g(\check{X}_1, \check{X}_2, \check{Y})| < \eta. \tag{2.29}$$

Denoting by $(\check{X}_1^{(n)}, \check{X}_2^{(n)}, \check{Y}^{(n)})$ the representative point of $(X_1^{(n)}, X_2^{(n)}, Y^{(n)})$ where $X_1^{(n)}$, $X_2^{(n)}$ and $Y^{(n)}$ are the type variables of $x_1^n \in \mathcal{X}_2^n$, $x_2^n \in \mathcal{X}_2^n$ and $y^n \in \mathcal{Y}^n$ respectively, we set an acceptance region

$$\mathcal{A}_n = \left\{(x_1^n, x_2^n, y^n) : g(\check{X}_1^{(n)}, \check{X}_2^{(n)}, \check{Y}^{(n)}) \leq r + 2\eta\right\}.$$

More precisely, our decoding scheme is as follows. Upon receiving $(M_1, M_2, M)$, find the representative point and its joint distribution. Then calculate the $g$-distance from this joint distribution to $P_{X_1X_2Y}$. If the $g$-distance is less than or equal to $r + 2\eta$, then we decide $H_0$ is true and vice versa. In other words, we first find the region of joint distributions which has a $g$-distance to $P_{X_1X_2Y}$ less than or equal to $r+2\eta$, which is visualized in Figure 2.4 as $\mathcal{H}_{r+2\eta}$. Then after knowing the joint distribution of the representative point, we can tell whether it is in $\mathcal{H}_{r+2\eta}$ or not. If it is in $\mathcal{H}_{r+2\eta}$, we decide $H_0$ is true and vice versa. In Figure 2.4, we use a square region to denote all possible joint distributions of the representative points.

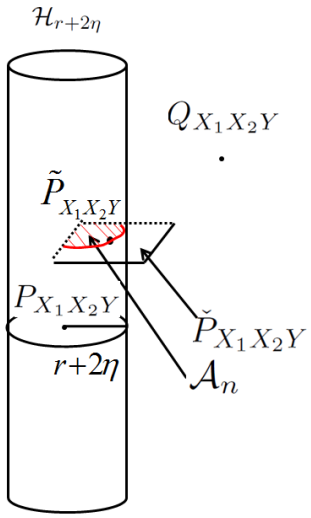



Figure 2.4: Visualization of acceptance region

Now we analyze the two types of error probability. For any $\rho > 0$ set

$$\xi_\rho = \left\{(x_1^n, x_2^n, y^n) : g(X_1^{(n)}, X_2^{(n)}, Y^{(n)}) \le \rho\right\};$$

then in view of (2.29) it is clear that

$$\xi_{r+\eta} \subset \mathcal{A}_n \subset \xi_{r+3\eta} \tag{2.30}$$

It is easy to see that $(x_1^n, x_2^n, y^n) \in \xi_{r+\eta}$ if $(x_1^n, x_2^n, y^n) \in S_{r+\eta}^n(X_1X_2Y)$, that is $S_{r+\eta}^n(X_1X_2Y)$

$\subset \xi_{r+\eta}$, which yields

$$1 - \alpha_n = P^n_{X_1X_2Y}(\mathcal{A}_n) \geq 1 - \exp(-nr)$$

for $n$ large enough. Hence, the constraint (2.9) is satisfied.

On the other hand, from the second inclusion in (2.30),

$$\begin{aligned}
\beta_n &= Q^n_{X_1X_2Y}(\mathcal{A}_n) \\
&\leq Q^n_{X_1X_2Y}(\xi_{r+3\eta}) \\
&\leq \sum_{\substack{X_1^{(n)}X_2^{(n)}Y^{(n)} \\ g(X_1^{(n)}, X_2^{(n)}, Y^{(n)}) \leq r+3\eta}} \exp\left(-nD\left(X_1^{(n)}X_2^{(n)}Y^{(n)}||Q_{X_1X_2Y}\right)\right) \\
&\leq (n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|} \max_{\substack{X_1^{(n)}X_2^{(n)}Y^{(n)} \\ g(X_1^{(n)}, X_2^{(n)}, Y^{(n)}) \leq r+3\eta}} \exp\left(-nD\left(X_1^{(n)}X_2^{(n)}Y^{(n)}||Q_{X_1X_2Y}\right)\right) \\
&\leq (n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|} \exp\left(-n\left(\min_{\substack{\tilde{X}_1\tilde{X}_2\tilde{Y} \\ g(\tilde{X}_1, \tilde{X}_2, \tilde{Y}) \leq r+3\eta}} D\left(\tilde{X}_1\tilde{X}_2\tilde{Y}||Q_{X_1X_2Y}\right)\right)\right).
\end{aligned}$$

Therefore,

$$\sigma(0,0,0,r) \geq \min_{\tilde{P}_{X_1X_2Y} \in \mathcal{H}_{r+3\eta}} D\left(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right),$$

which establishes (2.23) if we let $\eta \to 0$. □

As $\sigma(0,0,r) = \sigma(0,0,\log|\mathcal{Y}|,r)$. From Theorem 2.2, we have

$$\sigma(0,0,r) = \sigma(0,0,\log|\mathcal{Y}|,r) \geq \sigma(0,0,0,r) \geq \sigma_{opt}.$$

Coupled with Theorem 2.1, we have:

**Theorem 2.3.** Let $P_{X_1X_2Y}$ be arbitrary and $Q_{X_1X_2Y} > 0$. For zero-rate compression in

$S_{X_1X_2|Y}$ with $R_1 = R_2 = 0$ and type 1 error constraint (2.9), the best type 2 error exponent

$$\sigma(0, 0, r) = \sigma_{opt}. \tag{2.31}$$

where $\sigma_{opt}$ is defined as (2.20).

**Proposition 2.1.** Given $P_{X_1X_2Y}$ and $Q_{X_1X_2Y}$, the problem of finding $\sigma_{opt}$ defined in (2.20) is a convex optimization problem.

*Proof.* First, given $Q_{X_1X_2Y}$, it is easy to verify that the objective function $D(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y})$ in (2.20) is a convex function of $\tilde{P}_{X_1X_2Y}$.

Then, we show that the feasible set $\mathcal{H}_r$ defined in (2.21) is also convex. Suppose $\tilde{P}'_{X_1X_2Y} \in \mathcal{H}_r$ and $\tilde{P}''_{X_1X_2Y} \in \mathcal{H}_r$, and $\tilde{P}'_{X_1X_2Y}$ has the same marginal PMFs with $\hat{P}'_{X_1X_2Y} \in \varphi_r$, and $\tilde{P}''_{X_1X_2Y}$ has the same marginal PMFs $\hat{P}''_{X_1X_2Y} \in \varphi_r$. Setting

$$\tilde{P}'''_{X_1X_2Y} = \pi\tilde{P}'_{X_1X_2Y} + (1-\pi)\tilde{P}''_{X_1X_2Y},$$

for $0 \leq \pi \leq 1$, we will show that $\tilde{P}'''_{X_1X_2Y} \in \mathcal{H}_r$, i.e. $\mathcal{H}_r$ is a convex set. As we have

$$\tilde{P}'''_{X_1} = \pi\tilde{P}'_{X_1} + (1-\pi)\tilde{P}''_{X_1} = \pi\hat{P}'_{X_1} + (1-\pi)\hat{P}''_{X_1},$$

and similar results with $\tilde{P}'''_{X_2}$ and $\tilde{P}'''_{Y}$, we can conclude that $\tilde{P}'''_{X_1X_2Y}$ has the same marginal distribution as $\pi\hat{P}'_{X_1X_2Y}+(1-\pi)\hat{P}''_{X_1X_2Y}$. Due to the convexity of $D(\hat{P}_{X_1X_2Y}||P_{X_1X_2Y})$ with respect to $\hat{P}_{X_1X_2Y}$ for a given $P_{X_1X_2Y}$, we have $(\pi\hat{P}'_{X_1X_2Y} + (1-\pi)\hat{P}''_{X_1X_2Y}) \in \varphi_r$. This implies that $\tilde{P}'''_{X_1X_2Y} \in \mathcal{H}_r$, and hence $\mathcal{H}_r$ is a convex set.

As the result, for any given $P_{X_1X_2Y}$ and $Q_{X_1X_2Y}$, characterizing $\sigma_{opt}$ is a convex optimization problem and can be solved efficiently. □

### 2.4.2 General Case

The results of the previous section can be extended to the general case with $L$ terminals. We have the following theorem, whose proof follows the similar steps as in those in Section 2.4.1 and hence is omitted for conciseness.

**Theorem 2.4.** Let $P_{X_1,\cdots,X_LY}$ be arbitrary and $Q_{X_1,\cdots,X_LY} > 0$. For zero-rate compression in $S_{X_1\cdots X_L|Y}$ with $R_i = 0$, $i = 1,\cdots,L$ and type 1 error constraint (2.9), the best type 2 error exponent

$$\sigma(0,\cdots,0,r) = \min_{\tilde{P}_{X_1\cdots X_LY}\in\mathcal{H}_r} D\left(\tilde{P}_{X_1\cdots X_LY}||Q_{X_1\cdots X_LY}\right) \tag{2.32}$$

where

$$\mathcal{H}_r = \left\{\tilde{P}_{X_1\cdots X_LY} : \tilde{P}_{X_i} = \hat{P}_{X_i}, \tilde{P}_Y = \hat{P}_Y,\ i = 1,\cdots,L \text{ for some } \hat{P}_{X_1\cdots X_LY} \in \varphi_r\right\}, \tag{2.33}$$

$$\varphi_r = \left\{\hat{P}_{X_1\cdots X_LY} : D(\hat{P}_{X_1\cdots X_LY}||P_{X_1\cdots X_LY}) \le r\right\}. \tag{2.34}$$

Similar to (2.20), characterizing (2.32) is a convex optimization problem, hence it can be solved efficiently.

## 2.5 Testing against Independence with Constant-type Constraints

In this section, we consider the scenario with positive communication rate constraints, i.e., $R_l > 0$, $l = 1,\cdots,L$, under the constant-type constraint on the type 1 error probability. As the general case is a very complex problem even for $L = 1$ [15], we focus on the testing against independence case in which we are interested in determining whether $X_1,\cdots,X_L$

and $Y$ are independent or not. Hence, the two hypotheses are

$$H_0 : P_{X_1 \cdots X_L Y}, \qquad H_1 : Q_{X_1 \cdots X_L Y} = P_{X_1 \cdots X_L} P_Y.$$

Note that the marginal distribution of $(X_1, \cdots, X_L)$ and $Y$ are the same under both hypotheses in the case of testing against independence.

To facilitate the presentation, in the following, we only provide details for the $L = 2$ case. The results can be extended to the general $L$ case with proper modifications. For $L = 2$, our goal is to characterize $\theta(R_1, R_2, \epsilon)$ under $\alpha_n \leq \epsilon$ and communication constraints (2.3).

Compared with the zero-rate compression case discussed in Section 2.4, in this scenario, each terminal can convey more information to the decision maker as the communication rate constraint $R_l > 0$ is less strict. Before presenting the formal proof, we first describe high level ideas on how to exploit the more flexible rate constraints (terms in the following will be made precise in the proof). For a given rate constraint $R_l$, terminal $\mathcal{X}_l$ first generates a quantization codebook containing $2^{nR_l}$ quantization sequences. After observing $x_l^n$, terminal $\mathcal{X}_l$ picks one sequence $u_l^n$ from the quantization codebook to describe $x_l^n$ and sends this sequence to the decision maker. After receiving the descriptions from terminals, the decision maker will declare that the hypothesis $H_0$ is true if the descriptions from these terminals and the side-information at the decision maker are correlated. Otherwise, the decision maker will declare $H_1$. The following theorem provides details of the scheme and error probability analysis.

**Theorem 2.5.** In system $S_{X_1 X_2|Y}$ with $R_l > 0,\ l = 1, 2$, constraint on type 1 error probability (2.8) and communication constraints (2.3), the error exponent of the type 2 error probability is lower bounded by

$$\theta(R_1, R_2, \epsilon) \geq \max_{P_{U_1|X_1} P_{U_2|X_2}} I(U_1 U_2; Y), \tag{2.35}$$

in which the maximization is over $P_{U_l|X_i}$'s such that $I(U_l; X_l) \leq R_l$ and $|\mathcal{U}_l| \leq |\mathcal{X}_l| + 1$.

*Proof.* In the following, $\eta > \eta' > \eta'' > \eta'''$ are given small numbers.

**Codebook generation.** Fix a conditional PMF $P_{U_1U_2|X_1X_2Y} = P_{U_1|X_1}P_{U_2|X_2}$ that attains the maximum in (2.35). Let $P_{U_1}(u_1) = \sum_{x_1} P_{X_1}(x_1)P_{U_1|X_1}(u_1|x_1)$ and $P_{U_2}(u_2) = \sum_{x_2} P_{X_2}(x_2)P_{U_2|X_2}(u_2|x_2)$. Randomly and independently generate $\lfloor 2^{nR_1} \rfloor$ sequences $u_1^n(m_1)$, $m_1 \in \{1, \cdots, \lfloor 2^{nR_1} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_1}(u_{1i})$. Randomly and independently generate $\lfloor 2^{nR_2} \rfloor$ sequences $u_2^n(m_2)$, $m_2 \in \{1, \cdots, \lfloor 2^{nR_2} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_2}(u_{2i})$. These sequences constitute the codebook $c$, which is revealed to all terminals. We use $\mathcal{C}$ to denote the set of all possible codebooks.

**Encoding for terminal $\mathcal{X}_1$.** After observing sequence $x_1^n$, terminal $\mathcal{X}_1$ finds a $u_1^n(m_1)$ such that $(x_1^n, u_1^n(m_1)) \in T_{\eta'''}^n(U_1X_1)$, and sends the index $m_1$ to terminal $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, sends 0.

**Encoding for terminal $\mathcal{X}_2$.** Similarly, after observing a sequence $x_2^n$, terminal $\mathcal{X}_2$ finds a $u_2^n(m_2)$ such that $(x_2^n, u_2^n(m_2)) \in T_{\eta'''}^n(U_2X_2)$, then it sends the index $m_2$ to terminal $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends 0.

**Testing.** Upon receiving $m_1$ and $m_2$, terminal $\mathcal{Y}$ sets the acceptance region $\mathcal{A}_n$ for $H_0$ to

$$\mathcal{A}_n = \left\{(m_1, m_2, y^n) : (u_1^n(m_1), u_2^n(m_2), y^n) \in T_\eta^n(U_1U_2Y)\right\}. \tag{2.36}$$

This implies that terminal $\mathcal{Y}$ decides $\hat{H} = H_0$ if and only if no 0 is received and $(u_1^n(m_1), u_2^n(m_2), y^n) \in T_\eta^n(U_1U_2Y)$.

**Error probability analysis.** Terminal $\mathcal{Y}$ chooses $\hat{H} \neq H_0$ if and only if one or more of the following events occur:

$$\varepsilon_1 = \left\{(U_1^n(m_1), X_1^n) \notin T_{\eta'''}^n \text{ for all } m_1 \in [1 : \lfloor 2^{nR_1} \rfloor]\right\},$$

$$\varepsilon_2 = \left\{(U_2^n(m_2), X_2^n) \notin T_{\eta'''}^n \text{ for all } m_2 \in [1 : \lfloor 2^{nR_2} \rfloor]\right\},$$
$$\varepsilon_3 = \left\{(U_1^n(M_1), U_2^n(M_2), Y^n) \notin T_\eta^n(U_1U_2Y)\right\}.$$

Hence, $\mathcal{A}_n = (\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3)^c$.

For any particular codebook $c \in \mathcal{C}$, we use $\alpha_{nc}$ and $\beta_{nc}$ to denote the type 1 and the type 2 error probabilities respectively. In the following, we will first compute the probabilities of two types of errors averaged over all possible codebooks:

$$\mathbb{E}\{\alpha_{nc}\} = \sum_{c\in\mathcal{C}} \alpha_{nc}\text{Pr}(c), \tag{2.37}$$
$$\mathbb{E}\{\beta_{nc}\} = \sum_{c\in\mathcal{C}} \beta_{nc}\text{Pr}(c). \tag{2.38}$$

We will then argue that there exists a particular codebook $c^*$ that has the desired properties.

***a) Type 1 error probability:*** To analyze the type 1 error probability, we have

$$\begin{aligned}\mathbb{E}\{\alpha_{nc}\} &= P_{X_1X_2Y}^n(\mathcal{A}_n^c) = P_{X_1X_2Y}^n(\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3) \\ &\le P_{X_1X_2Y}^n(\varepsilon_1) + P_{X_1X_2Y}^n(\varepsilon_2) + P_{X_1X_2Y}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3).\end{aligned}$$

We now bound each term.

(1) By the covering lemma [11, Section 3.7], $P_{X_1X_2Y}^n(\varepsilon_1) \to 0$ as $n \to \infty$ if $R_1 \ge I(U_1; X_1) + \delta(\eta)$ and $P_{X_1X_2Y}^n(\varepsilon_2) \to 0$ as $n \to \infty$ if $R_2 \ge I(U_2; X_2) + \delta(\eta)$.

(2) To bound the last term, we need three steps, each of which uses a version of the Markov lemma [11, Section 12.1].

*Step 1*: Show that $(U_2^n(M_2), X_1^n, X_2^n) \in T_{\eta''}^n(U_2X_1X_2)$ with a probability tends to 1 as $n$ increases.

Since $X_2^n|\{U_2^n(M_2) = u_2^n, X_1^n = x_1^n\} \sim \prod_{i=1}^n P_{X_2|X_1}(x_{2i}|x_{1i})$ and $\eta'' > \eta'''$, by the Markov lemma, $\text{Pr}\{(U_2^n(M_2), X_1^n, X_2^n) \notin T_{\eta''}^n\}$ tends to zero as $n \to \infty$.

*Step 2*: Show that $(U_1^n(M_1), U_2^n(M_2), X_1^n, X_2^n) \in T_{\eta'}^n(U_1U_2X_1X_2)$ with a probability tends to 1 as $n$ increases.

From the distribution we draw $U_1^n(M_1)$ and $U_2^n(M_2)$, we have the Markov chain

$$U_2^n(M_2) \leftrightarrow X_2^n \leftrightarrow X_1^n \leftrightarrow U_1^n(M_1).$$

As $(u_2^n, x_1^n, x_2^n) \in T_{\eta''}^n(U_2X_1X_2)$ and from the Markov chain we know that

$$\Pr\{U_1^n(M_1) = u_1^n | U_2^n(M_2) = u_2^n, X_1^n = x_1^n, X_2^n = x_2^n\} = \Pr\{U_1^n(M_1) = u_1^n | x_1^n\}.$$

By the covering lemma, $\Pr\{(x_1^n, U_1^n) \in T_{\eta''}^n(U_1X_1)\}$ converges to 1 as $n \to \infty$, that is $\Pr\{U_1^n(M_1) = u_1^n | x_1^n\}$ satisfies the first condition in the Markov lemma. Then we show that it also satisfies the second condition in the Markov lemma.

For all $u_1^n \in T_{\eta''}^n(U_1|x_1^n)$,

$$\begin{aligned}
&\Pr\{U_1^n(M_1) = u_1^n | X_1^n = x_1^n\} \\
&= \Pr\{U_1^n(M_1) = u_1^n, U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n) | X_1^n = x_1^n\} \\
&= \Pr\{U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n) | X_1^n = x_1^n\} \\
&\quad \times \Pr\{U_1^n(M_1) = u_1^n | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n\} \\
&\le \Pr\{U_1^n(M_1) = u_1^n | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n\} \\
&= \sum_{m_1} \Pr\{U_1^n(M_1) = u_1^n, M_1 = m_1 | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n\} \\
&= \sum_{m_1} \Pr\{U_1^n(M_1) = u_1^n | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n, M_1 = m_1\} \\
&\quad \times \Pr\{M_1 = m_1 | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n\} \\
&\overset{(a)}{=} \sum_{m_1} \Pr\{U_1^n(m_1) = u_1^n | U_1^n(m_1) \in T_{\eta''}^n(U_1|x_1^n)\} \\
&\quad \times \Pr\{M_1 = m_1 | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n\} \\
&\overset{(b)}{\le} \sum_{m_1} \Pr\{M_1 = m_1 | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n\} 2^{-n(H(U_1|X_1) - \delta(\eta''))}
\end{aligned}$$

$$= 2^{-n(H(U_1|X_1)-\delta(\eta''))},$$

where (a) follows since

$$\begin{aligned}
&\Pr\{U_1^n(M_1) = u_1^n | U_1^n(M_1) \in T_{\eta''}^n(U_1|x_1^n), X_1^n = x_1^n, M_1 = m_1\} \\
&= \Pr\{U_1^n(m_1) = u_1^n | U_1^n(m_1) \in T_{\eta''}^n(U_1|X_1^n = x_1^n), X_1^n = x_1^n, M_1 = m_1\} \\
&= \Pr\{U_1^n(m_1) = u_1^n | U_1^n(m_1) \in T_{\eta''}^n(U_1|x_1^n)\}.
\end{aligned}$$

(b) follows from properties of typical sequences. Similarly, we can also prove that for every $u_1^n \in T_{\eta''}^n(U_1|x_1^n)$ and $n$ sufficiently large,

$$\Pr\{U_1^n(M_1) = u_1^n | X_1^n = x_1^n\} \geq (1-\eta'')2^{-n(H(U_1|X_1)+\delta(\eta''))}.$$

Hence, this satisfies the second condition in the Markov Lemma. By the Markov lemma, we have $(U_1^n(M_1), U_2^n(M_2), X_1^n, X_2^n) \in T_{\eta'}^n(U_1U_1X_1X_2)$.

*Step 3:* Show that$(Y^n, U_1^n(M_1), U_2^n(M_2)) \in T_\eta^n(U_1U_2Y)$ with a probability tends to 1 as $n$ increases.

First, $(U_1^n(M_1), U_2^n(M_2)) \leftrightarrow (X_1^n, X_2^n) \leftrightarrow Y^n$ forms a Morkov chain as $(U_1^n(M_1), U_2^n(M_2))$ is a function of $(X_1^n, X_2^n)$. According to *Step 1* and *Step 2*, we have $(U_1^n(M_1), U_2^n(M_2), X_2^n, X_1^n) \in T_{\eta'}^n(U_1U_2X_1X_2)$, and $Y^n$ is drawn $\sim \prod_{i=1}^n P_{Y|X_1X_2}(y_i|x_{1i}, x_{2i})$, hence, by the Markov lemma, we have $(Y^n, U_1^n(M_1), U_2^n(M_2)) \in T_\eta^n(U_1U_2Y)$ with a probability tends to 1 as $n$ increases. This implies that $P_{X_1X_2Y}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3)$ tends to 0 as $n$ increases.

Combining all steps above, we have that $\alpha_n \downarrow 0$ as $n$ increases, hence the type 1 error probability constraint is satisfied.

***b) Type 2 error probability:*** For the type 2 error probability, assume in this case that $H_1$

is true. Then

$$\begin{aligned}\mathbb{E}\{\beta_{nc}\} &= (P_{X_1X_2}P_Y)^n(\mathcal{A}_n) = (P_{X_1X_2}P_Y)^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3^c) \\ &= (P_{X_1X_2}P_Y)^n(\varepsilon_1^c) \times (P_{X_1X_2}P_Y)^n(\varepsilon_2^c) \times (P_{X_1X_2}P_Y)^n(\varepsilon_3^c|\varepsilon_1^c \cap \varepsilon_2^c)\end{aligned}$$

We now bound each factor.

(1) By the covering lemma, $(P_{X_1X_2}P_Y)^n(\varepsilon_1^c) \to 1$ as $n \to \infty$ if $R_1 \geq I(U_1; X_1) + \delta(\eta)$ and $(P_{X_1X_2}P_Y)^n(\varepsilon_2^c) \to 1$ as $n \to \infty$ if $R_2 \geq I(U_2; X_2) + \delta(\eta)$.

(2) For the third term, we have

$$\begin{aligned}&(P_{X_1X_2}P_Y)^n(\varepsilon_3^c|\varepsilon_1^c \cap \varepsilon_2^c) \\ &= \sum_{(u_1^n,u_2^n,y^n)\in T_\eta^n} (P_{X_1X_2}P_Y)^n\{U_1^n(M_1) = u_1^n, U_2^n(M_2) = u_2^n, Y^n = y^n|\varepsilon_1^c \cap \varepsilon_2^c\} \\ &= \sum_{(u_1^n,u_2^n,y^n)\in T_\eta^n} P_{X_1X_2}^n\{U_1^n(M_1) = u_1^n, U_2^n(M_2) = u_2^n|\varepsilon_1^c \cap \varepsilon_2^c\} \times P_Y^n\{Y^n = y^n|\varepsilon_1^c \cap \varepsilon_2^c\} \\ &\leq 2^{n(H(U_1U_2Y)+\delta(\eta))}2^{-n(H(U_1U_2)-\delta(\eta'))}2^{-n(H(Y)-\delta(\eta'))} \\ &= 2^{-n(I(U_1U_2;Y)-\delta(\eta))}.\end{aligned}$$

Combining the bounds on the three factors, we have

$$\mathbb{E}\{\beta_{nc}\} \leq 2^{-n(I(U_1U_2;Y)-\delta(\eta))}.$$

***c) Existence of a particular codebook:*** In summary, combining a) and b) above, we know that, if $R_1 \geq I(U_1; X_1)$ and $R_2 \geq I(U_2; X_2)$, we have

$$\mathbb{E}\{\alpha_{nc}\} = \sum_{c\in\mathcal{C}} \alpha_{nc}\text{Pr}(c) \leq \epsilon, \tag{2.39}$$

$$\mathbb{E}\{\beta_{nc}\} = \sum_{c\in\mathcal{C}} \beta_{nc}\text{Pr}(c) \leq 2^{-n(I(U_1U_2;Y)-\delta(\epsilon))}. \tag{2.40}$$

Let $\epsilon_0 = \epsilon/3$ and define

$$\begin{aligned}\mathcal{C}_1 &\triangleq \{c:\ \alpha_{nc} \le 3\epsilon_0\},\\ \mathcal{C}_2 &\triangleq \left\{c:\ \beta_{nc} \le 3\times 2^{-n(I(U_1U_2;Y)-\delta(\epsilon_0))}\right\}.\end{aligned}$$

As (2.39) and (2.40) are true for any $\epsilon$, then (2.39) and (2.40) hold for $\epsilon_0 = \epsilon/3$ when $n$ is sufficiently large. Then for the type 1 error probability, we have

$$\begin{aligned}\epsilon_0 \ge \mathbb{E}\{\alpha_{nc}\} &= \sum_{c\in\mathcal{C}} \alpha_{nc}\text{Pr}(c)\\ &= \sum_{c\in\mathcal{C}_1} \alpha_{nc}\text{Pr}(c) + \sum_{c\in\bar{\mathcal{C}}_1} \alpha_{nc}\text{Pr}(c)\\ &\ge \sum_{c\in\mathcal{C}_1} \alpha_{nc}\text{Pr}(c) + 3\epsilon_0\text{Pr}\{\bar{\mathcal{C}}_1\}.\end{aligned}$$

This implies $\text{Pr}\{\bar{\mathcal{C}}_1\} \le 1/3$, i.e.,

$$\text{Pr}\{\mathcal{C}_1\} \ge 2/3.$$

Similarly for the type 2 error probability, we have

$$\begin{aligned}&\text{Pr}\left\{\mathcal{C}_2\right\}\\ &= \text{Pr}\left\{c:\ \beta_{nc} \le 2^{-n(I(U_1U_2;Y)-\delta(\epsilon_0)+\frac{\log 3}{n})}\right\}\\ &= \text{Pr}\left\{c:\ \beta_{nc} \le 2^{-n(I(U_1U_2;Y)-\delta(\epsilon_0^*))}\right\}\\ &\ge 2/3.\end{aligned}$$

Therefore

$$\text{Pr}\{\mathcal{C}_1 \cap \mathcal{C}_2\} \ge 1/3,$$

which implies that there exists a codebook $c^*$ such that

$$\begin{aligned} \theta(R_1, R_2, \epsilon) &\geq I(U_1U_2; Y), \\ R_1 &\geq I(U_1; X_1), \\ R_2 &\geq I(U_2; X_2). \end{aligned}$$

Now we have shown that

$$\alpha_n = P^n_{X_1X_2Y}(\mathcal{A}^c_n) \leq \epsilon, \tag{2.41}$$

$$\beta_n = Q^n_{X_1X_2Y}(\mathcal{A}_n) \leq 2^{-n(I(U_1U_2;Y)-\delta(\eta))}, \tag{2.42}$$

if the conditions specified in the theorem are satisfied. Hence, we have (2.35). □

Finally, we establish an upper bound on the type 2 error exponent that any scheme can achieve.

**Theorem 2.6.** In system $S_{X_1X_2|Y}$ with $R_l \geq 0,\ l = 1, 2$, the constraint on type 1 error probability (2.8) and the communication constraints (2.3), the best error exponent for type 2 error probability

$$\lim_{\epsilon\to 0} \theta(R_1, R_2, \epsilon) \leq \max_{U_1U_2} I(U_1U_2; Y) \tag{2.43}$$

in which the maximization is over $U_l$'s such that $R_i \geq I(U_l; X_l)$, $|\mathcal{U}_l| \leq |\mathcal{X}_l| + 1$, $U_1 \to X_1 \to (X_2, Y)$ and $U_2 \to X_2 \to (X_1, Y)$.

*Proof.* We will show that for any encoding and decoding scheme that satisfies the type 1 error constraint $\alpha_n \leq \epsilon$ and rate constraints (2.3), the type 2 error exponent must satisfy (2.43).

First, for any scheme that satisfies the type 1 error and rate constraints, we have

$$
\begin{aligned}
&D(P_{M_1M_2Y^n}||P_{M_1M_2}P_{Y^n}) \\
&= \sum_{(m_1,m_2,y^n)\in\mathcal{A}_n} P_{M_1M_2Y^n}\log\frac{P_{M_1M_2Y^n}}{P_{M_1M_2}P_{Y^n}} + \sum_{(m_1,m_2,y^n)\in\mathcal{A}_n^c} P_{M_1M_2Y^n}\log\frac{P_{M_1M_2Y^n}}{P_{M_1M_2}P_{Y^n}} \\
&\overset{(a)}{\geq} (1-\alpha_n)\log\frac{1-\alpha_n}{\beta_n} + \alpha_n\log\frac{\alpha_n}{1-\beta_n} \\
&= (1-\alpha_n)\log\frac{1}{\beta_n} + \alpha_n\log\frac{1}{1-\beta_n} - H(\alpha_n) \\
&\geq (1-\alpha_n)\log\frac{1}{\beta_n} - H(\alpha_n) \\
&\overset{(b)}{\geq} (1-\epsilon)\log\frac{1}{\beta_n} - H(\alpha_n).
\end{aligned}
$$

where $M_l = f_l(X_l^n),\ l = 1, 2$, $\alpha_n$ and $\beta_n$ are defined in (2.6) and (2.7), and $H(\alpha_n)$ is

$$
H(\alpha_n) \triangleq -(1-\alpha_n)\log(1-\alpha_n) - \alpha_n\log\alpha_n. \tag{2.44}
$$

In the derivation above, $(a)$ is true due to the log sum inequality [11] and $(b)$ follows by the constraint (2.8).

Hence we have the following upper bound

$$
\begin{aligned}
\lim_{\epsilon\to 0}\theta(R_1,R_2,\epsilon) &\leq \lim_{n\to\infty}\frac{1}{n}D(P_{M_1M_2Y^n}||P_{M_1M_2}P_{Y^n}) \\
&= \lim_{n\to\infty}\frac{1}{n}I(M_1M_2;Y^n) \\
&= \lim_{n\to\infty}\frac{1}{n}(H(Y^n) - H(Y^n|M_1M_2)) \\
&= H(Y) - \lim_{n\to\infty}\frac{1}{n}H(Y^n|M_1M_2).
\end{aligned} \tag{2.45}
$$

If we simplify $\frac{1}{n}H(Y^n|M_1M_2)$, we obtain the desired bound. In Appendix A.2, we show that

$$
\frac{1}{n}H(Y^n|M_1M_2) = H(Y|U_1U_2),
$$

for properly chosen $U_1U_2$ satisfying the conditions specified in the statement of theorem. Combing this with (2.45), we obtain the desired result. □

## 2.6 Numerical Results

In this section, we provide numerical results to illustrate the application of the theories developed in Section 2.4 and Section 2.5.

### 2.6.1 Numerical Results for Testing with Zero-rate Compression under Exponential-type Constraints

In Figure 2.5, we illustrate $\sigma_{opt}$, namely the optimal type 2 error exponent characterized in Theorem 2.3, as a function of the type 1 error exponent constraint $r$. For comparison, we also plot the corresponding curve for the centralized case. In the figure, the solid line represents $\sigma_{opt}$ and the dashed line is the optimal type 2 error exponent for the centralized case. In generating Figure 2.5, we set $X_1$, $X_2$ and $Y$ as binary random variables. Furthermore, we set

$$P_{X_1X_2Y} = \left\{\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right\}$$

and

$$Q_{X_1X_2Y} = \left\{\frac{1}{12}, \frac{1}{12}, \frac{5}{72}, \frac{7}{72}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}.$$

It is easy to verify that $D(P_{X_1X_2Y}||Q_{X_1X_2Y}) = 0.0624$. From Figure 2.5, we can see that the type 2 error exponent obtained in the distributed case is smaller than that of the centralized case for every $r$. This is reasonable as in the centralized case, the decision maker has full access to all observations and hence makes less error. Furthermore, the type 2 error exponents for both settings are close to 0 when $r > 0.062$, which makes sense as when $r > D(P_{X_1X_2Y}||Q_{X_1X_2Y})$, no matter what observation is observed, the decision maker decides $H_0$ is true.

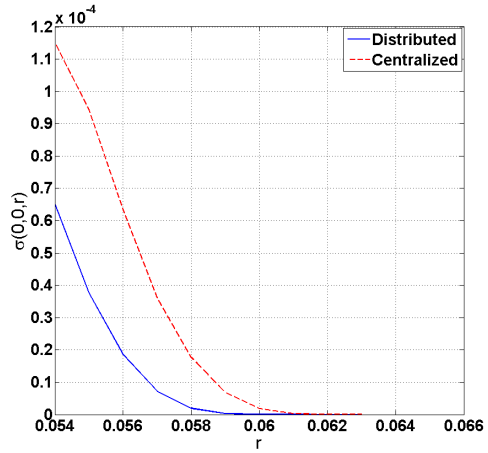



Figure 2.5: $\sigma(0,0,r)$ vs $r$ with $D\left(P_{X_1X_2Y}||Q_{X_1X_2Y}\right) = 0.0624$

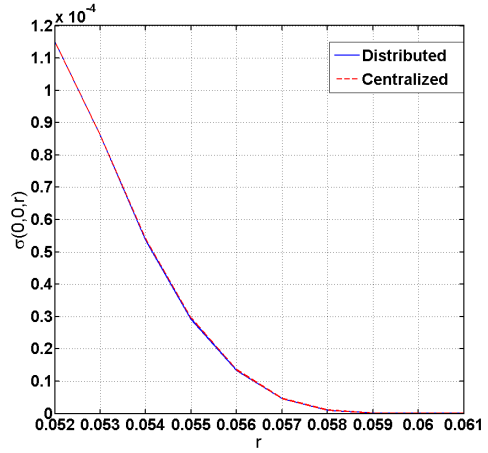



Figure 2.6: $\sigma(0,0,r)$ vs $r$ with $D\left(P_{X_1X_2Y}||Q_{X_1X_2Y}\right) = 0.0588$

Figure 2.6 illustrates $\sigma_{opt}$ for different PMFs. In generating Figure 2.6, we keep $P_{X_1X_2Y}$ same as above, but change $Q_{X_1X_2Y}$ to

$$Q_{X_1X_2Y} = \left\{\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}.$$

In this case, $D(P_{X_1X_2Y}||Q_{X_1X_2Y}) = 0.0588$. From Figure 2.6, we can see that the type 2 error exponent obtained in the distributed setting is quite close to that of the centralized case. This implies that, for certain PMFs, the distributed setting with a proper zero-rate compression can achieve a performance close to that of the centralized setting.

### 2.6.2 Numerical Results for Testing against Independence under Constant-type Constraints

In Figure 2.7, we illustrate $\theta(R_1, R_2, \epsilon)$ discussed in Theorem 2.5 as a function of the rate constraints. In generating this figure, we again set $X_1$, $X_2$ and $Y$ to be binary random variables and set

$$P_{X_1X_2Y} = \left\{\frac{1}{6}, \frac{1}{3}, \frac{1}{12}, \frac{1}{6}, 0, 0, \frac{1}{8}, \frac{1}{8}\right\},$$

from which one can calculate $Q_{X_1X_2Y} = P_{X_1X_2}P_Y$. Furthermore, to make the computation feasible, we assume $|\mathcal{U}_l| = |\mathcal{X}_l| = 2$ in the simulation. In order to visualize the result better, we make $R_1 = R_2 = R$. Hence, we demonstrate a lower bound on the type 2 error exponent achievable using our scheme.

From Figure 2.7, we can see that the type 2 error exponent increases as $R$ increases, which makes sense as the constraint is relaxed, the decision maker can get more information about $X_1^n$ and $X_2^n$, and thus make less error. Furthermore, when $R$ is large enough, the decision maker can fully recover $X_1^n$ and $X_2^n$, which is then the same as the centralized setting. According to Stein's lemma, in the centralized setting, the type 2 error exponent equals $D(P_{X_1X_2Y}||P_{X_1X_2}P_Y)$. In our simulation, $D(P_{X_1X_2Y}||P_{X_1X_2}P_Y) = 0.2229$, and we

can see that the maximum value in Figure 2.7 is quite close to $0.2229$.

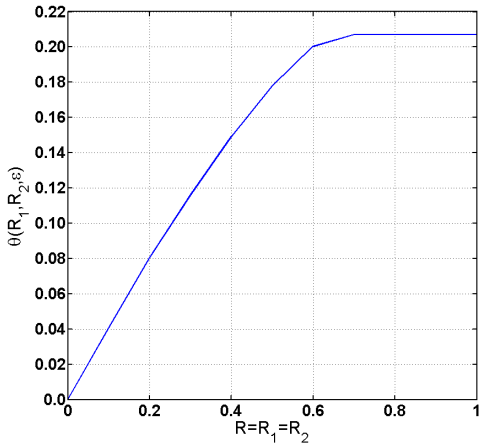



Figure 2.7: $\theta(R_1, R_2, \epsilon)$ vs $R = R_1 = R_2$ with $D(P_{X_1X_2Y}||Q_{X_1X_2Y}) = 0.2229$

## 2.7 Conclusion

In this chapter, we have discussed distributed inference problems with non-interactive encoders. Using properties of $r$-divergence sequences, we have characterized the best error exponent of the type 2 error probability under the zero-rate compression and exponential-type type 1 error probability constraints. Furthermore, we have discussed the problem of testing against independence under the constant-type constraint on the type 1 error probability. We have derived a lower bound and upper-bound on the type 2 error exponent.

# Chapter 3

# Distributed Testing with Cascaded Encoders

## 3.1 Introduction

In this chapter, we discuss the distributed hypothesis testing with cascaded encoders.

We first focus on the problem of testing against independence. This work builds upon Section 2.5, in which we studied the testing against independence case with non-interactive communication. Compared with Section 2.5, this work allows cascaded communication for terminals $\mathcal{X}_l, l = 1, \cdots, L$, so that terminal $\mathcal{X}_l$ can utilize the information from terminals $\mathcal{X}_{l'}, l' = 1, \cdots, l-1$, when it performs encoding. The cascaded communication results in two major differences with the cases using non-interactive communication in [1] and Section 2.5. First, in the non-interactive communication case, one typically converts the testing against independence problem to the problem of source coding with a helper [1], then uses the corresponding results in the source coding with a helper problem to characterize the type 2 error exponent. However, if we follow a similar strategy, then the problem will be related to a source coding with multiple helpers problem, which is still an open problem in network information theory. Second, in the existing work with multiple terminals under

non-interactive communication as studied in Section 2.5, the type 2 error exponent is not fully characterized. However, in our cascaded communication case, as terminals are allowed to use the received massages to perform encoding, we are able to fully characterize the type 2 error exponent for certain scenarios.

We then extend the study to the case with general hypotheses. The problem with non-interactive communication under the same hypotheses was first proposed and studied in [1] and a tighter lower bound was derived in [13]. Different from these works, in which it is assumed that data related to all $X_l, l = 1, \cdots, L$ is stored in one terminal $\mathcal{X}$ (and hence there are two terminals $\mathcal{X}$ and $\mathcal{Y}$ in the model studied in [13] and [1]), we allow data related to $X_l, l = 1, \cdots, L$ to be stored in multiple terminals, and we allow cascaded communications among encoders for encoding. As these two extensions make this problem more complex and no upper bound is derived even for the case with non-interactive communications, in this chapter, we only give a lower bound on the type 2 error exponent given the constraints on the type 1 error probability and communication rates.

Finally, we compare performances of schemes with cascaded and non-interactive communications. Intuitively, compared with the scheme with non-interactive communication in Chapter 2, the decision maker can potentially obtain more information in the cascaded communication case and hence is expected to make a better decision. We show that this is indeed the case by giving an explicit example in which our scheme with cascaded communication achieves a larger type 2 error exponent under different communication rate constraints. On the other hand, we prove that, compared with non-interactive communication, cascaded communication does not offer any improvement in the type 2 error exponent for the zero-rate data compression case.

The remainder of the chapter is organized as follows. In Section 3.2, we introduce the model studied in this chapter. In Section 3.3, we study the problem of testing against independence and extends the result to general case. In Section 3.5, we compare the performances of schemes with cascaded communication and schemes with non-interactive com-

munication. Finally, we offer some concluding remarks in Section 3.6.

## 3.2 Model

In this section, we present our model and summarize the difference between our model and the existing work with non-interactive communication schemes.

Different with the model in Section 2.2, we consider a model illustrated in Figure 3.1. These terminals broadcast messages in a sequential order from terminal 1 until terminal $L$,

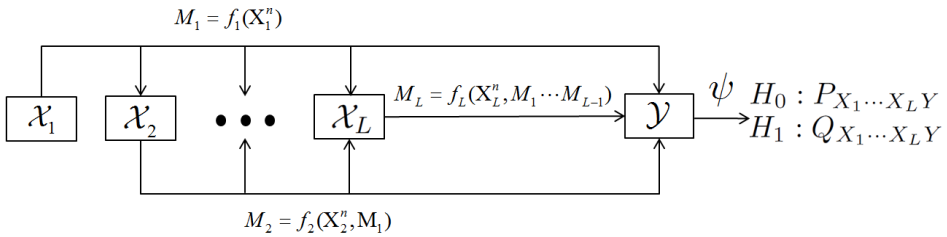



Figure 3.1: Model

and each terminal will use all messages received so far along with its own observations for encoding. More specifically, terminal $\mathcal{X}_1$ will first broadcast its encoded message, which depends only on $X_1^n$, and then terminal $\mathcal{X}_2$ will broadcast its encoded message, which now depends on not only its own observations $X_2^n$ but also the message received from terminal $\mathcal{X}_1$. The process continues until terminal $\mathcal{X}_L$, who will use messages received from $\mathcal{X}_1$ until $\mathcal{X}_{L-1}$ and its own observations $X_L^n$ for encoding. Finally, terminal $\mathcal{Y}$ decides which hypothesis is true based on its own information and the received messages from terminal $\mathcal{X}_1, \cdots, \mathcal{X}_L$.

The main difference between our model and the non-interactive communication model considered in Chapter 2 is that, in the non-interactive communication model, the encoding function of each user relies only on its own observations. That is, the encoding function at terminal $\mathcal{X}_l$ in the non-interactive communication model is given as

$$f_l : \mathcal{X}_l^n \to \mathcal{M}_l = \{1, 2, \ldots, M_l\},\ l = 1, \cdots, L. \tag{3.1}$$

However, in the cascaded case, the encoding function of each user relies not only on its own observations, but also the messages received from other terminals. More specifically, terminal $\mathcal{X}_1$ uses an encoder

$$f_1 : \mathcal{X}_1^n \to \mathcal{M}_1 = \{1, 2, \ldots, M_1\}, \tag{3.2}$$

and terminal $\mathcal{X}_l,\ l = 2, \cdots, L$ uses an encoder

$$f_l : (\mathcal{X}_l^n, \mathcal{M}_1, \cdots, \mathcal{M}_{l-1}) \to \mathcal{M}_l = \{1, 2, \ldots, M_l\}, \tag{3.3}$$

with rates $R_l$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log M_l \leq R_l, \quad l = 1, \cdots, L. \tag{3.4}$$

All remaining definitions are the same as Section 2.2.

In the following, we will use $\theta_{\text{non-interactive}}(R_1, \cdots, R_L, \epsilon)$ to denote the corresponding type 2 error exponent under constant-type constraint on the type 1 error probability in the non-interactive model.

## 3.3 Main Results

In this section, we focus on a special case: testing against independence, in which we are interested in determining whether $(X_1, \cdots, X_L)$ and $Y$ are independent or not.

To simplify our presentation, we first present the results and detailed proof for $L = 2$ case in Section 3.3.1, and then extend the results of the $L = 2$ case to the general case with $L \geq 2$ terminals in Section 3.3.2.

### 3.3.1 $L = 2$ Case

In this subsection, we study the $L = 2$ case in detail. Our goal is to characterize the type 2 error exponent $\theta(R_1, R_2, \epsilon)$ under $\alpha_n \leq \epsilon$. We will show this in two parts. First, we design a scheme and characterize the corresponding error exponent. Then we will show that the scheme is optimal.

Compared with the non-interactive scenario considered in Chapter 2, in our model, $\mathcal{X}_2$ can use the message $f_1(X_1^n)$ from terminal $\mathcal{X}_1$ to perform the encoding. Hence, the coding scheme will be more complex while terminal $\mathcal{Y}$ could potentially receive more information. In the following, we first design a scheme and characterize its error exponent.

**Theorem 3.1.** For the test against independence with $L = 2$ cascaded encoders, the best error exponent for the type 2 error probability satisfies

$$\theta(R_1, R_2, \epsilon) \geq \max_{U_1U_2 \in \varphi_0} I(U_1U_2; Y) \tag{3.5}$$

where

$$\begin{aligned} \varphi_0 = \Big\{ & U_1U_2 : R_1 \geq I(U_1; X_1),\ R_2 \geq I(U_2; X_2|U_1), \\ & U_1 \leftrightarrow X_1 \leftrightarrow (X_2, Y), \qquad (3.6) \\ & U_2 \leftrightarrow (X_2, U_1) \leftrightarrow (X_1, Y), \qquad (3.7) \\ & |\mathcal{U}_1| \leq |\mathcal{X}_1| + 1, |\mathcal{U}_2| \leq |\mathcal{X}_2| \cdot |\mathcal{U}_1| + 1 \Big\}. \end{aligned}$$

*Proof.* In the following, $\eta > \eta' > \eta'' > \eta'''$ are given small numbers.

**Codebook generation.** Fix a joint distribution attaining the maximum in (3.5), which satisfies $P_{U_1U_2|X_1X_2Y} = P_{U_1|X_1}P_{U_2|U_1X_2}$. Let $P_{U_1}(u_1) = \sum_{x_1} P_{X_1}(x_1)P_{U_1|X_1}(u_1|x_1)$, and $P_{U_2|U_1}(u_2|u_1) = \sum_{x_2} P_{X_2|U_1}(x_2|u_1)P_{U_2|U_1X_2}(u_2|u_1, x_2)$. Randomly and independently generate $\lfloor 2^{nR_1} \rfloor$ sequences $u_1^n(m_1)$, $m_1 \in \{1, \cdots, \lfloor 2^{nR_1} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_1}(u_{1i})$. For each $u_1^n(m_1)$, randomly and independently generate $\lfloor 2^{nR_2} \rfloor$ sequences $u_2^n(m_2)$, $m_2 \in$

$\{1, \cdots, \lfloor 2^{nR_2} \rfloor\}$ each according to $\prod_{i=1}^{n} P_{U_2|U_1}(u_{2i}|u_{1i})$. These sequences constitute the codebook $c$, which is revealed to all terminals. This process is shown in Figure 3.2. We use $\mathcal{C}$ to denote the set of all possible codebooks.

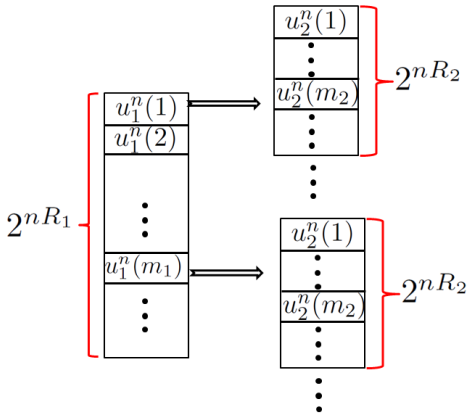



Figure 3.2: Codebook generation

**Encoding for terminal $\mathcal{X}_1$.** Given a sequence $x_1^n$, terminal $\mathcal{X}_1$ finds a $u_1^n(m_1)$ such that $(x_1^n, u_1^n(m_1)) \in T_{\eta''}^n(X_1U_1)$, then it sends the index $m_1$ to both terminal $\mathcal{X}_2$ and $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends 0.

**Encoding for terminal $\mathcal{X}_2$.** If $m_1 = 0$ is received from terminal $\mathcal{X}_1$, terminal $\mathcal{X}_2$ sends $m_2 = 0$ to terminal $\mathcal{Y}$. If $m_1 \neq 0$ is received, given $x_2^n$ and $m_1$, terminal $\mathcal{X}_2$ finds a $u_2^n(m_2)$ such that $(u_1^n(m_1), u_2^n(m_2), x_2^n) \in T_{\eta''}^n(U_1U_2X_2)$ and sends the index $m_2$ to terminal $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends 0.

**Testing.** Upon receiving messages from terminal $\mathcal{X}_1$ and $\mathcal{X}_2$, terminal $\mathcal{Y}$ sets the acceptance region $\mathcal{A}_n$ for $H_0$ to

$$\mathcal{A}_n = \left\{(m_1, m_2, y^n) : (u_1^n(m_1), u_2^n(m_2), y^n) \in T_\eta^n(U_1U_2Y)\right\}.$$

This implies that terminal $\mathcal{Y}$ decides $\hat{H} = H_0$ if and only if no 0 is received and $(u_1^n(m_1), u_2^n(m_2), y^n) \in T_\eta^n(U_1U_2Y)$.

**Analysis of two types of errors.** Terminal $\mathcal{Y}$ chooses $\hat{H} = H_1$ if and only if one or more of the following events occur:

$$\varepsilon_1 = \left\{(U_1^n(m_1), X_1^n) \notin T_{\eta'''}^n(U_1X_1) \text{ for all } m_1 \in [1 : \lfloor 2^{nR_1} \rfloor]\right\},$$
$$\varepsilon_2 = \left\{(U_1^n(M_1), U_2^n(m_2), X_2^n) \notin T_{\eta''}^n(U_1U_2X_2) \text{ for all } m_2 \in [1 : \lfloor 2^{nR_2} \rfloor]\right\},$$
$$\varepsilon_3 = \left\{(U_1^n(M_1), U_2^n(M_2), Y^n) \notin T_{\eta}^n(U_1U_2Y)\right\}.$$

Here, we can see that $\mathcal{A}_n^c = \varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3$.

Using the definition in (2.37) and (2.38), we will then argue that there exists a particular codebook $c^*$ that has the desired properties.

***a) Type 1 error probability:*** To compute the type 1 error probability, we assume that $H_0$ is true. Then

$$\begin{aligned}\mathbb{E}\{\alpha_{nc}\} &= P_{X_1X_2Y}^n(\mathcal{A}_n^c) \\ &= P_{X_1X_2Y}^n(\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3) \\ &\le P_{X_1X_2Y}^n(\varepsilon_1) + P_{X_1X_2Y}^n(\varepsilon_1^c \cap \varepsilon_2) + P_{X_1X_2Y}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3).\end{aligned}$$

We now bound each term.

(1) By the covering lemma [11, Section 3.7], $P_{X_1X_2Y}^n(\varepsilon_1) \to 0$ as $n \to \infty$ if $R_1 \ge I(U_1; X_1) + \delta(\eta''')$.

(2) Since $\eta'' > \eta'''$, $\varepsilon_1^c = \{(U_1^n(M_1), X_1^n) \in T_{\eta'''}^n(U_1X_1)\}$ and $X_2^n|\{X_1^n, U_1^n\} = X_2^n|X_1^n \sim \prod_{i=1}^n P_{X_2|X_1}(x_{2i}|x_{1i})$, by the conditional typicality lemma [11, Section 2.5], then $\Pr\{(U_1^n(M_1), X_1^n, X_2^n) \in T_{\eta''}^n(U_1X_1X_2)\} \to 1$, thus $\Pr\{(U_1^n(M_1), X_2^n) \in T_{\eta''}^n(U_1X_2)\} \to 1$ as $n \to \infty$. Therefore, again by the covering lemma, $P_{X_1X_2Y}^n(\varepsilon_1^c \cap \varepsilon_2) \to 0$ as $n \to \infty$ if $R_2 \ge I(U_2; X_2|U_1) + \delta(\eta'')$.

(3) To bound the last term, we need two steps.

*Step 1*: Since $X_2^n, Y^n | \{X_1^n = x_1^n, U_1^n(M_1) = u_1^n\} \sim \prod_{i=1}^n P_{X_2Y|X_1}(x_{2i}, y_i|x_{1i})$, we can show that $\Pr\{(X_1^n, X_2^n, U_1^n(M_1), Y^n) \in T_{\eta'}^n(X_1X_2U_1Y)\} \to 1$ using the conditional typicality lemma.

*Step 2*: Since we have the Markov chain $U_2^n(M_2) \leftrightarrow (X_2^n, U_1^n(M_1)) \leftrightarrow (X_1^n, Y^n)$ and $(X_1^n, X_2^n, U_1^n(M_1), Y^n) \in T_{\eta'}^n(X_1X_2U_1Y)$ by *Step 1*, we can show that $\Pr\{(U_1^n(M_1), U_2^n(M_2), X_1^n, X_2^n, Y^n) \in T_{\eta}^n(U_1U_2X_1X_2Y)\} \to 1$ as $n \to \infty$ using Markov lemma [11, Section 12.1]. By the covering lemma, we have $\lim_{n\to\infty} \Pr\{(U_2^n(M_2), U_1^n(m_1), X_2^n) \in T_{\eta''}^n(U_2U_1X_2)\} = 1$, that is, $P_{U_2|U_1X_2}^n$ satisfies the first condition in the Markov lemma. Now we prove that the second condition holds.

For all $u_2^n \in T_{\eta''}^n(U_2|x_2^n, u_1^n)$,

$$
\begin{aligned}
&\Pr\{U_2^n(M_2) = u_2^n | X_2^n = x_2^n, U_1^n(M_1) = u_1^n\} \\
&= \Pr\{U_2^n(M_2) = u_2^n, U_2^n(M_2) \in T_{\eta''}^n(U_2|x_2^n, u_1^n) | X_2^n = x_2^n, U_1^n(M_1) = u_1^n\} \\
&\leq \Pr\{U_2^n(M_2) = u_2^n | U_2^n(M_2) \in T_{\eta''}^n(U_2|x_2^n, u_1^n), X_2^n = x_2^n, U_1^n(M_1) = u_1^n\} \\
&= \sum_{m_2} \Pr\{U_2^n(M_2) = u_2^n, M_2 = m_2 | U_2^n(M_2) \in T_{\eta''}^n(U_2|x_2^n, u_1^n), X_2^n = x_2^n, U_1^n(M_1) = u_1^n\} \\
&= \sum_{m_2} \Pr\{U_2^n(M_2) = u_2^n | U_2^n(M_2) \in T_{\eta''}^n(U_2|x_2^n, u_1^n)\} \\
&\qquad \cdot \Pr\{M_2 = m_2 | U_2^n(M_2) \in T_{\eta''}^n(U_2|x_2^n, u_1^n), X_2^n = x_2^n, U_1^n(M_1) = u_1^n\} \\
&\leq 2^{-n(H(U_2|X_2,U_1) - \delta(\eta''))},
\end{aligned}
$$

Hence, this satisfies the second condition in the Markov Lemma. By the Markov lemma, $\Pr\{(U_2^n(M_2), X_2^n, U_1^n(M_1), Y^n) \in T_{\eta}^n(U_2X_2U_1Y)\} \to 1$, i.e. $\Pr\{(U_2^n(M_2), U_1^n(M_1), Y^n) \in T_{\eta}^n(U_2U_1Y)\} \to 1$ as $n \to \infty$. Therefore, $P_{X_1X_2Y}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3) \to 0$ as $n \to \infty$.

***b) Type 2 error probability:*** To calculate the type 2 error probability, assume in this case that $H_1$ is true, then we have

$$\mathbb{E}\{\beta_{nc}\} = (P_{X_1X_2}P_Y)^n(\mathcal{A}_n)$$

$$= (P_{X_1X_2}P_Y)^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3^c)$$
$$= (P_{X_1X_2}P_Y)^n(\varepsilon_1^c) \cdot (P_{X_1X_2}P_Y)^n(\varepsilon_2^c|\varepsilon_1^c) \cdot (P_{X_1X_2}P_Y)^n(\varepsilon_3^c|\varepsilon_1^c \cap \varepsilon_2^c).$$

We now bound each factor.

(1) By the covering lemma, $(P_{X_1X_2}P_Y)^n(\varepsilon_1^c) \to 1$ as $n \to \infty$, if $R_1 \geq I(U_1; X_1) + \delta(\eta''')$.

(2) The second term is the same as that of $H_0$ as it depends only on $P^n_{X_1X_2}$. Hence again by the covering lemma, $(P_{X_1X_2}P_Y)^n(\varepsilon_2^c|\varepsilon_1^c) \to 1$ as $n \to \infty$ if $R_2 \geq I(U_2; X_2|U_1) + \delta(\eta'')$.

(3) For the third term, we have

$$\begin{aligned}
&(P_{X_1X_2}P_Y)^n(\varepsilon_3^c|\varepsilon_1^c \cap \varepsilon_2^c) \\
&= \sum_{(u_1^n,u_2^n,y^n)\in T_\eta^n(U_1U_2Y)} (P_{X_1X_2}P_Y)^n\{U_1^n(M_1) = u_1^n, U_2^n(M_2) = u_2^n, Y^n = y^n|\varepsilon_1^c \cap \varepsilon_2^c\} \\
&\leq 2^{n(H(U_1U_2Y)+\delta(\eta))}2^{-n(H(U_1U_2)-\delta(\eta'))}2^{-n(H(Y)-\delta(\eta'))} \\
&= 2^{-n(I(U_1U_2;Y)-\delta(\eta))}.
\end{aligned}$$

Combining the bounds on these three factors, we have

$$\mathbb{E}\{\beta_{nc}\} \leq 2^{-n(I(U_1U_2;Y)-\delta(\eta))}.$$

***c) Existence of a particular codebook:*** Similar to Section 2.5, we can show that there exists a codebook $c^*$ such that

$$\begin{aligned}
\alpha_{nc^*} &\leq \epsilon, \\
\beta_{nc^*} &\leq 2^{-n(I(U_1U_2;Y)-\delta(\epsilon_0^*))},
\end{aligned}$$

as long as

$$R_1 \geq I(U_1; X_1), \quad R_2 \geq I(U_2; X_2|U_1).$$

This completes the achievability proof.

□

Now we show that the scheme in Theorem 3.1 is optimal.

**Theorem 3.2.** In the testing against independence with $L = 2$ cascaded encoders, when the type 1 error constraint (2.8) is satisfied, the best error exponent for the type 2 error probability satisfies

$$\lim_{\epsilon \to 0} \theta(R_1, R_2, \epsilon) \leq \max_{U_1 U_2 \in \varphi_0} I(U_1 U_2; Y) \tag{3.8}$$

where $\varphi_0$ is defined in Theorem 3.1.

*Proof.* First, for any scheme $(f_1, f_2, \psi)$ that satisfies the type 1 error constraint (2.8) and rate constraints (3.4), we have

$$\begin{aligned} D\left(P_{M_1 M_2 Y^n} || P_{M_1 M_2} P_{Y^n}\right) &\overset{(a)}{\geq} (1-\alpha_n) \log \frac{1-\alpha_n}{\beta_n} + \alpha_n \log \frac{\alpha_n}{1-\beta_n} \\ &\overset{(b)}{\geq} (1-\epsilon) \log \frac{1}{\beta_n} - H(\alpha_n), \end{aligned}$$

in which $M_1 = f_1(X_1^n)$, $M_2 = f_2(X_2^n, M_1)$, $\alpha_n$ and $\beta_n$ are defined in (2.6) and (2.7) respectively. In the above derivation, $(a)$ is true due to the log sum inequality [11], and $(b)$ follows by the constraint (2.8). By the communication constraints (3.4), we have $H(M_l) \leq nR_l,\ l = 1, 2$.

Hence we have the following multi-letter expression of the upper bound

$$\lim_{\epsilon \to 0} \theta(R_1, R_2, \epsilon) \leq \lim_{n \to \infty} \frac{1}{n} D(P_{M_1 M_2 Y^n} || P_{M_1 M_2} P_{Y^n})$$

$$= \lim_{n\to\infty} \frac{1}{n}(I(M_1M_2;Y^n))$$
$$= H(Y) - \lim_{n\to\infty} \frac{1}{n}(H(Y^n|M_1M_2)). \tag{3.9}$$

Then, we single-letterize the upper bound in (3.9) in the following steps.

First consider

$$\begin{aligned} nR_1 &\geq H(M_1) \geq I(M_1; X_1^n X_2^n) \\ &= \sum_{i=1}^{n} I(M_1 X_1^{i-1} X_{2(i+1)}^n; X_{1i} X_{2i}) \\ &\geq \sum_{i=1}^{n} I(M_1 X_1^{i-1} X_{2(i+1)}^n; X_{1i}) \\ &\overset{(a)}{=} \sum_{i=1}^{n} I(U_{1i}; X_{1i}), \end{aligned}$$

where $(a)$ is true by identifying $U_{1i} = (M_1, X_1^{i-1}, X_{2(i+1)}^n)$ and noting that $U_{1i} \leftrightarrow X_{1i} \leftrightarrow (X_{2i}, Y_i)$ forms a Markov chain.

Next consider

$$\begin{aligned} nR_2 &\geq H(M_2) \geq I(M_2; X_1^n X_2^n Y^n | M_1) \\ &= \sum_{i=1}^{n} I(M_2; X_{1i} X_{2i} Y_i | M_1 X_1^{i-1} X_{2(i+1)}^n Y^{i-1}) \\ &\overset{(b)}{=} \sum_{i=1}^{n} I(M_2 Y^{i-1}; X_{1i} X_{2i} Y_i | M_1 X_1^{i-1} X_{2(i+1)}^n) \\ &\geq \sum_{i=1}^{n} I(M_2 Y^{i-1}; X_{2i} | M_1 X_1^{i-1} X_{2(i+1)}^n) \\ &\overset{(c)}{=} \sum_{i=1}^{n} I(U_{2i}; X_{2i} | U_{1i}), \end{aligned}$$

where $(b)$ is true since $Y^{i-1} \leftrightarrow (X_{2(i+1)}^n, X_1^{i-1}, M_1) \leftrightarrow (X_{1i}, X_{2i}, Y_i)$ forms a Markov chain,

which can be derived in the following way,

$$
\begin{aligned}
&(X_1^n, X_{1i}, X_{2i}, Y_i, X_{2(i+1)}^n) \leftrightarrow X_1^{i-1} \leftrightarrow Y^{i-1} \\
&\Rightarrow (M_1, X_{1i}, X_{2i}, Y_i, X_{2(i+1)}^n) \leftrightarrow X_1^{i-1} \leftrightarrow Y^{i-1} \\
&\overset{(d)}{\Rightarrow} (X_{1i}, X_{2i}, Y_i) \leftrightarrow (M_1, X_1^{i-1}, X_{2(i+1)}^n) \leftrightarrow Y^{i-1},
\end{aligned}
$$

in which $(d)$ follows by the weak union property of Markov chain [59]. $(c)$ follows by defining $U_{2i} = (M_2, Y^{i-1})$ and noting that $U_{2i} \leftrightarrow (U_{1i}, X_{2i}) \leftrightarrow (X_{1i}, Y_i)$ forms a Markov chain which is proved in Appendix B.1.

Finally, we consider

$$
\begin{aligned}
H(Y^n|M_1M_2) &= \sum_{i=1}^{n} H(Y_i|M_1M_2Y^{i-1}) \\
&\geq \sum_{i=1}^{n} H(Y_i|M_1M_2Y^{i-1}X_1^{i-1}X_{2(i+1)}^n) \\
&= \sum_{i=1}^{n} H(Y_i|U_{1i}U_{2i}).
\end{aligned}
$$

Define the time-sharing random variable $Q \sim \text{Unif}[1 : n]$ and independent of $(M_1, M_2, X_1^n, X_2^n, Y^n)$, and identify $U_1 = (U_{1Q}, Q),\ U_2 = (U_{2Q}, Q),\ X_1 = X_{1Q},\ X_2 = X_{2Q}$, and $Y = Y_Q$. Clearly, we have $U_1 \leftrightarrow X_1 \leftrightarrow (X_2, Y)$ and $U_2 \leftrightarrow (U_1, X_2) \leftrightarrow (X_1, Y)$ form two Markov chains. Hence we have shown

$$
\begin{aligned}
&R_1 \geq I(U_1; X_1), \quad R_2 \geq I(U_2; X_2|U_1), \\
&\lim_{\epsilon \to 0} \theta(R_1, R_2, \epsilon) \leq H(Y) - H(Y|U_1U_2) = I(Y; U_1U_2),
\end{aligned}
$$

for $P_{U_1U_2|X_1X_2Y} = P_{U_1|X_1}P_{U_2|U_1X_2}$. This completes the converse proof. □

Hence, we obtain a matching upper and lower bound on the type 2 error exponent which is shown in Theorem 3.3.

**Theorem 3.3.** In the testing against independence with $L = 2$ cascaded encoders, when the type 1 error constraint (2.8) is satisfied, the best error exponent for the type 2 error probability satisfies

$$\lim_{\epsilon \to 0} \theta(R_1, R_2, \epsilon) = \max_{U_1 U_2 \in \varphi_0} I(U_1 U_2; Y) \tag{3.10}$$

where $\varphi_0$ is defined in Theorem 3.1.

### 3.3.2 General $L$ Case

The results in the previous subsection can be extended to the general case with $L$ terminals and a decision maker $\mathcal{Y}$. The result is shown in the following theorem.

**Theorem 3.4.** In the testing against independence with $L$ cascaded encoders, the best type 2 error exponent satisfies

$$\theta(R_1, \cdots, R_L, \epsilon) \geq \max_{U_1 \cdots U_L \in \varphi} I(U_1 \cdots U_L; Y), \tag{3.11}$$

in which

$$\begin{aligned} \varphi = \Big\{ & U_1 \cdots U_L : R_1 \geq I(U_1; X_1), R_l \geq I(U_l; X_l | U_1 \cdots U_{l-1}), \\ & U_1 \leftrightarrow X_1 \leftrightarrow (X_2, \cdots, X_L, Y), && (3.12) \\ & U_l \leftrightarrow (X_l, U_1, \cdots, U_{l-1}) \leftrightarrow (X_1, \cdots, X_{l-1}, X_{l+1}, \cdots, X_L, Y), && (3.13) \\ & |\mathcal{U}_1| \leq |\mathcal{X}_1| + 1, \\ & |\mathcal{U}_l| \leq |\mathcal{X}_l| \cdot |\mathcal{U}_{l-1}| \cdots |\mathcal{U}_1| + 1, l = 2, \cdots, L \Big\}. \end{aligned}$$

*Proof.* The proof can be found in Appendix B.2. □

## 3.4 General PMF Case

In this section, we extend our study to the general PMF case (i.e., not necessarily for the test against independence anymore) with $L$ terminals:

$$H_0 : P_{X_1 \cdots X_L Y}, \quad H_1 : Q_{X_1 \cdots X_L Y}.$$

A detailed proof is given for the $L = 2$ case, and the result can be extended to the general case with $L$ terminals.

**Theorem 3.5.** For the case with general hypothesis $P_{X_1X_2Y}$ vs $Q_{X_1X_2Y}$ with $L = 2$ interactive encoders, the best error exponent of the type 2 error probability satisfies

$$\theta(R_1, R_2, \epsilon) \geq \max_{U_1U_2 \in \varphi_0} \min_{\tilde{P}_{U_1U_2X_1X_2Y} \in \xi_0} D\left(\tilde{P}_{U_1U_2X_1X_2Y} || Q_{U_1U_2X_1X_2Y}\right) \tag{3.14}$$

where $\varphi_0$ is defined in Theorem 3.1,

$$\xi_0 = \left\{\tilde{P}_{U_1U_2X_1X_2Y} : \tilde{P}_{U_1X_1} = P_{U_1X_1}, \tilde{P}_{U_1U_2X_2} = P_{U_1U_2X_2}, \tilde{P}_{U_1U_2Y} = P_{U_1U_2Y}\right\}.$$

and $Q_{U_1|X_1} = P_{U_1|X_1}$, $Q_{U_2|U_1X_1X_2} = P_{U_2|U_1X_1X_2}$.

*Proof.* In the following, $\epsilon > \epsilon' > \epsilon'' > \epsilon'''$ are given small numbers.

**Codebook generation.** Fix a joint distribution attaining the maximum if (3.14), which satisfies $P_{U_1U_2|X_1X_2Y} = P_{U_1|X_1}P_{U_2|U_1X_2}$. Let $P_{U_1}(u_1) = \sum_{x_1} P_{X_1}(x_1)P_{U_1|X_1}(u_1|x_1)$, and $P_{U_2|U_1}(u_2|u_1) = \sum_{x_2} P_{X_2|U_1}(x_2|u_1)P_{U_2|U_1X_2}(u_2|u_1, x_2)$. Randomly and independently generate $||f_1|| = 2^{n(I(U_1;X_1)+\eta)}$ sequences $u_1^n(m_1)$, $m_1 \in \{1, \cdots, ||f_1||\}$ each according to $\prod_{i=1}^n P_{U_1}(u_{1i})$. For each $u_1^n(m_1)$, randomly and independently generate $||f_2|| = 2^{n(I(U_2;X_2|U_1)+\eta)}$ sequences $u_2^n(m_2)$, $m_2 \in \{1, \cdots, ||f_2||\}$ each according to $\prod_{i=1}^n P_{U_2|U_1}(u_{2i}|u_{1i})$. These sequences constitute the codebook $c$, which is revealed to all terminals. We use $\mathcal{C}$ to denote the set of all possible codebooks.

**Encoding for terminal $\mathcal{X}_1$.** Given a sequence $x_1^n$, terminal $\mathcal{X}_1$ finds a $u_1^n(m_1)$ such that $(x_1^n, u_1^n(m_1)) \in T_{\epsilon'''}^{(n)}(X_1U_1)$, then it sends the index $m_1$ to both terminal $\mathcal{X}_2$ and $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends $0$.

**Encoding for terminal $\mathcal{X}_2$.** If $0$ is received from terminal $\mathcal{X}_1$, terminal $\mathcal{X}_2$ sends $0$ to terminal $\mathcal{Y}$. If $m_1 \neq 0$ is received, given $x_2^n$ and $m_1$, terminal $\mathcal{X}_2$ finds a $u_2^n(m_2)$ such that $(u_1^n(m_1), u_2^n(m_2), x_2^n) \in T_{\epsilon''}^{(n)}(U_1U_2X_2)$ and sends the index $m_2$ to terminal $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends $0$.

**Testing.** Upon receiving messages from terminal $\mathcal{X}_1$ and $\mathcal{X}_2$, terminal $\mathcal{Y}$ sets the acceptance region $\mathcal{A}_n$ for $H_0$ to

$$\mathcal{A}_n = \left\{(m_1, m_2, y^n) : (u_1^n(m_1), u_2^n(m_2), y^n) \in T_\epsilon^{(n)}(U_1U_2Y)\right\}.$$

This implies that terminal $\mathcal{Y}$ decides $\hat{H} = H_0$ if and only if no $0$ is received and $(u_1^n(m_1), u_2^n(m_2), y^n) \in T_\epsilon^{(n)}(U_1U_2Y)$.

**Analysis of two types of errors.** Terminal $\mathcal{Y}$ chooses $\hat{H} \neq H_0$ if and only if one or more of the following events occur:

$$\begin{aligned}
\varepsilon_1 &= \left\{(U_1^n(m_1), X_1^n) \notin T_{\epsilon'''}^{(n)}(U_1X_1) \text{ for all } m_1 \in [1 : 2^{nR_1}]\right\}, \\
\varepsilon_2 &= \left\{(U_1^n(M_1), U_2^n(m_2), X_2^n) \notin T_{\epsilon''}^{(n)}(U_1U_2X_2) \text{ for all } m_2 \in [1 : 2^{nR_2}]\right\}, \\
\varepsilon_3 &= \left\{(U_1^n(M_1), U_2^n(M_2), Y^n) \notin T_\epsilon^{(n)}(U_1U_2Y)\right\}.
\end{aligned}$$

Hence, we can see that $\mathcal{A}_n^c = \varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3$.

Using the definition in (2.37) and (2.38), we will then argue that there exists a particular codebook $c^*$ that has the desired properties. The analysis of the type 1 error probability is similar to the analysis in Section 3.3.1, but the analysis of the type 2 error probability is substantially more involved.

***a) Type 1 error probability:*** To compute the type 1 error probability, we assume that $H_0$ is true. Then

$$\begin{aligned}\mathbb{E}\{\alpha_{nc}\} &= P^n_{X_1X_2Y}(\mathcal{A}_n^c) = P^n_{X_1X_2Y}(\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3) \\ &\leq P^n_{X_1X_2Y}(\varepsilon_1) + P^n_{X_1X_2Y}(\varepsilon_1^c \cap \varepsilon_2) + P^n_{X_1X_2Y}(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3).\end{aligned} \tag{3.15}$$

From (3.15) we can see that type 1 error probability only relies on $P_{X_1X_2Y}$, which is the same as in Section 3.3.1. Hence, the analysis of the type 1 error probability is the same as that in Section 3.3.1 as the rate constraints $R_1 \geq I(U_1;X_1)$ and $R_2 \geq I(U_2;X_2|U_1)$ are satisfied. The detailed analysis is omitted here.

***b) Type 2 error probability:*** To calculate the type 2 error probability, assume in this case that $H_1$ is true. For $m_1 \in [1 : M_1]$, $m_2 \in [1 : M_2]$, and $y^n \in T_\epsilon^{(n)}(Y|u_1^n(m_1), u_2^n(m_2))$, define

$$S_{m_1,m_2}(y^n) = \{u_1^n(m_1)\} \times \{u_2^n(m_2)\} \times T_\epsilon^{(n)}(X_1|u_1^n(m_1)) \times T_\epsilon^{(n)}(X_2|u_1^n(m_1)u_2^n(m_2)) \times \{y^n\},$$

and

$$\varphi_n = \bigcup_{m_1=1}^{||f_1||} \bigcup_{m_2=1}^{||f_2||} \bigcup_{y^n \in T_\epsilon^{(n)}(Y|u_1^n(m_1),u_2^n(m_2))} S_{m_1,m_2}(y^n).$$

Suppose $U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}$ is a type variable of $(u_1^n, u_2^n, x_1^n, x_2^n, y^n) \in S_{m_1,m_2}(x_2^n, y^n)$, then

$$Q^n_{X_1X_2Y}(x_1^n, x_2^n, y^n) = \exp\left[-n\left(H\left(X_1^{(n)}X_2^{(n)}Y^{(n)}\right) + D\left(X_1^{(n)}X_2^{(n)}Y^{(n)}||Q_{X_1X_2Y}\right)\right)\right].$$

Denoting $N\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right)$ the number of those elements $(u_1^n, u_2^n, x_1^n, x_2^n, y^n) \in$

$\varphi_n$ that have $\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right)$ as their type variable, it follows that

$$N\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right) \leq \exp\Big[n\Big(I\left(X_1;U_1\right)+I\left(X_2;U_2|U_1\right) \\ +H\left(X_1^{(n)}X_2^{(n)}|U_1^{(n)}U_2^{(n)}Y^{(n)}\right) \\ +H\left(Y|U_1U_2\right)+2\eta+2\epsilon\Big)\Big].$$

Hence,

$$\mathbb{E}\{\beta_{nc}\} = Q_{X_1X_2Y}^n(\mathcal{A}_n) \\ \leq \sum_{U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}} \exp\left[-n\left(k\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right)-2\eta-2\epsilon\right)\right], \quad (3.16)$$

where

$$k\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right) = H\left(X_1^{(n)}X_2^{(n)}Y^{(n)}\right)+D\left(X_1^{(n)}X_2^{(n)}Y^{(n)}||Q_{X_1X_2Y}\right) \\ -I\left(X_1;U_1\right)-I\left(X_2;U_2|U_1\right)-H\left(Y|U_1U_2\right) \\ -H\left(X_1^{(n)}X_2^{(n)}|U_1^{(n)}U_2^{(n)}Y^{(n)}\right), \quad (3.17)$$

and the sum is taken over all possible type variables of elements $(u_1^n, u_2^n, x_1^n, x_2^n, y^n) \in \varphi_n$. Hence, we have $(u_1^n(m_1), x_1^n) \in T_\epsilon^{(n)}(U_1X_1)$, $(u_1^n(m_1), u_2^n(m_2), x_2^n) \in T_\epsilon^{(n)}(U_1U_2X_2)$, and $(u_1^n(m_1), u_2^n(m_2), y^n) \in T_\epsilon^{(n)}(U_1U_2Y)$. This implies that the sum ranges over all possible type variables $\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right)$ such that, for all $u_1 \in \mathcal{U}_1$, $u_2 \in \mathcal{U}_2$, $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, and $y \in \mathcal{Y}$,

$$|P_{U_1^{(n)}X_1^{(n)}}(u_1x_1) - P_{U_1X_1}(u_1x_1)| \leq \epsilon P_{U_1X_1}(u_1x_1),$$
$$|P_{U_1^{(n)}U_2^{(n)}X_2^{(n)}}(u_1u_2x_2) - P_{U_1U_2X_2}(u_1u_2x_2)| \leq \epsilon P_{U_1U_2X_2}(u_1u_2x_2),$$
$$|P_{U_1^{(n)}U_2^{(n)}Y^{(n)}}(u_1u_2y) - P_{U_1U_2Y}(u_1u_2y)| \leq \epsilon P_{U_1U_2Y}(u_1u_2y).$$

Thus, we can rewrite (3.17) as

$$\begin{aligned} k\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right) = H\left(\tilde{X}_1\tilde{X}_2\tilde{Y}\right) + D\left(\tilde{X}_1\tilde{X}_2\tilde{Y}||Q_{X_1X_2Y}\right) \\ -I\left(\tilde{X}_1;\tilde{U}_1\right) - I\left(\tilde{X}_2;\tilde{U}_2|\tilde{U}_1\right) - H\left(\tilde{Y}|\tilde{U}_1\tilde{U}_2\right) \\ -H\left(\tilde{X}_1\tilde{X}_2|\tilde{U}_1\tilde{U}_2\tilde{Y}\right) + \delta(\epsilon), \end{aligned} \tag{3.18}$$

with some variable $\tilde{U}_1\tilde{U}_2\tilde{X}_1\tilde{X}_2\tilde{Y}$ such that

$$\tilde{P}_{U_1X_1} = P_{U_1X_1},\ \tilde{P}_{U_1U_2X_2} = P_{U_1U_2X_2}, \tilde{P}_{U_1U_2Y} = P_{U_1U_2Y},$$

where $\delta(\epsilon) \to 0$. Through some calculation, we can get

$$k\left(U_1^{(n)}U_2^{(n)}X_1^{(n)}X_2^{(n)}Y^{(n)}\right) = D\left(\tilde{P}_{U_1U_2X_1X_2Y}||Q_{U_1U_2X_1X_2Y}\right) + \delta(\epsilon),$$

where $Q_{U_1|X_1} = P_{U_1|X_1}$, $P_{U_2|U_1X_2} = Q_{U_2|U_1X_2}$.

Thus, by (3.16) and (3.19), we have

$$\begin{aligned} &\mathbb{E}\{\beta_{nc}\} \le (n+1)^{|\mathcal{U}_1|\cdot|\mathcal{U}_2|\cdot|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdot|\mathcal{Y}|} \\ &\max_{\tilde{P}_{U_1U_2X_1X_2Y}\in\xi(U_1U_2)} \exp\left[-n\left(D\left(\tilde{P}_{U_1U_2X_1X_2Y}||Q_{U_1U_2X_1X_2Y}\right) + \delta(\epsilon) - 2\eta - 2\epsilon\right)\right]. \end{aligned}$$

***c) Existence of a particular codebook:*** Using similar arguments in Section 3.3.1, we can see that there exists a codebook $c^*$ such that

$$\begin{aligned} &\alpha_{nc^*} \le \epsilon, \\ &\beta_{nc^*} \le (n+1)^{|\mathcal{U}_1|\cdot|\mathcal{U}_2|\cdot|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdot|\mathcal{Y}|} \\ &\max_{\tilde{P}_{U_1U_2X_1X_2Y}\in\xi(U_1U_2)} \exp\Big[-n\left(D\left(\tilde{P}_{U_1U_2X_1X_2Y}||Q_{U_1U_2X_1X_2Y}\right) + \delta(\epsilon)\right)\Big], \end{aligned}$$

as long as

$$R_1 \geq I(U_1; X_1), \quad R_2 \geq I(U_2; X_2|U_1).$$

As $\epsilon > 0$ is arbitrary, it is concluded that

$$\theta(R_1, R_2, \epsilon) \geq \min_{\tilde{P}_{U_1U_2X_1X_2Y} \in \xi(U_1U_2)} D\left(\tilde{P}_{U_1U_2X_1X_2Y} || Q_{U_1U_2X_1X_2Y}\right).$$

This completes the achievability proof. □

The achievable scheme could be potentially improved by employing binning scheme [31, 36]. However, the obtained error exponent form is very complicated and hence we omit it. The achieve scheme above can also be easily extended to the general $L$ case.

**Theorem 3.6.** For the case with general hypothesis $P_{X_1\cdots X_LY}$ vs $Q_{X_1\cdots X_LY}$ with $L$ interactive encoders, the best error exponent of the type 2 error probability satisfies

$$\theta(R_1, \cdots, R_2, \epsilon) \geq \max_{U_1\cdots U_L \in \varphi} \min_{\tilde{P}_{U_1\cdots U_LX_1\cdots X_LY} \in \xi} D\left(\tilde{P}_{U_1\cdots U_LX_1\cdots X_LY} || Q_{U_1\cdots U_LX_1\cdots X_LY}\right),$$

where $\varphi$ is defined in Theorem 3.4 and

$$\xi = \left\{\tilde{P}_{U_1\cdots U_LX_1\cdots X_LY} : \tilde{P}_{U_1\cdots U_lX_l} = P_{U_1\cdots U_lX_l}, \tilde{P}_{U_1\cdots U_LY} = P_{U_1\cdots U_LY},\ l = 1, \cdots, L\right\}.$$

## 3.5 Comparison with the Non-interactive Communication Model

In this section, we compare the performance achieved by the cascaded communication scheme and that of the non-interactive communication scheme. We will provide concrete examples to show that for certain PMF and positive communication rates, the scheme with cascaded

communication outperforms that of the non-interactive communication scheme. On the other hand, we will also prove that when the communication rates go to zero (zero-rate compression), the performance of the cascaded communication scheme is the same as that of the non-interactive communication scheme, and hence the cascaded scheme does not improve the performance in these scenarios.

### 3.5.1 Example When the Cascaded Scheme Is Better Than the Non-interactive Scheme

Here, we provide an example in which the error exponent achieved using the cascaded scheme is larger than that can be achieved using the non-interactive scheme. The example is about the testing against independence case. The testing against independence problem with non-interactive communications is studied in Section 2.5, which provides a lower and an upper bound on the type 2 error exponent of non-interactive schemes. As the lower and upper bounds in Section 2.5 do not match with each other, in this part, we compare the type 2 error exponent achieved by the scheme with cascaded communications shown in the proof of Theorem 3.1, with the upper bound on the type 2 error exponent of the non-interactive scheme derived in Theorem 2.5.

In the example, we let $X_1,\ X_2$ and $Y$ be binary random variables with joint PMF $P_{X_1X_2Y}$, which is shown in Table 3.1. For testing against independence case, we have

Table 3.1: The joint PMF $P_{X_1X_2Y}$

| $X_1X_2Y$ | 000 | 010 | 100 | 110 |
|---|---|---|---|---|
| $P_{X_1X_2Y}$ | 0.0704 | 0.2108 | 0.0015 | 0.3233 |
| $X_1X_2Y$ | 001 | 011 | 101 | 111 |
| $P_{X_1X_2Y}$ | 0.2206 | 0.0667 | 0.0046 | 0.1021 |

$Q_{X_1X_2Y} = P_{X_1X_2}P_Y$, which can be easily calculated from Table 3.1. With given communication constraint $R = R_1 = R_2$, we use Theorem 3.1 to find the best value of the type 2 error exponent that we can achieve using our cascaded scheme. For comparison, we also use

Theorem 2.5 to find an upper bound on the type 2 error exponent of the non-interactive case. For $R = 0.48$, we list the conditional distributions $P_{U_1|X_1}$ and $P_{U_2|X_2}$ for non-interactive case in Table 3.2 and the conditional distributions $P_{U_1|X_1}$ and $P_{U_2|X_2U_1}$ for the cascaded case in Table 3.3.

Table 3.2: $P_{U_1|X_1}$ and $P_{U_2|X_2}$ for non-interactive case when $R = 0.48$

| $U_1\|X_1$ | $0\|0$ | $1\|0$ | $0\|1$ | $1\|1$ |
|---|---|---|---|---|
| $P_{U_1\|X_1}$ | 0.9991 | 0.0009 | 0.1564 | 0.8436 |
| $U_2\|X_2$ | $0\|0$ | $1\|0$ | $0\|1$ | $1\|1$ |
| $P_{U_2\|X_2}$ | 0.9686 | 0.0314 | 0.0357 | 0.9643 |

Table 3.3: $P_{U_1|X_1}$ and $P_{U_2|X_2U_1}$ for cascaded case when $R = 0.48$

| $U_1\|X_1$ | $0\|0$ | $1\|0$ | $0\|1$ | $1\|1$ |
|---|---|---|---|---|
| $P_{U_1\|X_1}$ | 0.0155 | 0.9845 | 0.5829 | 0.4171 |
| $U_2\|X_2U_1$ | $0\|00$ | $1\|00$ | $0\|01$ | $1\|01$ |
| $P_{U_2\|X_2U_1}$ | 0.0636 | 0.9364 | 0.9727 | 0.0273 |
| $U_2\|X_2U_1$ | $0\|10$ | $1\|10$ | $0\|11$ | $1\|11$ |
| $P_{U_2\|X_2U_1}$ | 0.9898 | 0.0102 | 0.0005 | 0.9995 |

The simulation results for different $R$s are shown in Figure 3.3. From Figure 3.3, we can see that the type 2 error exponents in both cases increase with the increasing value of $R$, which makes sense as the more information we can send, the less errors we will make. We also observe that the type 2 error exponent achieved using our cascaded communication scheme is even larger than an upper bound on the type 2 error exponent of any non-interactive schemes. Hence, we confirm the intuitive idea that more information offered by the cascaded communication facilitates a better decision making for certain testing against independence cases with positive communication rates.

We also list the error exponents for $R \geq 0.46$ in Table 3.4 since it is not obvious to see the increase in the performance of both cascaded and non-interactive communication. Note that when $R$ is large enough, we can let $U_1 = X_1$ and $U_2 = X_2$. And we have $\theta_{\text{non-ineractive}} \leq I(X_1X_2;Y) = \theta$. To achieve this maximum value, the constraints of $R_1$ and $R_2$ can be simplified in the following:

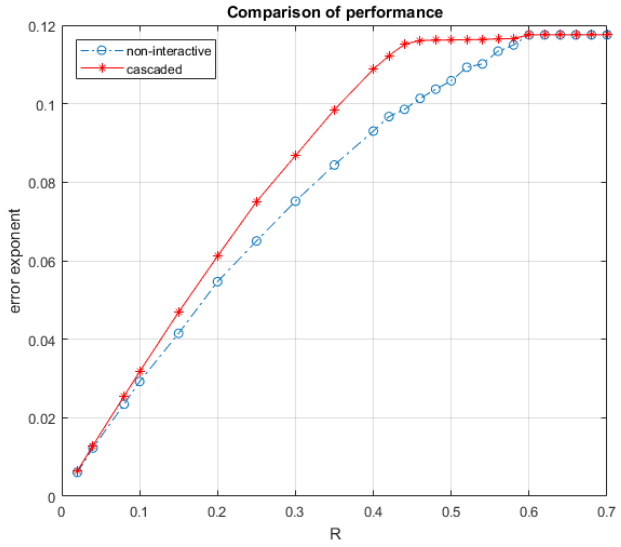



Figure 3.3: Simulation results

- Non-interactive case:

$$R_1 \geq H(X_1), \quad R_2 \geq H(X_2). \tag{3.19}$$

- Cascaded case:

$$R_1 \geq H(X_1), \quad R_2 \geq H(X_2|X_1). \tag{3.20}$$

We also list the values these theoretic limits in Table 3.5.

Table 3.4: Error exponents for $R \geq 0.42$

| $R$ | 0.42 | 0.46 | 0.50 | 0.52 |
|---|---|---|---|---|
| $\theta_{\text{non-interactive}}$ | 0.096724 | 0.10136 | 0.10585 | 0.10930 |
| $\theta$ | 0.11222 | 0.11612 | 0.11625 | 0.11630 |
| $R$ | 0.54 | 0.58 | 0.62 | 0.66 |
| $\theta_{\text{non-interactive}}$ | 0.11013 | 0.11496 | 0.11754 | 0.11754 |
| $\theta$ | 0.11640 | 0.11661 | 0.11755 | 0.11755 |
| $R$ | 0.68 | 0.70 | 0.72 | 0.74 |
| $\theta_{\text{non-interactive}}$ | 0.11754 | 0.11754 | 0.11754 | 0.11754 |
| $\theta$ | 0.11755 | 0.11755 | 0.11755 | 0.11755 |

Table 3.5: Theoretic limits for $U_1 = X_1$ and $U_1 = X_2$

| $I(X_1X_2;Y)$ | $H(X_1)$ | $H(X_2)$ | $H(X_2\|X_1)$ |
|---|---|---|---|
| 0.1187 | 0.6838 | 0.5127 | 0.4259 |

From Table 3.4 and Table 3.5, we can see that the increasing speed of $\theta$ decreases when $R \geq 0.42$ as $R$ is large enough for terminal $\mathcal{X}_2$. This same happens for $\theta_{\text{non-interactive}}$ when $R \geq 0.52$. Furthermore, both $\theta_{\text{non-interactive}}$ and $\theta$ approach the best possible value of $I(X_1, X_2; Y) = 0.1187$ as $R \to 0.68$. We note that there is a slight gap between the theoretic limit and simulation results. This is due to the precision of the numerical simulation.

### 3.5.2 Example When the Cascaded Scheme Has the Same Performance as that of the Non-interactive Scheme

In this subsection, we provide an example for which the cascaded scheme has the same performance as that of the non-interactive scheme. In particular, we will prove that, under zero-rate data compression, i.e. $R_l = 0,\ l = 1, \cdots, L$, cascaded communication does not improve the performance.

In the non-interactive communication scenario with zero-rate compression, a matching upper bound and lower bound on the type 2 error exponent was provided in [34, Theorem 2] when $Q_{X_1 \cdots X_L Y} > 0$. If we can prove an upper bound on the type 2 error exponent for the cascaded communication case that is no larger the error exponent shown in [34], then we can arrive at the conclusion that the cascaded communication won't help under the zero-rate compression case.

For reference, we state the error exponent of the non-interactive scheme characterized in [34] in the following.

**Theorem 3.7.** ([34]) Let $P_{X_1 \cdots X_L Y}$ be arbitrary and $Q_{X_1 \cdots X_L Y} > 0$, for all $\epsilon \in [0, 1)$, the type 2 error exponent for zero-rate compression under $\alpha_n \leq \epsilon$ with $L$ non-interactive encoders is

given by

$$\theta_{\text{non-interactive}}(0,\cdots,0,\epsilon) = \min_{\tilde{P}_{X_1\cdots X_L Y}\in\mathcal{L}} D\left(\tilde{P}_{X_1\cdots X_L Y}||Q_{X_1\cdots X_L Y}\right) \tag{3.21}$$

where

$$\mathcal{L} = \left\{\tilde{P}_{X_1\cdots X_L Y} : \tilde{P}_{X_l} = P_{X_l},\ l = 1,\cdots,L,\ \tilde{P}_Y = P_Y\right\}.$$

In the following, we provide an upper bound on the type 2 error exponent for the cascaded case.

**Theorem 3.8.** Let $P_{X_1\cdots X_L Y}$ be arbitrary and $Q_{X_1\cdots X_L Y} > 0$, for all $\epsilon \in [0,1)$, the best type 2 error exponent for zero-rate compression under $\alpha_n \leq \epsilon$ with $L$ cascaded encoders satisfies

$$\theta(0,\cdots,0,\epsilon) \leq \min_{\tilde{P}_{X_1\cdots X_L Y}\in\mathcal{L}} D\left(\tilde{P}_{X_1\cdots X_L Y}||Q_{X_1\cdots X_L Y}\right) \tag{3.22}$$

where $\mathcal{L}$ is defined in Theorem 3.7.

*Proof.* Please see Appendix B.3. □

Comparing Theorem 3.8 with Theorem 3.7, we can see that the upper bound on the type 2 error exponent for the cascaded communication scheme is the same as the type 2 error exponent achievable by the non-interactive communication scheme. This implies that the performance of the cascaded communication scheme is the same as that of the non-interactive communication scheme in the zero-rate data compression case.

The conclusion that cascaded communication does not improve the type 2 error exponent under the zero-rate data compression case also holds when we have the exponential-type constraint on the type 1 error probability defined in (2.9).

In the cascaded communication case, based on the results in Theorem 3.8, we can use a similar strategy as in [15] to convert the problem under the exponential-type constraint (2.9)

to the corresponding problem under the constraint in (2.8). As the converting strategy is independent of the communication style, it will be the same as that in Section 2.4. Then an upper bound on the type 2 error exponent under the exponential-type constraint can be easily derived without going into details, shown in the sequel.

**Theorem 3.9.** Let $P_{X_1\cdots X_L Y}$ be arbitrary and $Q_{X_1\cdots X_L Y} > 0$, the best type 2 error exponent for zero-rate compression case under (2.9) with $L$ cascaded encoders satisfies

$$\sigma(0,\cdots,0,r) \leq \min_{\tilde{P}_{X_1\cdots X_L Y}\in\mathcal{H}_r} D\left(\tilde{P}_{X_1\cdots X_L Y}||Q_{X_1\cdots X_L Y}\right) \tag{3.23}$$

where

$$\mathcal{H}_r = \Big\{\tilde{P}_{X_1\cdots X_L Y} : \tilde{P}_{X_l} = \hat{P}_{X_l}, \tilde{P}_Y = \hat{P}_Y, \, l = 1,\cdots,L \\ \text{for some } \hat{P}_{X_1\cdots X_L Y} \in \varphi_r\Big\}, \tag{3.24}$$

$$\varphi_r = \left\{\hat{P}_{X_1\cdots X_L Y} : D(\hat{P}_{X_1\cdots X_L Y}||P_{X_1\cdots X_L Y}) \leq r\right\}. \tag{3.25}$$

Comparing Theorem 3.9 with Theorem 2.4, where a matching upper and lower bound is provided for the non-interactive scheme, we can conclude that there is no gain in performance on the type 2 error exponent under zero-rate compression with the exponential-type constraint on the type 1 error probability.

## 3.6 Conclusion

In the chapter, we have considered distributed testing problems with cascaded encoders. We have first investigated the special case of testing against independence. We have designed a scheme to benefit from the extra information provided by cascaded communications, and have shown that the proposed scheme is optimal when certain Markovian relation exists. We have then derived a lower bound on the type 2 error exponent for cases with general

hypotheses. Compared with existing results in the non-interactive communication cases, we have shown that cascaded communication does provide performance gain under certain PMFs and positive communication rates but does not offer gain under zero-rate data compression scenarios.

# Chapter 4

# Distributed Identity Inference with Data Compression

## 4.1 Introduction

In this chapter, we extend our study to the case with model uncertainties.

We first focus on the zero-rate compression problem under exponential-type constraint on the type 1 error probability. Compared with [34], in which a composite hypothesis testing problem was studied under zero-rate compression and a constant-type constraint on the type 1 error probability, our exponential-type error probability constraint is much stricter and a more complex coding/decoding scheme is needed. Interestingly, the encoding scheme in [16] is universal, i.e., it does not depend on the the distributions under either hypothesis. The decoding scheme in [16], however, depends on the knowledge of distribution in $H_0$ and hence is not universal. As $H_0$ is composite in our case, the decoding scheme in[16] is not applicable anymore. By devising a new universal decoding scheme and providing a matching upper-bound, we fully characterize the type 2 error exponent under the zero-rate compression and the exponential-type constraint on the type 1 error probability.

We then extend our study to identity testing problem with positive-rate compression,

which has not been studied in the area of distributed hypothesis testing. This problem is related to parameter estimation with multi-terminal compression [14], in which the decision maker tries to estimate the unknown parameter in the joint distribution of the data at different terminals after receiving the compressed messages from each terminal. In [14], the authors first provided a universal coding scheme for encoding, then employed minimum-entropy decoding to recover the messages sent by terminals, and finally derived the asymptotic distribution of the joint type of the sequences and messages to get the maximum-likelihood estimators. We can follow their scheme and calculate the two types of error probabilities by utilizing the asymptotic distribution of the joint types of sequences and messages. However, due to complex derivations involved, the obtained type 2 error exponent bound is very complicated and does not provide meaningful insights. Instead of following this route, we take an alternative approach to obtain meaningful performance bound. The enabling observation of our scheme is that our goal is only to determine whether the data $(X^n, Y^n)$ is generated by a PMF in $H_0$ or not, but we do not care which particular PMF in $H_0$ is used to generate the data. Hence, it is not necessary to estimate the unknown parameter. Based on this observation, we employ a universal encoding scheme similar to the one used in [14] but design a different universal decoding scheme that determines the hypothesis without first making parameter estimations, and further characterize the type 2 error exponent of this scheme. Using this idea, we first investigate the constant-type error probability constraint case. We establish a lower bound on the type 2 error exponent for general PMF. We then investigate the special case of testing against independence, in which we are interested in whether $X$ and $Y$ are independent or not. Due to this special requirement, we can simplify our coding/decoding scheme further and also establish a matching upper bound on the type 2 error exponent, and hence fully characterize the performance for this special case. We then extend our study to the more challenging case with exponential-type constraint on the type 1 error probability, and design a scheme to provide a lower bound on the type 2 error exponent.

The remainder of the chapter is organized as follows. In Section 4.2, we introduce the

model studied in this chapter. In Section 4.3, we provide an important lemma that play an important part in the proof. In Section 4.4, we present the results of identity testing with zero-rate compression. Section 4.5 show the results of identity testing with positive-rate compression. Finally, we offer some concluding remarks in Section 4.6.

## 4.2 Model

To simplify our presentation, we assume we only have terminal $\mathcal{X}$ and terminal $\mathcal{Y}$. In this chapter, our goal is to determine whether the true joint distribution is the same as the given distribution $Q_{XY}$ or far away from it. We interpret this problem as a hypothesis testing problem with a composite null hypothesis and a simple alternative hypothesis:

$$H_0 : P_{XY} \in \Pi \quad \text{vs} \quad H_1 : Q_{XY}, \tag{4.1}$$

where $\Pi = \{P_{XY} \in \mathcal{P}_{XY} : ||P_{XY} - Q_{XY}||_1 \geq \lambda\}$ and $\lambda$ is some fixed positive number. The model is shown in Figure 4.1. As discussed in the Section 1.4, the other formulation with

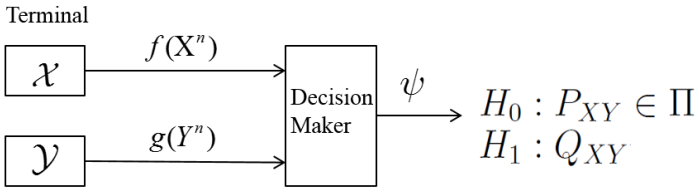



Figure 4.1: Model

simple null and composite alternative hypothesis can be analyzed following existing work and hence is not discussed in this chapter.

In a typical identity testing problem, one determines which hypothesis is true under the assumption that $(X^n, Y^n)$ are fully available at the decision maker. In this chapter, we consider a setting in which terminal $\mathcal{X}$ only observes $X^n$ and terminal $\mathcal{Y}$ observes only $Y^n$. Terminals $\mathcal{X}$ and $\mathcal{Y}$ are allowed to send encoded messages to the decision maker. And the decision maker decides which hypothesis is true using the encoded messages directly, which

makes the problem more complex. We denote the system as $S_{XY}$ in the sequel.

More specifically, the system consists of 2 encoders $f$ and $g$, one at terminal $\mathcal{X}$ and the other one at terminal $\mathcal{Y}$, and one decision function $\psi$ at the decision maker. After observing the data sequence $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$, encoder $f$ or $g$ transforms the sequence $x^n$ or $y^n$ into a message $f(x^n)$ or $g(y^n)$ taking values from the message set $\mathcal{M}_n$ and $\mathcal{N}_n$

$$f : \mathcal{X}^n \to \mathcal{M}_n = \{1, 2, \ldots, M_n\}, \tag{4.2}$$

$$g : \mathcal{Y}^n \to \mathcal{N}_n = \{1, 2, \ldots, N_n\}, \tag{4.3}$$

with rate constraint:

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n \leq R_1, \tag{4.4}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log N_n \leq R_2. \tag{4.5}$$

Using the messages $f(X^n)$ and $g(Y^n)$, the decision maker will use the decision function $\psi$ to determine which hypothesis is true:

$$\psi : (\mathcal{M}_n, \mathcal{N}_n) \to \{H_0, H_1\}. \tag{4.6}$$

For any given decision function $\psi$, one can define the acceptance region as

$$\mathcal{A}_n = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \psi(f(x^n), g(y^n)) = H_0\}.$$

For any given $f$, $g$ and $\psi$, the type 1 error probability $\alpha_n$ and the type 2 error probability $\beta_n$ are defined as following:

$$\alpha_n = \sup_{P_{XY}\in\Pi} P^n_{XY}(\mathcal{A}_n^c), \text{ and } \beta_n = Q^n_{XY}(\mathcal{A}_n). \tag{4.7}$$

Based on the definition of the encoding functions $f$, $g$ and the decoding function $\psi$, we can define the type 2 error exponents under two types of constraints on the type 1 error probability similarily in Section 2.2.

## 4.3 Preliminaries

It will be clear in the sequel, in our schemes and analysis, we would like to cover $\Pi$ using regions with small area. To serve our purpose, we would like each region to have diminishing small area as $n$ increases, at the same time we would also like to control the growth of the number of regions needed to cover $\Pi$. The following particular way of covering $\Pi$ strikes a desirable balance for our purpose. In particular, we let $\Lambda_n(\mathcal{XY})$ be the set of all possible types of $(X^n, Y^n)$. First, choose the center points $t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})$, where $\Lambda_n^{\Pi}(\mathcal{XY}) = \Pi \cap \Lambda_n(\mathcal{XY})$. It is easy to get the $|\Lambda_n^{\Pi}(\mathcal{XY})| \leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|}$. Then define the $\zeta$-set around each center point

$$\mathcal{N}_{t_{XY}} \triangleq \left\{\tilde{t}_{XY} \in \mathcal{P}_{XY} : ||\tilde{t}_{XY} - t_{XY}||_1 \leq \zeta\right\}, \tag{4.8}$$

where $\zeta = \frac{1}{n}$. We can prove the following lemma.

**Lemma 4.1.** If we choose $t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})$ as the center points, and define the $\zeta$-set as (4.8), then we have

$$\bigcup_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})} \mathcal{N}_{t_{XY}} \supseteq \Pi. \tag{4.9}$$

*Proof.* Please see Appendix C.1. □

## 4.4 Identity Testing under Zero-rate Data Compression

In this section, we will characterize the error exponent of the type 2 error probability with zero-rate data compression.

As discussed in the Section 1.4, the identity testing under zero-rate data compression with constant-type constraint was studied in [34]. Hence, here we will focus on the identity testing problem with an exponential-type constraint on the type 1 error. As $H_0$ is a composite hypothesis, we provide a universal encoding and decoding scheme to establish a lower bound on the error exponent of the type 2 error. We further establish a matching upper bound and hence fully characterize the error exponent of the type 2 error for this scenario.

**Theorem 4.1.** Let $P_{XY} \in \Pi$ be arbitrary and $Q_{XY} > 0$. For zero-rate data compression in $S_{XY}$ and the type 1 error constraint (2.9), the error exponent satisfies

$$\sigma(0, 0, r) = \sigma_{opt} \tag{4.10}$$

in which

$$\sigma_{opt} \triangleq \inf_{P_{XY} \in \Pi} \min_{\tilde{P}_{XY} \in \mathcal{H}_r} D\left(\tilde{P}_{XY} || Q_{XY}\right) \tag{4.11}$$

with

$$\mathcal{H}_r = \left\{ \tilde{P}_{XY} : \tilde{P}_X = \hat{P}_X, \tilde{P}_Y = \hat{P}_Y \text{ for some } \hat{P}_{XY} \in \varphi_r \right\},$$

$$\varphi_r = \left\{ \hat{P}_{XY} : D(\hat{P}_{XY} || P_{XY}) \leq r \right\}.$$

*Proof.* **Achievability:**

We first show the achievability by providing a universal coding scheme.

*Step 1:Encoding.*

Divide the $(|\mathcal{X}| + |\mathcal{Y}|)$ dimensional unit cube into equal-sized $M_n \cdot N_n$ small cells with each edge of length $\kappa_n$ along the first $|\mathcal{X}|$ components, and each edge of length $\tau_n$ along the $|\mathcal{Y}|$ components, where

$$\kappa_n = M_n^{-1/|\mathcal{X}|}, \quad \tau_n = N_n^{-1/|\mathcal{Y}|},$$

in which

$$M_n \to \infty, N_n \to \infty, \tag{4.12}$$

but $\frac{1}{n}\log M_n \to 0$ and $\frac{1}{n}\log N_n \to 0$, as $n \to \infty$ (i.e., zero-rate compression for all two terminals).

Choose and fix a representative point in each cell for every set of variables $(\tilde{X}, \tilde{Y})$. Then in a given cell, we make its representative variable set $(\check{X}, \check{Y})$ correspond in such a way that$((\check{P}_X)_{x\in\mathcal{X}}, (\check{P}_Y)_{y\in\mathcal{Y}})$ is the representative point of $((\tilde{P}_X)_{x\in\mathcal{X}}, (\tilde{P}_Y)_{y\in\mathcal{Y}})$. For each terminal, after observing its sequence, it determines its type and then finds the index of the corresponding edge. Each terminal then sends the index to the decision maker. After receiving all the indexes, the decision maker can determine the cell index. Since we have assumed (4.12), we see that with any $\eta > 0$

$$|\tilde{P}_X - \check{P}_X| < \eta, \quad x \in \mathcal{X}, \tag{4.13}$$

$$|\tilde{P}_Y - \check{P}_Y| < \eta, \quad y \in \mathcal{Y}, \tag{4.14}$$

for sufficiently large $n$.

We note that the encoding scheme is universal and does not depend on the knowledge of $P_{XY}$.

*Step 2: Acceptance region and type 1 error analysis*

For the decision maker, it needs to design a universal acceptance region so that the type 1 error constraint is satisfied regardless of what the true value of $P_{XY}$ is. One can certainly design an individual acceptance region that satisfy the type 1 error constraint for each possible value of $P_{XY} \in \Pi$ using the approach in the simple hypothesis case, then take the union of these individual regions as the final acceptance region. This will clearly satisfy the type 1 error constraint regardless the true value of $P_{XY}$. This approach will work if there are a finite number of possible $P_{XY}$s. However, in our case, there are infinitely many possible $P_{XY}$s in

$\Pi$. This approach will lead to very loose performance bound. In the following, we design a new approach that will lead to performance bound matching with the converse bound to be presented below.

According to Lemma 4.1, we can choose $t_{XY}$ as the *center points* and design $\mathcal{N}_{t_{XY}}$ to cover the set $\Pi$. In our approach, we first design individual acceptance region $\mathcal{A}_n^{t_{XY}}$ for each center point $t_{XY}$ such that type 1 error probability constraint is satisfied. Then, we show that the acceptance region $\mathcal{A}_n^{t_{XY}}$ can be applied to each distribution $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$ to satisfy the type 1 error probability constraint. Hence, by taking union of these acceptance regions (the number of regions grows in polynomial order of $n$) , we can find $\mathcal{A}_n$ such that the type 1 error probability constraint is satisfied and a tight lower bound on the type 2 error exponent can be imposed.

*Step 2.1: Acceptance region for $t_{XY}$.*

Set

$$f_{t_{XY}}\left(\tilde{X},\tilde{Y}\right) = \min_{\substack{\hat{P}_{XY} \\ \hat{P}_X = \tilde{P}_X \\ \hat{P}_Y = \tilde{P}_Y}} D\left(\hat{P}_{XY}||t_{XY}\right)$$

which is continuous in $\left((\tilde{P}_X)_{x\in\mathcal{X}}, (\tilde{P}_Y)_{y\in\mathcal{Y}}\right)$. Furthermore, the continuity of $f_{t_{XY}}\left(\tilde{X},\tilde{Y}\right)$ in $\left(\tilde{X},\tilde{Y}\right)$ yields

$$\left| f_{t_{XY}}\left(\tilde{X},\tilde{Y}\right) - f_{t_{XY}}\left(\check{X},\check{Y}\right)\right| < \eta. \tag{4.15}$$

Denoting by $(\check{X}^{(n)}(x^n), \check{Y}^{(n)}(y^n))$ the representative point of $(X^{(n)}(x^n), Y^{(n)}(y^n))$ where $X^{(n)}(x^n)$ and $Y^{(n)}(y^n)$ are the type variables of $x^n \in \mathcal{X}^n$, and $y^n \in \mathcal{Y}^n$ respectively, we set an acceptance region based on $t_{XY}$:

$$\mathcal{A}_n^{t_{XY}} = \left\{(x^n,y^n) : f_{t_{XY}}\left(\check{X}^{(n)}(x^n), \check{Y}^{(n)}(y^n)\right) \le r + 3\eta\right\}.$$

For any $\rho > 0$ set

$$\xi_\rho^{t_{XY}} = \left\{(x^n,y^n) : f_{t_{XY}}(\tilde{X},\tilde{Y}) \le \rho\right\};$$

then in view of (4.15) it is clear that

$$\xi_{r+2\eta}^{t_{XY}} \subset \mathcal{A}_n^{t_{XY}} \subset \xi_{r+4\eta}^{t_{XY}}. \tag{4.16}$$

*Step 2.2: Error analysis for $t_{XY}$.*

It is easy to see that $(x^n, y^n) \in \xi_{r+2\eta}^{t_{XY}}$ if $(x^n, y^n) \in S_{r+2\eta}^n(t_{XY})$, that is $S_{r+2\eta}^n(t_{XY}) \subset \xi_{r+2\eta}^{t_{XY}}$, which yields

$$1 - \alpha_n^{t_{XY}} = t_{XY}\left(X^n Y^n \in \mathcal{A}_n^{t_{XY}}\right) \geq 1 - \exp(-nr)$$

for $n$ large enough. Hence, the constraint (2.9) is satisfied.

*Step 2.3: Error analysis for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}} \setminus t_{XY}$.*

First, we have the following inequality:

$$f_{t_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right) \overset{(a)}{\leq} f_{\tilde{t}_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right) + \delta(\zeta) \tag{4.17}$$

$$\leq D\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)||\tilde{t}_{XY}\right) + \delta(\zeta), \tag{4.18}$$

where $(a)$ is true due to the continuity of the divergence, shown in Appendix C.2. And $\delta(\zeta)$ is of the same order of $\zeta$, i.e. $\delta(\zeta) \sim O(\frac{1}{n})$.

Then, it is easy to get that $(x^n, y^n) \in \xi_{r+2\eta}^{t_{XY}}$ if $(x^n, y^n) \in S_{r+2\eta-\delta(\zeta)}^n(\tilde{t}_{XY})$, that is $S_{r+2\eta-\delta(\zeta)}^n(\tilde{t}_{XY}) \subset \xi_{r+2\eta}^{t_{XY}}$, which yields

$$1 - \alpha_n^{\tilde{t}_{XY}} = \tilde{t}_{XY}\left(X^n Y^n \in \mathcal{A}_n^{t_{XY}}\right) \geq 1 - \exp(-nr)$$

for $n$ large enough and $\frac{1}{\eta} \sim o(n)$. Hence, the constraint (2.9) is satisfied.

*Step 3: Acceptance region for all $P_{XY} \in \Pi$.*

According to the results from *Step 2.1* to *step 2.3*, we can set the acceptance region as

$$\mathcal{A}_n = \bigcup_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{X}\mathcal{Y})} \mathcal{A}_n^{t_{XY}}. \tag{4.19}$$

It is obvious that

$$\alpha_n = \sup_{P_{XY} \in \Pi} P_{XY}^n(\mathcal{A}_n^c) \leq \sup_{P_{XY} \in \Pi} P_{XY}^n \left( \bigcap_{t_{XY} \in \Lambda_n^{\Pi} \mathcal{X}\mathcal{Y})} (\mathcal{A}_n^{t_{XY}})^c \right) \leq \exp(-nr).$$

*Step 4: Type 2 error exponent analysis.*

From the second inclusion in (4.16), we have

$$\begin{aligned}
Q_{XY}^n(\mathcal{A}_n) &\leq Q_{XY}^n \left( \bigcup_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{X}\mathcal{Y})} \xi_{r+4\eta}^{t_{XY}} \right) \\
&\leq \sum_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{X}\mathcal{Y})} Q_{XY}^n \left( \xi_{r+4\eta}^{t_{XY}} \right) \\
&\leq \sum_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{X}\mathcal{Y})} \sum_{\substack{\tilde{P}_{XY} \\ f_{t_{XY}}(\tilde{X}, \tilde{Y}) \leq r + 4\eta}} \exp\left( -nD\left( \tilde{P}_{XY} || Q_{XY} \right) \right) \\
&\leq \left( (n+1)^{2|\mathcal{X}||\mathcal{Y}|} \right) \cdot \sup_{P_{XY} \in \Pi} \max_{\substack{\tilde{P}_{XY} \\ f_{p_{XY}}(\tilde{X}, \tilde{Y}) \leq r + 4\eta}} \exp\left( -nD\left( \tilde{P}_{XY} || Q_{XY} \right) \right) \\
&\leq \left( (n+1)^{2|\mathcal{X}||\mathcal{Y}|} \right) \cdot \exp\left( -n \left( \inf_{P_{XY} \in \Pi} \min_{\substack{\tilde{P}_{XY} \\ f_{P_{XY}}(\tilde{X}, \tilde{Y}) \leq r + 4\eta}} D\left( \tilde{P}_{XY} || Q_{XY} \right) \right) \right).
\end{aligned}$$

Therefore

$$\beta_n = Q_{XY}^n(\mathcal{A}_n) \leq \left( (n+1)^{2|\mathcal{X}||\mathcal{Y}|} \right) \cdot \exp\left( -n \left( \inf_{P_{XY} \in \Pi} \min_{\substack{\tilde{P}_{XY} \\ f_{P_{XY}}(\tilde{X}, \tilde{Y}) \leq r + 4\eta}} D\left( \tilde{P}_{XY} || Q_{XY} \right) \right) \right).$$

Thus

$$\sigma(0,0,r) \geq \inf_{P_{XY}\in\Pi} \min_{\tilde{P}_{XY}\in\mathcal{H}_{r+4\eta}} D\left(\tilde{P}_{XY}||Q_{XY}\right),$$

which establishes the lower bound in Theorem 4.1 if we let $\eta \to 0$.

**Converse**

Here, we establish an upper bound on the error exponent that any scheme can achieve. Following the similar strategy as in [15], we can first convert a problem with the exponential-type constraint to a corresponding problem with the constant-type constraint. We can then obtain an upper bound on the error exponent using the results in [34] for the constant-type constraint. To invoke "Blowing up lemma" [2] in the proof of in [34], the positive condition $Q_{XY} > 0$ is needed.

Let $\mathcal{A}_n$ be an arbitrary acceptance region such that

$$\alpha_n \leq \exp(-nr), \quad r > 0 \tag{4.20}$$

where

$$\alpha_n = \sup_{P_{XY}\in\Pi} P_{XY}^n(\mathcal{A}_n^c). \tag{4.21}$$

Equations (4.20) and (4.21) imply that

$$\inf_{P_{XY}\in\Pi} P_{XY}^n(\mathcal{A}_n) \geq 1 - \exp[-n(r-\gamma)] \quad \forall n \geq n_0 \tag{4.22}$$

where $\gamma > 0$ is an arbitrarily small constant.

Next, for each $P_{XY} \in \Pi$, select an arbitrary "internal point" $P_{X_0Y_0} \in \varphi_r$, where $\varphi_r$ is specified in (4.1). Then clearly

$$D(P_{X_0Y_0}||P_{XY}) < r. \tag{4.23}$$

Define

$$\hat{T}_n(\delta) = \{\text{joint types } \hat{P}_n \text{ on } \mathcal{X}^n \times \mathcal{Y}^n : D(\hat{P}_n||P_{X_0Y_0}) < \delta\} \tag{4.24}$$

where $\delta > 0$ is an arbitrary constant. Then, in view of (4.23) and the uniform continuity of the divergence, for all $\hat{P}_n \in \hat{T}_n(\delta)$ it holds that

$$c_n \equiv D(\hat{P}_n||P_{XY}) < r - 2\gamma, \tag{4.25}$$

provided that we take $\gamma > 0$ and $\delta > 0$ sufficiently small. Consequently, according to Lemma 2.2, we have

$$|\mathcal{A}_n(\hat{P}_n)| \geq (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\gamma))|S_0(\hat{P}_n)| \tag{4.26}$$

for all $\hat{P}_n \in \hat{T}_n(\delta)$. Now we define the set

$$T_n(\delta) = \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : X^{(n)}Y^{(n)} \in \hat{T}_n(\delta)\right\} \tag{4.27}$$

and consider an i.i.d. random sequence of length $n$ generated according to the probability distribution $P_{X_0Y_0}$. Then, from (4.26), we have

$$\begin{aligned}
P^n_{X_0Y_0}(\mathcal{A}_n) &\geq P^n_{X_0Y_0}(\mathcal{A}_n \cap T_n(\delta)) \\
&= \sum_{\hat{P}_n \in \hat{T}_n(\delta)} P^n_{X_0Y_0}(\mathcal{A}_n \cap S_0(\hat{P}_n)) \\
&= \sum_{\hat{P}_n \in \hat{T}_n(\delta)} P^n_{X_0Y_0}(\mathcal{A}_n(\hat{P}_n)) \\
&\geq (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\gamma)) \cdot \sum_{\hat{P}_n \in \hat{T}_n(\delta)} P^n_{X_0Y_0}(S_0(\hat{P}_n)) \\
&= (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\gamma)) P_{X_0^{(n)}Y_0^{(n)}}(\hat{T}_n(\delta)) \\
&\geq (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\gamma)) \cdot (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\delta)).
\end{aligned} \tag{4.28}$$

Now consider the zero-rate $(R_1 = 0, R_2 = 0, R \geq 0)$ hypothesis testing problem with

$$H_0 : P_{X_0Y_0} \in \Pi \quad \text{vs} \quad H_1 : Q_{XY}. \tag{4.29}$$

Then, for this hypothesis testing problem, if we use the same acceptance region $\mathcal{A}_n$ as above, the type 1 error probability

$$\begin{aligned}
\alpha_n^{(0)} &= 1 - \inf_{P_{XY} \in \Pi} P_{X_0Y_0}^n(\mathcal{A}_n) \\
&\leq 1 - (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\gamma)) \cdot (1 - (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\delta)) \\
&\leq \epsilon.
\end{aligned}$$

Hence, for the hypothesis testing problem (4.29), the acceptance region $\mathcal{A}_n$ satisfies the constant-type type 1 error probability constraint.

As a special case of Theorem 4 in [34], we know that the type 2 error exponent

$$\theta(0, 0, \epsilon) \leq \inf_{P_{XY} \in \Pi} \min_{\tilde{P}_{XY} \in \mathcal{L}_0} D\left(\tilde{P}_{XY} || Q_{XY}\right), \tag{4.30}$$

where

$$\mathcal{L}_0 = \left\{ \tilde{P}_{XY} : \tilde{P}_X = P_{X_0}, \tilde{P}_Y = P_{Y_0} \right\}.$$

On the other hand, we note that $P_{X_0Y_0}$ was arbitrary as far as condition (4.23) is satisfied. Therefore, in the light of the definition of $\mathcal{H}_r$, we see that the infimum of the right-hand side in (4.30) over all possible internal points $P_{X_0Y_0}$ satisfying (4.23) coincides with

$$\inf_{P_{XY} \in \Pi} \min_{\tilde{P}_{XY} \in \mathcal{H}_r} D\left(\tilde{P}_{XY} || Q_{XY}\right).$$

Thus (4.30) reduces to

$$\sigma(0,0,r) \leq \inf_{P_{XY}\in\Pi} \min_{\tilde{P}_{XY}\in\mathcal{H}_r} D\left(\tilde{P}_{XY}||Q_{XY}\right).$$

□

## 4.5 Identity Testing under Positive Rate Compression

In this section, we investigate the identity testing under positive rate compression constraints (4.4) and (4.5). We first establish a lower bound on the type 2 error exponent under constant-type constraint (2.8) for general PMF. We then investigate a special case of identity testing against independence. For this special case, we can establish matching lower bounds and upper bounds under constant-type constraint (2.8). Finally, we provide a lower bound on the type 2 error exponent under the exponential-type constraint (2.9).

### 4.5.1 Results with Constant-type Constraint

As $H_0$ is composite, we need a universal encoding and decoding scheme. The universal encoding method is similar with that in [14], but a different decoding scheme is needed. In [14], their goal was to estimate the unknown parameter in the joint distribution of $(X, Y)$. Therefore, the authors utilized a complex decoding method to get *maximum-likelihood* estimators. They first use the *minimum-entropy* decoding method to recover the massages sent by each terminal, and then derived an asymptotic distribution for the joint types of sequences $(X^n, Y^n)$ and messages $(U^n, V^n)$, and finally used score vector of the observable type to find the *maximum-likelihood* estimators. Following their schemes with necessary changes, we can calculate the two types of error probabilities by utilizing the asymptotic distribution of the joint types of sequences and messages. However, due to complex derivations involved, the obtained type 2 error exponent formulator is very complicated and does

not provide meaningful insights. Furthermore, our goal is to decide whether the joint distribution of $(X, Y)$ is $Q_{XY}$ or not, which makes the estimation the unknown parameter not necessary. Hence, we can simplify our decoder design and performance analysis. In the following method, we will show a simple decoding method using the result in Lemma 4.1.

We first focus on the case with general PMF stated in (4.1).

Let $\mathcal{U}$ be an arbitrary finite set. For each distribution $P_X$ on $\mathcal{X}$, let $\omega(\cdot|\cdot; P_X)$ be any stochastic mapping from $\mathcal{X}$ to $\mathcal{U}$, i.e. $\omega(u|x; P_X)$ be the probability of $u \in \mathcal{U}$ given $x \in \mathcal{X}$. Similarly, for each distribution $P_Y$ on $\mathcal{Y}$, let $\varrho(\cdot|\cdot; P_Y)$ be any stochastic mapping from $\mathcal{Y}$ to $\mathcal{V}$, i.e. $\varrho(v|y; P_Y)$ be the probability of $v \in \mathcal{V}$ given $y \in \mathcal{Y}$.

**Theorem 4.2.** For $R_1 \geq 0$, $R_2 \geq 0$, we have

$$\theta(R_1, R_2, \epsilon) \geq \inf_{P_{XY} \in \Pi} \max_{(\omega,\varrho) \in \varphi_{P_{XY}}} \min_{\tilde{P}_{UVXY} \in \xi_{P_{XY}}} D\left(\tilde{P}_{UVXY} || Q_{UVXY}\right), \tag{4.31}$$

where

$$\begin{aligned} \varphi_{P_{XY}} = \Big\{ (\omega, \varrho) : & R_1' \geq I(X;U), R_2' \geq I(Y;V), \\ & R_1' - R_1 \leq I(U;V), \\ & R_2' - R_2 \leq I(U;V), \\ & R_1' - R_1 + R_2' - R_2 \leq I(U;V) \\ & P_{U|X} = \omega(u|x; P_X), P_{V|Y} = \varrho(v|y; P_Y), \\ & U \leftrightarrow X \leftrightarrow Y \leftrightarrow V \Big\}, \end{aligned} \tag{4.32}$$

and

$$\xi_{P_{XY}} = \left\{ \tilde{P}_{UVXY} : \tilde{P}_{UX} = P_{UX}, \tilde{P}_{VY} = P_{VY}, \tilde{P}_{UV} = P_{UV} \right\}. \tag{4.33}$$

Note that $\varphi_{P_{XY}}$ denotes the set of $(\omega, \varrho)$ when the distribution of $(X, Y)$ is $P_{XY}$, and similar

for $\xi_{P_{XY}}$.

*Proof.* **Step 1: Encoding.**

In this step, we show a universal encoding scheme.

*Step 1.1: Codebook generation.*

For each $t_X \in \Lambda_n(\mathcal{X})$, i.e. the type of $X^n$, generate $2^{nR_1'}$ sequences $u_{s_1}^n,\ s_1 \in \{1, \cdots, 2^{nR_1'}\}$ randomly and independently according to $P_U^{t_X}(u) = \sum_{x\in\mathcal{X}} \omega(u|x;t_X)t_X(x)$ for some $\omega(u|x;t_X) \in \varphi_{t_{XY}}$. Then randomly assign to every sequence a bin index $m_1 \in \{1, \cdots, 2^{nR_1}\}$, this bin is denoted as $\mathcal{B}(m_1)$. Similarly, for each $t_Y \in \Lambda_n(\mathcal{Y})$, i.e. the type of $Y^n$, generate $2^{nR_2'}$ sequences $v_{s_2}^n,\ s_2 \in \{1, \cdots, 2^{nR_2'}\}$ randomly and independently according to $P_V^{t_Y}(v) = \sum_{y\in\mathcal{Y}} \varrho(v|y;t_Y)t_Y(y)$ for some $\varrho(v|y;t_Y) \in \varphi_{t_{XY}}$. Then randomly assign to every sequence a bin index $m_2 \in \{1, \cdots, 2^{nR_2}\}$, this bin is denoted as $\mathcal{B}(m_2)$.

*Step 1.2: Encoding.*

Given a sequence $x^n$, terminal $\mathcal{X}$ finds it type and chooses $u_{s_{10}}^n$ generated according to $P_U^{t_X}$, that jointly typical with $x^n$. Then terminals $\mathcal{X}$ sends the bin index $m_1$ to the decision maker. Similarly, given a sequence $y^n$, terminal $\mathcal{Y}$ finds it type and chooses $v_{s_{20}}^n$ generated according to $P_V^{t_Y}$, that jointly typical with $y^n$. Then terminals $\mathcal{Y}$ sends the bin index $m_2$ to the decision maker.

**Step 2: Testing.**

Upon receiving $m_1$ and $m_2$, the decision maker needs to design a universal acceptance region so that the type 1 error constraint is satisfied. Based on Lemma 4.1, for each center point $t_{XY}$, we set an individual acceptance region $\mathcal{A}_n^{t_{XY}}$ such that the type 1 error probability constraint under $t_{XY}$ is satisfied. Then we show that using the same acceptance region $\mathcal{A}_n^{t_{XY}}$, the type 1 error probability constraint can also be satisfied for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$. Finally, we make the final acceptance region as the union of a finite number of acceptance regions $\mathcal{A}_n^{t_{XY}}$ and show that the type 1 error probability constraint under $H_0$ is satisfied.

*Step 2.1: Acceptance region for $t_{XY}$.*

Set

$$\mathcal{A}_n^{t_{XY}} = \left\{(m_1, m_2) : \exists \text{ unique } s_1 \in \mathcal{B}(m_1), s_2 \in \mathcal{B}(m_2) : (u_{s_1}^n, v_{s_2}^n) \in T_\epsilon^n(t_{UV})\right\}, \quad (4.34)$$

where the joint typicality is according to

$$t_{UV}(uv) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \omega(u|x; t_X) \cdot \varrho(v|y; t_Y) \cdot t_{XY}. \quad (4.35)$$

The decision maker chooses $\hat{H} \neq H_0$ if and only if one or more of the following events occur:

$$\begin{aligned}
\varepsilon_1 &= \left\{(U_{s_1}^n, X^n) \notin T_{\epsilon''}^n \text{ for all } s_1 \in [1 : 2^{nR_1'}]\right\}, \\
\varepsilon_2 &= \left\{(V_{s_2}^n, Y^n) \notin T_{\epsilon''}^n \text{ for all } s_2 \in [1 : 2^{nR_2'}]\right\}, \\
\varepsilon_3 &= \left\{\exists \text{ none or more than one } (s_1, s_2) \in \mathcal{B}(m_1) \times \mathcal{B}(m_2) : (U_{s_1}^n, V_{s_2}^n) \notin T_\epsilon^n\right\}.
\end{aligned}$$

Hence, $\mathcal{A}_n = (\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3)^c$.

To analyze the type 1 error probability, we have

$$\begin{aligned}
\alpha_n &= t_{XY}^n(\mathcal{A}_n^c) \\
&= t_{XY}^n(\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3) \\
&\leq t_{XY}^n(\varepsilon_1) + t_{XY}^n(\varepsilon_2) + t_{XY}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3).
\end{aligned}$$

*Step 2.2: Type 1 error analysis for $t_{XY}$.*

When the true $P_{XY} = t_{XY}$, we can bound each term.

(1) By the covering lemma [11, Section 3.7], $t_{XY}^n(\varepsilon_1) \to 0$ as $n \to \infty$ if $R_1' \geq I(U; X) + \delta(\epsilon)$ and $t_{XY}^n(\varepsilon_2) \to 0$ as $n \to \infty$ if $R_2' \geq I(V; Y) + \delta(\epsilon)$.

(2) To bound the last term, we first show that there exists $(S_1, S_2) \in \mathcal{B}(M_1) \times \mathcal{B}(M_2)$ such

that $(U_{S_1}^n, V_{S_2}^n) \notin T_\epsilon^n$; then we show that such $(S_1, S_2)$ are unique when

$$R_1' - R_1 \leq I(U;V) - \delta(\epsilon), \tag{4.36}$$

$$R_2' - R_2 \leq I(U;V) - \delta(\epsilon), \tag{4.37}$$

$$R_1' - R_1 + R_2' - R_2 \leq I(U;V) - \delta(\epsilon). \tag{4.38}$$

For the existence of $(S_1, S_2)$, we first show that

$$(V_{S_2}^n, X^n, Y^n) \in T_{\epsilon'}^n,\ S_2 \in \mathcal{B}(M_2); \tag{4.39}$$

then we show that for $(S_1, S_2) \in \mathcal{B}(M_1) \times \mathcal{B}(M_2)$

$$(U_{S_1}^n, V_{S_2}^n, X^n, Y^n) \in T_\epsilon^n. \tag{4.40}$$

Since $V \leftrightarrow Y \leftrightarrow X$, we can prove that $(V_{S_2}^n, X^n, Y^n) \in T_{\epsilon'}^n,\ S_2 \in \mathcal{B}(M_2)$ by the Markov lemma [11, Section 12.1]. We can also prove that $(U_{S_1}^n, V_{S_2}^n, X^n, Y^n) \in T_\epsilon^n$ for $(S_1, S_2) \in \mathcal{B}(M_1) \times \mathcal{B}(M_2)$ using Markov lemma [11, Section 12.1] since $U \leftrightarrow X \leftrightarrow Y \leftrightarrow V$, details are shown in Appendix C.3.

To show $(S_1, S_2)$ are unique and $(S_1, S_2) = (S_{10}, S_{20})$, we have three situations:

$$\mathcal{S}_1 = \left\{(U_{S_1}^n, V^n(S_{20})) \in T_\epsilon^n,\ S_1 \neq S_{10} \text{ for some } S_1 \in \mathcal{B}(M_1)\right\}, \tag{4.41}$$

$$\mathcal{S}_2 = \left\{(U^n(S_{10}), V^n(S_2)) \in T_\epsilon^n,\ S_2 \neq S_{20} \text{ for some } S_2 \in \mathcal{B}(M_2)\right\}, \tag{4.42}$$

$$\begin{aligned}\mathcal{S}_3 = \big\{&(U_{S_1}^n, V^n(S_2)) \in T_\epsilon^n,\ S_1 \neq S_{10},\ S_2 \neq S_{20},\\ &\text{for some } (S_1, S_2) \in \mathcal{B}(M_1) \times \mathcal{B}(M_2)\big\}.\end{aligned} \tag{4.43}$$

The probability of a particular $U_{S_1}^n,\ S_1 \neq S_{10}$ that is jointly typical with $V^n(S_{20})$ can be bounded as

$$t_{XY}\left((U_{S_1}^n, V^n(S_{20})) \in T_\epsilon^n\right) \leq 2^{-n(I(U;V)-\delta(\epsilon))}. \tag{4.44}$$

Hence the error probability is

$$\begin{aligned} t_{XY}(\mathcal{S}_1) &\leq \sum_{S_1 \in \mathcal{B}(M_1), S_1 \neq S_{10}} t_{XY}((U_{S_1}^n, V^n(S_{20})) \in T_\epsilon^n) \\ &\leq 2^{n(R_1' - R_1)} 2^{-n(I(U;V)-\delta(\epsilon))}, \end{aligned} \tag{4.45}$$

which tends to 0 as $n \to \infty$ if $R_1' - R_1 \leq I(U;V)$.

In an analogous manner, we have

$$t_{XY}(\mathcal{S}_2) \leq 2^{n(R_2' - R_2)} 2^{-n(I(U;V)-\delta(\epsilon))}, \tag{4.46}$$

and

$$t_{XY}(\mathcal{S}_3) \leq 2^{n(R_1' - R_1 + R_2' - R_2)} 2^{-n(I(U;V)-\delta(\epsilon))}. \tag{4.47}$$

Hence, we have shown such $(s_1, s_2)$ are unique when

$$\begin{aligned} R_1' - R_1 &\leq I(U;V) - \delta(\epsilon), \\ R_2' - R_2 &\leq I(U;V) - \delta(\epsilon), \\ R_1' - R_1 + R_2' - R_2 &\leq I(U;V) - \delta(\epsilon). \end{aligned}$$

*Step 2.3: Type 1 error analysis for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}} \setminus t_{XY}$.*

As for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$, $||\tilde{t}_{XY} - t_{XY}||_1 \leq \frac{1}{n}$, $X^n Y^n$ generated according to $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$, has the type $t_X$ and $t_Y$. Hence, the $U^n$ and $V^n$ sequences are the same as what generated in *Stpe 2.3*. Due to the continuity of mutual information theory and entropy, we can show that if $R_1' \geq I(U;X) + \delta(\epsilon)$ and $R_2' \geq I(V;Y) + \delta(\epsilon)$, the constraint on the type 1 error probability is satisfied using the acceptance region $\mathcal{A}_n^{t_{XY}}$.

First we show the probability that $\tilde{t}_{XY}(X^nU^n \in T^n_{\epsilon'''}(t_{UX})) = t_{XY}(X^nU^n \in T^n_{\epsilon'''}(t_{UX}))$.

$$\begin{aligned}\tilde{t}_{XY}(X^nU^n \in T^n_{\epsilon'''}(t_{UX})) &= \sum_{x^n \in T^n_{\epsilon'''}(t_X)} \tilde{t}^n_X(x^n) \sum_{u^n \in T^n_{\epsilon'''}(U|x^n)} \tilde{t}^n_U(u^n) \\ &= 2^{n(H(X)+H(U|X)-D(t_X||\tilde{t}_X)-H(X)-H(U)+\delta(\epsilon'''))} \\ &= 2^{-n(I(U;X)-D(t_X||\tilde{t}_X)+\delta(\epsilon'''))} \\ &\overset{(a)}{=} 2^{-n(I(U;X)+\delta(\epsilon))},\end{aligned}$$

where $(a)$ is true due to $D(t_X||\tilde{t}_X) \to 0$ with order $O(\frac{1}{n})$ given $||\tilde{t}_X - t_X|| \leq \frac{1}{n}$.

Hence, we can show that if $R'_1 \geq I(U;X)+\delta(\epsilon)$, there exists at least one $u^n$ that is jointly typical with $x^n$ according to $t_{UX}$. Similarly, we can prove that if $R'_2 \geq I(V;Y)+\delta(\epsilon)$, there exists at least one $v^n$ that is jointly typical with $y^n$ according to $t_{VY}$.

Following similar steps, we can prove that there exists unique $(S_1, S_2) \in \mathcal{B}(M_1)\times\mathcal{B}(M_2)$ such that $(U^n_{S_1}, V^n_{S_2}, X^n, Y^n) \in T^n_\epsilon$. Therefore, the type 1 error probability constraint is satisfied.

*Step 2.4: Acceptance region for all $P_{XY} \in \Pi$.*

According to the results in *Step 2.2* and *Step 2.3*, we can set the acceptance region as

$$\mathcal{A}_n = \bigcup_{t_{XY} \in \Lambda^{\Pi}_n(\mathcal{XY})} \mathcal{A}^{t_{XY}}_n \tag{4.48}$$

It is obvious that

$$\alpha_n = \sup_{P_{XY}\in\Pi} P^n_{XY}(\mathcal{A}^c_n) \leq \epsilon.$$

*Step 2.5: The type 2 error exponent.*

$$Q^n_{XY}(\mathcal{A}_n) = Q^n_{XY}\left(\bigcup_{t_{XY} \in \Lambda^{\Pi}_n(\mathcal{XY})} \mathcal{A}^{t_{XY}}_n\right) \tag{4.49}$$

$$\leq \sum_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})} Q_{XY}^n(\mathcal{A}_n^{t_{XY}}) \tag{4.50}$$

$$\overset{(a)}{=} \sum_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})} \exp\left(-n\left(\max_{(\omega,\varrho)\in\varphi_{t_{XY}}} \min_{\tilde{P}_{UVXY}} D\left(\tilde{P}_{UVXY}||Q_{UVXY}\right)\right)\right) \tag{4.51}$$

$$\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})} \exp\left(-n\left(\max_{(\omega,\varrho)\in\varphi_{t_{XY}}} \min_{\tilde{P}_{UVXY}} D\left(\tilde{P}_{UVXY}||Q_{UVXY}\right)\right)\right) \tag{4.52}$$

$$\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\left(\inf_{P_{XY}\in\Pi} \max_{(\omega,\varrho)\in\varphi_{P_{XY}}} \min_{\tilde{P}_{UVXY}} D\left(\tilde{P}_{UVXY}||Q_{UVXY}\right)\right)\right), \tag{4.53}$$

where $(a)$ is the result in [13].

Hence, we obtain a lower bound on the type 2 error exponent.

□

We now focus on the special case of testing against independence, for which the hypotheses are

$$H_0 : P_{XY} \in \Pi^{\perp} \qquad \text{vs} \qquad H_1 : Q_Y Q_Y, \tag{4.54}$$

where $\Pi^{\perp} = \Pi \cap \{P_{XY} : P_X = Q_X, P_Y = Q_Y\}$. This special hypothesis has a two-fold meaning: whether $(X, Y)$ are independent or not and whether the joint distribution of $(X, Y)$ is $Q_X Q_Y$ or not. Due to the fact that the marginal distribution in this special case is the same for both hypotheses, we can simplify our encoding/decoding scheme and derive a matching upper bound on the type 2 error exponent, which allows us to fully characterize the optimal type 2 error exponent.

**Theorem 4.3.** For $R_1 \geq 0$, $R_2 \geq 0$, we have

$$\theta(R_1, R_2, \epsilon) = \inf_{P_{XY}\in\Pi^{\perp}} \max_{UV\in\varphi_{P_{XY}}} I(U;V). \tag{4.55}$$

where

$$\varphi_{P_{XY}} = \{UV : I(U;X) \leq R_1,\ I(V;Y) \leq R_2, U \leftrightarrow X \leftrightarrow Y \leftrightarrow V\}. \tag{4.56}$$

*Proof.* **Achievability:**

In the following, $\epsilon > \epsilon' > \epsilon'' > \epsilon'''$ are given small numbers.

**Step 1: Encoding.**

*Step 1.1: Codebook generation.*

Since all $P_{XY} \in \Pi^{\perp}$ have the same marginal distribution, $\omega(u|x; P_X)$ can be simplified to $P_{U|X}$ and $\varrho(v|y; P_Y)$ can be simplified to $P_{V|Y}$. Fix $P_{U|X}$, generate $2^{nR_1}$ sequences $u_{m_1}^n$, $m_1 \in \{1, \cdots, 2^{nR_1}\}$ randomly and independently according to $P_U(u) = \sum_{x \in \mathcal{X}} P_{U|X}(u|x)P_X(x)$. Similarly, fixed $P_{V|Y}$, generate $N(t_Y) = 2^{nR_2}$ sequences $v_{m_2}^n$, $m_2 \in \{1, \cdots, 2^{nR_2}\}$ randomly and independently according to $P_V(v) = \sum_{y \in \mathcal{Y}} P_{V|Y}(v|y)P_Y(y)$. These sequences constitute the codebook $c$, which is revealed to all terminals. We use $\mathcal{C}$ to denote the set of all possible codebooks.

*Step 1.2: Encoding.*

Given a sequence $x^n$, terminal $\mathcal{X}$ finds it type and chooses $u_{m_1}^n$, generated according to $P_U$, that jointly typical with $x^n$. Then terminals $\mathcal{X}$ send $m_1$ to the decision maker. Similarly, given a sequence $y^n$, terminal $\mathcal{Y}$ finds it type and chooses $v_{m_2}^n$, generated according to $P_V$, that jointly typical with $y^n$. Then terminals $\mathcal{Y}$ send $m_2$ to the decision maker.

**Step 2: Testing.**

Upon receiving $m_1$ and $m_2$, the decision maker needs to design a universal acceptance region so that the type 1 error constraint is satisfied. Similarly to the scheme for the general PMF case, we first set an individual acceptance region $\mathcal{A}_n^{t_{XY}}$ such that the type 1 error probability constraint under $t_{XY}$ is satisfied for each center point $t_{XY}$. Then we show that using the same acceptance region $\mathcal{A}_n^{t_{XY}}$, the type 1 error probability constraint can also be satisfied for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$. Finally, we make the final acceptance region as the union of a finite number of acceptance regions $\mathcal{A}_n^{t_{XY}}$ and show that the type 1 error probability constraint under $H_0$ is satisfied.

*Step 2.1: Acceptance region based on each center point $t_{XY}$.* Set

$$\mathcal{A}_n^{t_{XY}} = \left\{(m_1, m_2) : (u_{m_1}^n, v_{m_2}^n) \in T_\epsilon^n(t_{UV})\right\}, \tag{4.57}$$

where the joint typicality is according to

$$t_{UV}(uv) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_{U|X}(u|x) \cdot P_{V|Y}(v|y) \cdot t_{XY}(xy). \tag{4.58}$$

The decision maker chooses $\hat{H} \neq H_0$ if and only if one or more of the following events occur:

$$\begin{aligned}
\varepsilon_1 &= \left\{(U_{M_1}^n, X^n) \notin T_{\epsilon'''}^n \text{ for all } M_1 \in [1 : 2^{nR_1}]\right\}, \\
\varepsilon_2 &= \left\{(V_{M_2}^n, Y^n) \notin T_{\epsilon'''}^n \text{ for all } M_2 \in [1 : 2^{nR_2}]\right\}, \\
\varepsilon_3 &= \left\{(U_{M_1}^n, V_{M_2}^n) \notin T_\epsilon^n\right\}.
\end{aligned}$$

Hence, $\mathcal{A}_n = (\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3)^c$.

Using the definition in (2.37) and (2.38), we will then argue that there exists a particular codebook $c^*$ that has the desired properties.

To analyze the type 1 error probability, we have

$$\begin{aligned}
\mathbb{E}\{\alpha_{nc}\} &= t_{XY}^n(\mathcal{A}_n^c) \\
&= t_{XY}^n(\varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3) \\
&\leq t_{XY}^n(\varepsilon_1) + t_{XY}^n(\varepsilon_2) + t_{XY}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3).
\end{aligned}$$

*Step 2.2: Type 1 error analysis for $t_{XY}$*

When the true $P_{XY} = t_{XY}$, we can bound each term.

(1) By the covering lemma [11, Section 3.7], $t_{XY}^n(\varepsilon_1) \to 0$ as $n \to \infty$ if $R_1 \geq I(U;X) + \delta(\epsilon)$ and $t_{XY}^n(\varepsilon_2) \to 0$ as $n \to \infty$ if $R_2 \geq I(V;Y) + \delta(\epsilon)$.

(2) To bound the last term, we use a version of the Markov lemma [11, Section 12.1].

First, we show that $(V_{M_2}^n, X^n, Y^n) \in T_{\epsilon''}^n$ with a probability tends to 1 as $n$ increases.

Since $Y^n|\{V_{M_2}^n = v_{m_2}^n, X^n = x^n\} \sim \prod_{i=1}^n P_{Y|X}(y|x)$ and $\epsilon'' > \epsilon'''$, by the Markov lemma, $\Pr\{(V_{M_2}^n, X^n, Y^n) \notin T_{\epsilon''}^n\}$ tends to zero as $n \to \infty$.

Second, we show that $(U_{M_1}^n, V_{M_2}^n, X^n, Y^n) \in T_{\epsilon'}^n$ with a probability tends to 1 as $n$ increases similar as in Appendix C.3.

Hence, we have proven that the constraint on the type 1 error probability is satisfied.

*Step 2.3: Type 1 error analysis for $P_{XY} \in \mathcal{N}_{t_{XY}} \setminus t_{XY}$.*

As all $P_{XY} \in \Pi^{\perp}$ has the same marginal distribution as $Q_x Q_Y$, we can easily get that if $R_1 \geq I(U;X) + \delta(\epsilon)$, there exists at least one $u^n$ that is jointly typical with $x^n$ according to $t_{UX}$. Similarly, we can get that if $R_2 \geq I(V;Y) + \delta(\epsilon)$, there exists at least one $v^n$ that is jointly typical with $y^n$ according to $t_{VY}$.

From the way we generate the $(U^n, V^n)$ sequences, we can get that the Markov chain $V_{M_2}^n \leftrightarrow Y^n \leftrightarrow X^n \leftrightarrow U_{M_1}^N$ still holds. Then we can use the same steps as in *Step 2.2* to prove that the third term $P_{XY}^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3)$ still holds.

*Step 2.4: Acceptance region for all $P_{XY} \in \Pi^{\perp}$.*

According to the results in *Step 2.2* and *Step 2.3*, we can set the acceptance region as

$$\mathcal{A}_n = \bigcup_{t_{XY} \in \Lambda_n^{\Pi^{\perp}}(\mathcal{X}\mathcal{Y})} \mathcal{A}_n^{t_{XY}} \tag{4.59}$$

It is obvious that

$$\mathbb{E}\{\alpha_{nc}\} = \sup_{P_{XY} \in \Pi^{\perp}} P_{XY}^n(\mathcal{A}_n^c) \leq \epsilon.$$

*Step 2.5: The type 2 error exponent.*

For the type 2 error probability, assume in this case that $H_1$ is true. Hence,

$$(Q_X Q_Y)^n(\mathcal{A}_n) = (Q_X Q_Y)^n \left( \bigcup_{t_{XY} \in \Lambda_n^{\Pi^\perp}(\mathcal{XY})} \mathcal{A}_n^{t_{XY}} \right) \tag{4.60}$$

$$\leq \sum_{t_{XY} \in \Lambda_n^{\Pi^\perp}(\mathcal{XY})} (Q_X Q_Y)^n \left( \mathcal{A}_n^{t_{XY}} \right) \tag{4.61}$$

Then, for a particular $t_{XY}$, we have

$$\begin{aligned}
\mathbb{E}\{\beta_{nc}\} &= (Q_X Q_Y)^n(\varepsilon_1^c \cap \varepsilon_2^c \cap \varepsilon_3^c) \\
&= (Q_X Q_Y)^n(\varepsilon_1^c) \cdot (Q_X Q_Y)^n(\varepsilon_2^c) \cdot (Q_X Q_Y)^n(\varepsilon_3^c | \varepsilon_2^c \varepsilon_1^c)
\end{aligned}$$

We now bound each factor.

(1) By the covering lemma, $(Q_X Q_Y)^n(\varepsilon_1^c)$ tends to 1 as $n \to \infty$ if $R_1 \geq I(U;X) + \delta(\epsilon)$. Similarly, $(Q_X Q_Y)^n(\varepsilon_2^c)$ tends to 1 as $n \to \infty$ if $R_2 \geq I(V;Y) + \delta(\epsilon)$.

(2)

$$\begin{aligned}
(Q_X Q_Y)^n(\varepsilon_3^c | \varepsilon_2^c, \varepsilon_1^c) &= \sum_{(u^n, v^n) \in T_\epsilon(t_{UV})} Q_X^n(u^n | \varepsilon_2^c, \varepsilon_1^c) Q_Y^n(v^n | \varepsilon_2^c, \varepsilon_1^c) \\
&\overset{(a)}{\leq} 2^{n(H(UV)+\delta(\epsilon)-H(U)-\delta(\epsilon)-H(V)-\delta(\epsilon))} \\
&= 2^{-n(I(U;V)-\delta(\epsilon))},
\end{aligned} \tag{4.62}$$

where $(a)$ is true due to the fact that the marginal distributions for $X$ and $Y$ repecitively are the same in both hypotheses.

Hence, we give a lower bound on the type 2 error exponent.

$$(Q_X Q_Y)^n(\mathcal{A}_n) \leq \sum_{t_{XY} \in \Lambda_n^{\Pi^\perp}(\mathcal{XY})} (Q_X Q_Y)^n(\mathcal{A}_n^{t_{XY}}) \tag{4.63}$$

$$\leq \sum_{t_{XY} \in \Lambda_n^{\Pi^\perp}} 2^{-n(I(U;V)-\delta(\epsilon))} \tag{4.64}$$

$$\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{-n\left(\inf_{P_{XY} \in \Pi^\perp} I(U;V)-\delta(\epsilon)\right)}. \tag{4.65}$$

***c) Existence of a particular codebook:*** Similar to Section 2.5, we can show that there exists a codebook $c^*$ such that

$$\begin{aligned} \alpha_{nc^*} &\leq \epsilon, \\ \beta_{nc^*} &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{-n\left(\inf_{P_{XY} \in \Pi^\perp} I(U;V)-\delta(\epsilon)\right)}, \end{aligned}$$

as long as

$$R_1 \geq I(U_1; X_1), \quad R_2 \geq I(U_2; X_2|U_1).$$

This completes the achievability proof.

**Converse:**

We will show that for any encoding and decoding scheme that satisfies the type 1 error constraint $\alpha_n \leq \epsilon$ and rate constraints (4.4), the type 2 error exponent is upper bounded by right side of (4.55).

For each $P_{XY} \in \Pi^\perp$ and $k = 1, 2, \ldots$, define :

$$\theta_k(R_1, R_2) = \sup_{f,g} \left\{ \frac{1}{k} D(P_{f(X^k)g(Y^k)} || Q_{f(X^k)} Q_{g(Y^k)}) \Big| \right.$$
$$\left. \frac{1}{k} \log ||f|| \leq R_1, \frac{1}{k} \log ||g|| \leq R_2 \right\}, \tag{4.66}$$

For every $R_1 \geq 0$ and $R_2 \geq 0$, according to [1], we have the following

$$a) \limsup_{n\to\infty} \frac{1}{n} \log \beta_n \leq -\sup_k \theta_k(R_1, R_2) \text{ for all } \epsilon \in (0,1), \tag{4.67}$$

$$b) \lim_{\epsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \beta_n \geq -\sup_k \theta_k(R_1, R_2). \tag{4.68}$$

From above, we can see that when $\epsilon \to 0$,

$$\theta(R_1, R_2, \epsilon) = \inf_{P_{XY} \in \Pi^{\perp}} \sup_k \theta_k(R_1, R_2). \tag{4.69}$$

Moreover,

$$\frac{1}{n} D(P_{f(X^k)g(Y^k)} || Q_{f(X^k)} Q_{g(Y^k)}) = \frac{1}{n} I(f(X^n); g(Y^n)).$$

Then, we single-letterize the $\theta(R_1, R_2)$ in the following way. For $R_1$, we have

$$\begin{aligned} nR_1 &\geq H(M_1) \\ &\geq I(M_1; X_1^n) \\ &= \sum_{i=1}^{n} I(M_1; X_i | X^{i-1}) \\ &= \sum_{i=1}^{n} I(M_1, X^{i-1}; X_i) \\ &\overset{(a)}{=} \sum_{i=1}^{n} I(U_i; X_i). \end{aligned}$$

where $(a)$ is true by identifying $U_i = (M_1, X^{i-1})$ and noting that $U_i \to X_i \to Y_i$ forms a Markov chain as

$$\begin{aligned} &(X^n, X^{i-1}) \leftrightarrow X_i \leftrightarrow Y_i \\ &\Rightarrow (M_1, X^{i-1}) \leftrightarrow X_i \leftrightarrow Y_i. \end{aligned}$$

Similarly, we can get

$$nR_2 \overset{(b)}{\geq} \sum_{i=1} I(V_i; Y_i), \tag{4.70}$$

where $(b)$ is true by identifying $V_i = (M_2, Y^{i-1})$ and noting that $V_i \to Y_i \to X_i$ forms a

Markov chain as

$$(Y^n, Y^{i-1}) \leftrightarrow Y_i \leftrightarrow X_i \Rightarrow (M_2, Y^{i-1}) \leftrightarrow Y_i \leftrightarrow X_i.$$

Then we have

$$I(M_1; M_2) \leq \sum_{i=1} I(M_1 X^{i-1}; M_2 X^{i-1}) \tag{4.71}$$

$$= \sum_{i=1} I(U_i; V_i) \tag{4.72}$$

Using a time-sharing random variable $Q \sim \text{Unif}[1 : n]$, independent of $(X^n, Y^n, U^n, V^n)$ we obtain

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i; U_i|Q = i) = I(X_Q; U_Q|Q),$$
$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i; V_i|Q = i) = I(Y_Q; V_Q|Q),$$
$$\frac{1}{n}\sum_{i=1}^{n} I(U_i; V_i|Q = i) = I(U_Q; V_Q|Q),$$

Since $Q$ is independent of $X_Q$, we have

$$I(X_Q; U_Q|Q) = I(X_Q; U_Q, Q).$$

Thus, defining $X = X_Q, Y = Y_Q, U = (U_Q, Q)$ and $V = (V_Q, Q)$ and letting $n \to \infty$, we have shown that

$$R_1 \geq I(X; U),$$

$$R_2 \geq I(Y; V),$$

$$\lim_{\epsilon \to 0} \theta(R_1, R_2, \epsilon) \leq I(U; V).$$

for some conditional PMF $P_{U|X}$ and $P_{V|Y}$.

□

### 4.5.2 Results with Exponential-type Constraint

In this subsection, we consider the case with exponential-type constraint, i.e., we require that the exponent of the type 1 error should be larger than $r$. Due the complexity of the problem for the general case $S_{XY}$, we here give the result assuming terminal $\mathcal{Y}$ can communicate with a large rate $R_2 > \log|\mathcal{Y}|$ so that the decision maker has full information about $Y^n$. We will use $\sigma(R_1, r)$ to denote the corresponding type 2 error exponent.

Let $\mathcal{U}$ be an arbitrary finite set and $\mathcal{P}(\mathcal{U}|\mathcal{X})$be the set of all possible conditional probability distributions $(P_{U|X}(u|x))_{(u,x)\in\mathcal{U}\times\mathcal{X}}$ on $\mathcal{U}$ given values in $\mathcal{X}$. Let $\omega$ denote the continuous mapping from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{U}|\mathcal{X})$ and $\Phi$ be the set of all possible $\omega$.

**Theorem 4.4.** For $R_1 \geq 0$, $r \geq 0$, we have

$$\sigma(R_1, r) \geq \inf_{P_{XY}\in\Pi} \sup_{\omega\in\phi_{P_{XY}}(R_1,r)} \min_{\tilde{P}_{UXY}\in\Xi_{P_{XY}}(\omega)} D\left(\tilde{P}_{UXY}||Q_{UXY}\right). \tag{4.73}$$

where

$$\phi_{P_{XY}}(R_1, r) = \left\{\omega \in \Phi : \max_{\substack{\hat{X}\,:\,D(\hat{X}||X)\,\leq\, r \\ \hat{P}_{U|X}\,=\,\omega(\hat{X})}} I\left(\hat{U};\hat{X}\right) \leq R_1\right\},$$

$$\hat{\Xi}_{P_{XY}}(\omega) = \left\{\hat{P}_{UXY} : \begin{array}{c} D\left(\hat{P}_{UXY}||P_{UXY}\right) \leq r \\ P_{U|X} = \hat{P}_{U|X} = \omega(\hat{X}) \\ U \leftrightarrow X \leftrightarrow Y \end{array}\right\},$$

$$\Xi_{P_{XY}}(\omega) = \Big\{\tilde{P}_{UXY} : \tilde{P}_{UX} = \hat{P}_{UX}, \tilde{P}_{UY} = \hat{P}_{UY}, \text{ for some } \hat{P}_{UXY} \in \hat{\Xi}_{t_{XY}}(\omega)\Big\},$$

and $Q_{U|X} = \tilde{P}_{U|X}, Q_{UXY} = Q_{U|X}Q_{XY}$.

*Proof.* **Step 1: Encoding.**

In this step, we show a universal encoding scheme.

Given a sequence $x^n$, terminal $\mathcal{X}$ finds its type $t_X$. Let $\eta > 0$ be an arbitrary small number. For each type variable $\hat{X}^{(n)}$ with $\rho_n \equiv D(\hat{X}^{(n)}||t_X) \leq r + \eta$ choose a joint type variable $\hat{U}^{(n)}\hat{X}^{(n)}$ for $\mathcal{U}^n \times \mathcal{X}^n$ such that

$$\mu_n \equiv D\left(\hat{U}^{(n)}|\hat{X}^{(n)}||t_{U|X}\right) \leq \eta/3; \tag{4.74}$$

$$I\left(\hat{U}^{(n)}; \hat{X}^{(n)}\right) \leq R + \eta/3, \tag{4.75}$$

where $t_{U|X} = \omega(\hat{X}^{(n)})$. With this $\hat{U}^{(n)}\hat{X}^{(n)}$, we put

$$\hat{M} \equiv M(\hat{X}^{(n)}) \equiv \exp\left(n\left(I\left(\hat{U}^{(n)}; \hat{X}^{(n)}\right) + \eta/3\right)\right). \tag{4.76}$$

It is easily shown that there exists $M(\hat{X}^{(n)})$ $u_m^n$ such that for every $x \in S_0^n(\hat{X}^{(n)})$ we have some $u_i^n$ such that $(u_i^n, x^n) \in S_0^n(\hat{U}^{(n)}\hat{X}^{(n)})$, $i \in \{1, \cdots, M(\hat{X}^{(n)})\}$. Send the index $i \in \{1, \cdots, M(\hat{X}^{(n)})\}$ to the decision maker.

**Step 2: Testing.**

Upon receiving $i$, the decision maker needs to design a universal acceptance region so that the type 1 error constraint is satisfied. Based on Lemma 4.1, for each center point $t_{XY}$, we set an individual acceptance region $\mathcal{A}_n^{t_{XY}}$ such that the type 1 error probability constraint under $t_{XY}$ is satisfied. Then we show that using the same acceptance region $\mathcal{A}_n^{t_{XY}}$, the type 1 error probability constraint can also be satisfied for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$. Finally, we make the final acceptance region as the union of a finite number of acceptance regions $\mathcal{A}_n^{t_{XY}}$ and show that the type 1 error probability constraint under $H_0$ is satisfied.

*Step 2.1: Acceptance region for $t_{XY}$.*

To define an acceptance region $\mathcal{A}_n^{t_{XY}}$, we have the following steps.

(1) Region $\mathcal{C}_i$.

It can be easily shown that there exist $u_1, \cdots, u_M$ such that for every $x \in S_0^n(\hat{X}^{(n)})$, we have some $u_i$ jointly typical with $x^n$, i.e. $(u_i, x^n) \in S_0^n(\hat{X}^{(n)})$. Hence, there exist $\hat{M}$ disjoint subsets $\mathcal{C}_1, \cdots, \mathcal{C}_M \subset S_0^n(\hat{X}^{(n)})$ such that

$$S_0^n(\hat{X}^{(n)}) = \bigcup_{i=1}^{\hat{M}} \mathcal{C}_i, \tag{4.77}$$

and for every $x^n \in \mathcal{C}_i$, we have

$$(u_i^n, x^n) \in S_0^n(\hat{X}^{(n)}), \quad , i = 1, \cdots, \hat{M}. \tag{4.78}$$

(2) Region $B_i(x^n)$.

For each $i \in \{1, \cdots, \hat{M}\}$ and $x^n \in \mathcal{C}_i$, define

$$\begin{aligned} B_i(x^n) &= \left\{ y^n \in \mathcal{Y}^n : (u_i^n, x^n, y^n) \in S_{r+\eta}^n(t_{UXY}) \right\} \\ &= \left\{ y^n \in \mathcal{Y}^n : D(u_i^n, x^n, y^n || t_{UXY}) \leq r + \eta \right\}, \end{aligned} \tag{4.79}$$

where $t_{UXY} = t_{XY} P_{U|X}$. By simple derivations, we have

$$D\left(\hat{U}^{(n)}\hat{X}^{(n)} || t_{UX}\right) = D\left(\hat{X}^{(n)} || t_X\right) + D\left(\hat{U}^{(n)} | \hat{X}^{(n)} || t_{U|X}\right). \tag{4.80}$$

Hence, using (4.74), (4.80) and the Markov chain $U \leftrightarrow X \leftrightarrow Y$, we have

$$\begin{aligned} &\Pr\{Y^n \in B_i(x^n) | X^n = x^n\} \\ &= \Pr\{Y^n \in B_i(x^n) | U^n X^n = u_i^n x^n\} \\ &\geq 1 - (n+1)^{|\mathcal{U}| \cdot |\mathcal{X}| \cdot |\mathcal{Y}|} \exp[-n(r + \eta - \rho_n - \mu_n)] \\ &\geq 1 - \exp[-n(r + \eta/3 - \rho_n)]. \end{aligned} \tag{4.81}$$

(3) Region $B(\hat{X}^{(n)})$.

Set

$$B(\hat{X}^{(n)}) = \bigcup_{i=1}^{\hat{M}} \bigcup_{x^n \in \mathcal{C}_i} (\{x^n\} \times B_i(x^n)); \tag{4.82}$$

then we have

$$\begin{aligned}
&\Pr\{X^n Y^n \in B(\hat{X}^{(n)}))\} \\
&= \sum_{i=1}^{\hat{M}} \sum_{x^n \in \mathcal{C}_i} \Pr\{X^n = x^n\} \cdot \Pr\{Y^n \in B_i(x^n | X^n = x^n) \\
&\geq \sum_{i=1}^{\hat{M}} \sum_{x^n \in \mathcal{C}_i} \Pr\{X^n = x^n\} \cdot (1 - \exp[-n(r + \eta/3 - \rho_n)]) \\
&= \Pr\{X^n \in S_0^n(\hat{X}^n)\} \cdot (1 - \exp[-n(r + \eta/3 - \rho_n)]) \\
&= \Pr\{X^n \in S_0^n(\hat{X}^n)\} - \Pr\{X^n \in S_0^n(\hat{X}^n)\} \cdot \exp[-n(r + \eta/3 - \rho_n)] \\
&\overset{(a)}{\geq} \Pr\{X^n \in S_0^n(\hat{X}^n)\} - \exp[-n(r + \eta/3)],
\end{aligned}$$

where $(a)$ is true due to the fact that

$$\begin{aligned}
\Pr(X^n \in S_0^n(\hat{X}^{(n)})) &= \sum_{x^n \in S_0^n(\hat{X}^{(n)})} \Pr(X^n = x^n) \\
&\leq \exp\left(n\left(H\left(\hat{X}^{(n)}\right)\right)\right) \cdot \exp\left(-n\left(H\left(\hat{X}^{(n)}\right) + D\left(\hat{X}^{(n)} || t_X\right)\right)\right) \\
&\leq \exp(-n\rho_n).
\end{aligned}$$

(4) Region $\mathcal{B}_i$.

For each $i \in \{1, \cdots, \hat{M}\}$, set

$$\mathcal{B}_i = \{y^n \in \mathcal{Y}^n : (x^n, y^n) \in B_i(x^n) \text{ for some } x^n \in \mathcal{C}_i\} \tag{4.83}$$

and

$$A(\hat{X}^{(n)}) = \bigcup_{i=1}^{\hat{M}} (\mathcal{C}_i \times \mathcal{B}_i). \tag{4.84}$$

Then, since for every $x^n \in \mathcal{C}_i$ $B_i(x^n) \subset \mathcal{B}_i$, we have that $B(\hat{X}^{(n)}) \subset A(\hat{X}^{(n)})$. Hence, we have

$$\Pr(X^n Y^n \in A(\hat{X}^{(n)})) \geq \Pr\{X^n \in S_0^n(\hat{X}^n)\} - \exp[-n(r+\eta/3)]. \tag{4.85}$$

(5) Acceptance Region $\mathcal{A}_n^{t_{XY}}$.

Set the acceptance region as

$$\mathcal{A}_n^{t_{XY}} = \bigcup_{\substack{\hat{X}^{(n)}\,: \\ D(\hat{X}^{(n)}||X) \leq r+\eta}} A(\hat{X}^{(n)}) \tag{4.86}$$

(6) Type 1 error probability.

$$\begin{aligned}
1 - \alpha_n^{t_{XY}} &= \Pr(X^n Y^n \in \mathcal{A}_n^{t_{XY}}) \\
&= \sum_{\substack{\hat{X}^{(n)}\,: \\ D(\hat{X}^{(n)}||X) \leq r+\eta}} \Pr(A(\hat{X}^{(n)})) \\
&= \sum_{\substack{\hat{X}^{(n)}\,: \\ D(\hat{X}^{(n)}||X) \leq r+\eta}} \Pr(X^n \in S_0^n(\hat{X}^{(n)})) - \sum_{\substack{\hat{X}^{(n)}\,: \\ D(\hat{X}^{(n)}||X) \leq r+\eta}} \exp[(-n(r+\eta/3)) \\
&\geq 1 - (n+1)^{|\mathcal{X}|} \exp(-n(r+\eta)) - (n+1)^{|\mathcal{X}|} \exp(-n(r+\eta/3)) \\
&\geq 1 - \exp(-nr)
\end{aligned}$$

for $n$ large enough, which means the constraint on the type 1 error probability is satisfied.

*Step 2.3: Type 1 error analysis for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}} \setminus t_{XY}$.*

As for $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$, $||\tilde{t}_{XY} - t_{XY}||_1 \leq \frac{1}{n}$, $X^nY^n$ generated according to $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$, has the type $t_X$ and $t_Y$. Hence, the region of $\hat{X}^{(n)}$ is the same as the region generated by $t_{XY}$. And we can calculate the distance of $\hat{X}^{(n)}$ and $\tilde{t}_X$:

$$\begin{aligned}
D(\hat{X}^{(n)}||\tilde{t}_X) &= \sum_{x\in\mathcal{X}} \hat{t}_X(x) \log \frac{\hat{t}_X(x)}{\tilde{t}_X(x)} \\
&= \sum_{x\in\mathcal{X}} \hat{t}_X(x) \log \frac{\hat{t}_X(x)}{t_X(x)} + \sum_{x\in\mathcal{X}} \hat{t}_X(x) \log \frac{t_X(x)}{\tilde{t}_X(x)} \\
&\overset{(a)}{=} D(\hat{X}^{(n)}||t_X) + \sum_{x\in\mathcal{X}\setminus\{\tilde{t}_X(x)=0\}} \hat{t}_X(x) \log \frac{t_X(x)}{\tilde{t}_X(x)} \\
&\leq D(\hat{X}^{(n)}||t_X) + \sum_{x\in\mathcal{X}\setminus\{\tilde{t}_X(x)=0\}} \hat{t}_X(x) \log \frac{\tilde{t}_X(x)+\zeta}{\tilde{t}_X(x)} \\
&\overset{(b)}{\leq} D(\hat{X}^{(n)}||t_X) + \sum_{x\in\mathcal{X}\setminus\{\tilde{t}_X(x)=0\}} \hat{t}_X(x) \frac{\zeta}{\tilde{t}_X(x)} \\
&\leq D(\hat{X}^{(n)}||t_X) + \delta(\zeta), \qquad (4.87)
\end{aligned}$$

where $(a)$ is true as if $\exists x_0 \in \mathcal{X}$, such that $\tilde{t}_X(x_0) = 0$, then $t_X(x_0) < \frac{1}{n}$, which means $t_X(x_0) = 0$ for $n$-sequence; $(b)$ is true due to the inequality $\log(1 + c \cdot x) \leq c \cdot x$ for $x > 0$ and a constant $c > 0$. Furthermore, from the last inequality, we can see that $\delta(\zeta)$ is a function of $\zeta$ and $\delta(\zeta) \sim O(\frac{1}{n})$.

Using the inequality (4.87), we can verify that all the equations/inequalities in *Step 2.2* still hold.

*Step 2.4: Acceptance region for all $P_{XY} \in \Pi$.*

According to the results in *Step 2.2* and *Step 2.3*, we can set the acceptance region as

$$\mathcal{A}_n = \bigcup_{t_{XY}\in\Lambda_n^{\Pi}(\mathcal{X}\mathcal{Y})} \mathcal{A}_n^{t_{XY}}. \qquad (4.88)$$

It is obvious that

$$\alpha_n = \sup_{P_{XY}\in\Pi} P_{XY}^n(\mathcal{A}_n^c) \leq \exp(-nr).$$

*Step 2.5: The type 2 error exponent.*

For the type 2 error probability, we have

$$\begin{aligned}\beta_n &= Q_{XY}^n(\mathcal{A}_n) \\ &= Q_{XY}^n\left(\bigcup_{t_{XY}\in\Lambda_n^{\Pi}(\mathcal{XY})} \mathcal{A}_n^{t_{XY}}\right) \\ &\leq \sum_{t_{XY}\in\Lambda_n^{\Pi}(\mathcal{XY})} Q_{XY}^n(\mathcal{A}_n^{t_{XY}}). \end{aligned} \tag{4.89}$$

First, we analyze the type 2 error probability for a particular $\mathcal{A}_n^{t_{XY}}$. fix any type variable $\hat{X}^{(n)}$ for $\mathcal{X}^n$ such that $D(\hat{t}_X||t_X) \leq r+\eta$. With this $\hat{X}^{(n)}$, set

$$\mathcal{F}_i = \{u_i^n\} \times \mathcal{C}_i \times \mathcal{B}_i, \quad i = 1, \cdots, \hat{M}. \tag{4.90}$$

Let $\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}$ be any possible type variable for $\mathcal{F}_i$ where $\tilde{t}_X = \hat{t}_X$, $\tilde{P}_{U|X} = \omega(\hat{X}^{(n)})$. Define

$$\mathcal{F}_i(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}) = \left\{(u_i^n, x^n, y^n) \in \mathcal{F}_i : tp(u_i^n x^n y^n) = \tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}\right\}.$$

It can be checked by Lemma 2.2 that for each $i = 1, \cdots, \hat{M}$.

$$\left|\mathcal{F}_i(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)})\right| \leq \exp\left(-n\left(H\left(\tilde{Y}^{(n)}|\tilde{U}^{(n)}\right) + H\left(\tilde{X}^{(n)}|\tilde{U}^{(n)}\tilde{Y}^{(n)}\right)\right)\right).$$

We also have

$$Q_{XY}^n(x^n y^n) = \exp\left(-n\left(H\left(\tilde{X}^{(n)}\tilde{Y}^{(n)}\right) + D\left(\tilde{X}^{(n)}\tilde{Y}^{(n)}||Q_{XY}\right)\right)\right) \tag{4.91}$$

for any $(x^n, y^n)$ such that $tp(x^n y^n) = \tilde{X}^{(n)}\tilde{Y}^{(n)}$. Let

$$\mathcal{F}(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}) = \bigcup_{i=1}^{\hat{M}} \mathcal{F}_i \left(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}\right) \tag{4.92}$$

and it follows that

$$\begin{aligned}
\beta_n^{t_{XY}} \left(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}\right) &= Q_{XY}^n \left(\mathcal{F}\left(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}\right)\right) \\
&\leq \exp\left(-n\left(H\left(\tilde{X}^{(n)}\tilde{Y}^{(n)}\right) + D\left(\tilde{X}^{(n)}\tilde{Y}^{(n)}||Q_{XY}\right)\right)\right) \\
&\quad \cdot \exp\left(-n\left(H\left(\tilde{Y}^{(n)}|\tilde{U}^{(n)}\right) + H\left(\tilde{X}^{(n)}|\tilde{U}^{(n)}\tilde{Y}^{(n)}\right)\right)\right) \\
&\quad \cdot \exp\left(n\left(I\left(\tilde{X}^{(n)}; \tilde{U}^{(n)}\right) + \eta/3\right)\right) \\
&= \exp\left(-n\left(D\left(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}||Q_{UXY}\right) - \eta/3\right)\right)
\end{aligned}$$

where $Q_{U|X} = \tilde{P}_{U|X}$ and $Q_{UXY} = Q_{U|X}Q_{XY}$, i.e. $U \leftrightarrow X \leftrightarrow Y$ forms a Markov chain under $Q_{UXY}$. Thus the type 2 error probability for $\mathcal{A}_n^{t_{XY}}$ is

$$\begin{aligned}
\beta_n^{t_{XY}} &\leq \sum_{\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}} \exp\left(-n\left(D\left(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}||Q_{UXY}\right) - \eta/3\right)\right) \\
&\leq (n+1)^{|\mathcal{U}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\left(-n\left(D\left(\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}||Q_{UXY}\right) - \eta/3\right)\right)
\end{aligned} \tag{4.93}$$

Meanwhile, from the way of constructing $\mathcal{B}_i$ in (4.83), it follows that for any possible $\tilde{U}^{(n)}\tilde{X}^{(n)}\tilde{Y}^{(n)}$, there must exist some type variable $\hat{U}^{(n)}\hat{X}^{(n)}\hat{Y}^{(n)}$ for the subset $\{u_i^n\} \times \cup_{x^n \in \mathcal{C}_i}(\{x^n\} \times \mathcal{B}_i)$ such that

$$\begin{aligned}
&D\left(\hat{U}^{(n)}\hat{X}^{(n)}\hat{Y}^{(n)}||t_{UXY}\right) \leq r + \eta \\
&\hat{P}_{UX} = \tilde{P}_{UX},\ \hat{P}_{UY} = \tilde{P}_{UY}.
\end{aligned} \tag{4.94}$$

Note that $\hat{U}^{(n)}\hat{X}^{(n)}\hat{Y}^{(n)}$ has to satisfy (4.74), therefore, we define the following two sets:

$$\hat{\Xi}^{\eta}_{t_{XY}}(\omega) = \left\{ \hat{P}_{UXY} : \begin{array}{c} D\left(\hat{P}_{UXY}||t_{UXY}\right) \leq r+\eta \\ D\left(\hat{P}_{U|X}||t_{U|X}\right) \leq \eta/3 \\ t_{U|X} = \omega(\hat{X}) \\ U \leftrightarrow X \leftrightarrow Y \end{array} \right\},$$

$$\Xi^{\eta}_{t_{XY}}(\omega) = \left\{ \tilde{P}_{UXY} : \tilde{P}_{UX} = \hat{P}_{UX}, \tilde{P}_{UY} = \hat{P}_{UY}, \text{for some } \hat{P}_{UXY} \in \hat{\Xi}^{\eta}_{t_{XY}}(\omega) \right\}.$$

Then, (4.93) yields for each $\omega \in \phi_{t_{XY}}$,

$$\beta_n^{t_{XY}} \leq (n+1)^{|\mathcal{U}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\left( -n \left( \min_{\tilde{P}_{UXY} \in \Xi^{\eta}_{t_{XY}}} D\left(\tilde{P}_{UXY}||Q_{UXY}\right) - \eta/3 \right) \right),$$

where $Q_{U|X} = \tilde{P}_{U|X}, Q_{UXY} = Q_{U|X}Q_{XY}$. Hence, we have

$$\beta_n^{t_{XY}} \leq (n+1)^{|\mathcal{U}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\left( -n \left( \max_{\omega \in \phi_{t_{XY}}} \min_{\tilde{P}_{UXY} \in \Xi^{\eta}_{t_{XY}}} D\left(\tilde{P}_{UXY}||Q_{UXY}\right) \right) \right).$$

Then (4.89) yields

$$\begin{aligned} \beta_n &\leq \sum_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})} (n+1)^{|\mathcal{U}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|} \exp\left( -n \left( \max_{\omega \in \phi_{t_{XY}}} \min_{\tilde{P}_{UXY}} D\left(\tilde{P}_{UXY}||Q_{UXY}\right) \right) \right) && (4.95) \\ &\leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|}(n+1)^{|\mathcal{U}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|} \max_{t_{XY} \in \Lambda_n^{\Pi}(\mathcal{XY})} \\ &\qquad \exp\Big( -n \left( \max_{\omega \in \phi_{t_{XY}}} \min_{\tilde{P}_{UXY}} D\left(\tilde{P}_{UXY}||Q_{UXY}\right) \right) \Big) && (4.96) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}(n+1)^{|\mathcal{U}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|} \exp\Big( -n \Big( \inf_{P_{XY} \in \Pi} \sup_{\omega \in \phi_{P_{XY}}} \\ &\qquad \min_{\tilde{P}_{UXY}} D\left(\tilde{P}_{UXY}||Q_{UXY}\right) \Big) \Big), && (4.97) \end{aligned}$$

where $Q_{U|X} = \tilde{P}_{U|X}, Q_{UXY} = Q_{U|X}Q_{XY}$.

Hence, we give a lower bound on the type 2 error exponent.

□

## 4.6 Conclusion

In this chapter, we have studied the distributed identity testing problem, in which the decision maker should decide whether the distribution indirectly revealed from the compressed data from multiple distributed terminals is the same as or $\lambda$-far from a given distribution. Under zero-rate compression and exponential-type constraint on the type 1 error probability, we have fully characterized the type 2 error exponent by providing matching upper and lower bounds. We have also fully characterized the type 2 error exponent for the testing against independence case under constant-type constraint and positive transmission rate constraint. Finally, for the positive transmission rate case, we have established a lower bound on the type 2 error exponent for general PMF case under constant-type error probability constraint and exponential-type error probability constraint.

# Chapter 5

# Conclusion and Extensions

In this chapter, we summarize the contributions we have made in this dissertation and propose certain potential directions in the field of distributed hypothesis testing.

## 5.1 Conclusion

This dissertation has explored the distributed inference problems from information theoretic perspective.

First, we have discussed the distributed inference problems with non-interactive encoders. Using properties of $r$-divergence sequences, we have characterized the best error exponent of the type 2 error probability under both the zero-rate compression and exponential-type type 1 error probability constraints. Furthermore, we have discussed the problem of testing against independence under the constant-type constraint on the type 1 error probability. We have derived a lower bound and an upper bound on the type 2 error exponent.

Second, we have considered distributed testing problems with cascaded encoders. We have first investigated the special case of testing against independence. We have designed a scheme to benefit from the extra information provided by cascaded communications, and have shown that the proposed scheme is optimal when certain Markovian relation exists. We have then derived a lower bound on the type 2 error exponent for cases with general

hypotheses. Compared with the existing results in the non-interactive communication cases, we have shown that cascaded communication does achieve performance gain under certain PMFs and positive communication rates but it does not offer gain under zero-rate data compression scenarios.

Finally, we have studied the distributed identity testing problem, in which the decision maker decides whether the distribution indirectly revealed from the compressed data from multiple distributed terminals is the same as or $\lambda$-far from a given distribution. Under zero-rate compression and exponential-type constraint on the type 1 error probability, we have fully characterized the type 2 error exponent by providing matching upper and lower bounds. We have also fully characterized the type 2 error exponent for the testing against independence case under constant-type constraint and positive transmission rate constraint. Finally, for the positive transmission rate case, we have established a lower bound on the type 2 error exponent for general PMF case under constant-type error probability constraint and exponential-type error probability constraint.

## 5.2 Future Directions

Equipped with the techniques and results presented in this dissertation, we can extend the current research on the following directions. First, it will be interesting to design inference algorithms with more sophisticated interactive communication schemes. Second, it is of interest to investigate how to compress data in nonparametric scenarios with unknown learning tasks.

### 5.2.1 Distributed Inference with Sophisticated Interactive Schemes

In Chapter 3, we have discussed the distributed inference problem under a simple form of interaction among users, in which terminal $\mathcal{X}_l$ encodes messages based on its own data and the messages received from terminal $\mathcal{X}_{l'}$, $l' = 1, \cdots, l-1$. As a natural extension, one can con-

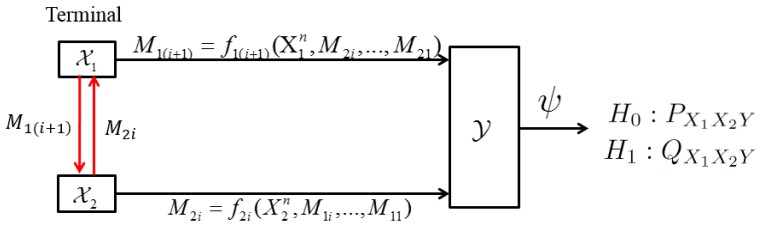



Figure 5.1: Sophisticated interactive communication

sider a more sophisticated form of interaction among users, in which terminal $\mathcal{X}_l$s can send multiple rounds of messages to each other. This model is related to [21, 48] and references therein, which study the multiple rounds of communication between two terminals $\mathcal{X}$ and $\mathcal{Y}$. Different from their works, in this model, the decision maker can utilize its own information and messages received from the interactive communication from terminals $\mathcal{X}_l$s. Moreover, when $L > 2$, the problem is much more complicated and it needs further exploration.

To simplify the presentation, we use the case of $L = 2$ to illustrate the main idea. Terminal $\mathcal{X}_1$ first encodes its local data to messages $M_{11}$ and broadcasts it. Terminal $\mathcal{X}_2$ utilizes the messages $M_{11}$ to encode its own information as $M_{21}$ and broadcasts it. This is called one round of interactive communication of $\mathcal{X}_1$ and $\mathcal{X}_2$. After receiving $M_{21}$, terminal $\mathcal{X}_1$ can further encode its local information as $M_{21}$ and broadcast it. This process continues for $N$ rounds. The encoding functions for $\mathcal{X}_1$ and $\mathcal{X}_2$ can be written as

$$f_{1i} : \{\mathcal{X}_1^n, \mathcal{M}_{2(i-1)}, \cdots, \mathcal{M}_{21}\} \to \mathcal{M}_{1i} = \{1, \cdots, M_{1i}\}, \tag{5.1}$$

$$f_{2i} : \{\mathcal{X}_2^n, \mathcal{M}_{1i}, \cdots, \mathcal{M}_{11}\} \to \mathcal{M}_{2i} = \{1, \cdots, M_{2i}\}. \tag{5.2}$$

After receiving all messages from terminals $\mathcal{X}_1$ and $\mathcal{X}_2$, the decision maker $\mathcal{Y}$ makes a decision about the joint PMF of $(X_1, X_2, Y)$ using a decoding function

$$\psi : \{\mathcal{M}_{11}, \mathcal{M}_{21}, \cdots, \mathcal{M}_{1N}, \mathcal{M}_{2N}\} \to \{H_0, H_1\}, \tag{5.3}$$

where $H_0 : P_{X_1X_2Y}$ and $H_1 : Q_{X_1X_2Y}$. This process is shown in Figure 5.1.

The goal of this problem is similar to what we have discussed in Chapter 3, i.e., to maximize the type 2 error exponent under the constraints on the type 1 error probability and communication rates. Moreover, one may want to compare the performance with the one in Chapter 3. Intuitively, the decision maker obtains more information through multiple rounds of communication between terminals $\mathcal{X}_1$ and $\mathcal{X}_2$ and thus it may achieve a better performance. Finally, based on the result of $L = 2$, one can try to see whether or not it is possible to generalize the results to any $L$ terminals.

### 5.2.2 Learning Task Oblivious Data Summarization

In this section, we point out another interesting topic: data summarization when the learning or inference task is oblivious.

As introduced in the Section 1.1, the massive volume of data produced nowadays bring challenges in the storage and process of the large dataset. To overcome this difficulty, we have discussed the way to distribute the data into multiple terminals and infer useful information from these distributed data using the computation power offered by these distributed machines. Another potential resolution is to have big data summarized so that they need less storage and extremely shorter time to get processed and retrieved. The summarized data will be a compact but still informative version of the entire data. Various techniques to summarize the data are introduced in [17] and references therein to fulfill different learning or inference tasks. However, users might not specify the learning task or users might perform more than one task on the processed data. In these cases, the techniques introduced for a specific task may not be optimal universally and new methods are needed to be explored. One possible way is to select a subset of data which can represent the whole dataset. This method is illustrated in Figure 5.2, in which one has a large dataset $V$, and selects a subset $S$ that contains enough information of $V$ for any learning task.

More specifically, one aims to select a subset of data without a specified learning task under a nonparametric model, i.e., no assumption is made about the distribution of the observed

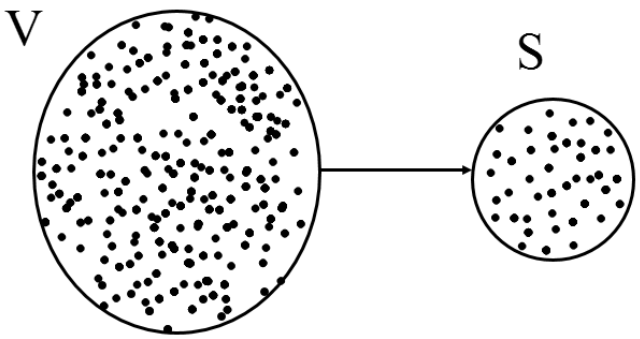



Figure 5.2: Data summarization

data. A reasonable approach is to select samples that preserve the distribution information revealed from the dataset as much as possible. Hence, one can first estimate the *probability density function (pdf)* of the observed data, then choose a subset such that the estimated pdf on the chosen subset approaches the original pdf on the whole dataset as close as possible.

Unfortunately, if one aims to choose the best subset of data point that minimizes the distance between the pdf estimated from the original dataset and the pdf estimated from the chosen dataset, the complexity is too high. In fact, the problem is an NP-hard problem, hence once the dataset size is large, the problem is not solvable. One possible approach to overcome this challenge is to transform the subset selection problem to a submodular maximization problem with cardinality constraint, which was first studied in [27]. Nemhauser and et al. proved that the greedy method can provide a good approximation to the optimal solution of the original NP-hard optimization problem within polynomial time complexity. Later in 2015, Mirzasoleiman et al. proposed a linear-time algorithm to maximize the monotone submodular function with cardinality constraint problem with near-optimal performance guarantee [26]. Hence, the ultimate goal in this problem is to design a proper submodular function to describe the distance between the two estimated pdfs and then use the stochastic greedy algorithm in [26] to solve it.

# Appendix A

# Appendix of Chapter 2

## A.1 Proof of Theorem 2.1

In this appendix, we present the proof of Theorem 2.1. In this proof, we need to show that for any encoding and decoding scheme that meets the type 1 error constraint, we have (2.19). Let $\mathcal{A}_n$ be an arbitrary acceptance region such that

$$\alpha_n \leq \exp(-nr), \quad r > 0 \tag{A.1}$$

where

$$\alpha_n = P^n_{X_1X_2Y}(\mathcal{A}_n^c). \tag{A.2}$$

Equations (A.1) and (A.2) imply that

$$P^n_{X_1X_2Y}(\mathcal{A}_n) \geq 1 - \exp(-n(r-\gamma)), \quad \forall n \geq n_0, \tag{A.3}$$

where $\gamma > 0$ is an arbitrarily small constant, and $n_0$ is a sufficiently large positive integer.

Next, select an arbitrary "internal point" $P_{X_{10}X_{20}Y_0} \in \varphi_r$, where $\varphi_r$ is specified in (2.22). Then clearly

$$D(P_{X_{10}X_{20}Y_0}||P_{X_1X_2Y}) < r. \tag{A.4}$$

Define

$$\hat{T}_n(\delta) = \{\text{joint types } \hat{P}_n \text{ on } \mathcal{X}_1^n \times \mathcal{X}_2^n \times \mathcal{Y}^n : D(\hat{P}_n||P_{X_{10}X_{20}Y_0}) < \delta\} \tag{A.5}$$

where $\delta > 0$ is an arbitrary constant. Then, in view of (A.4) and the uniform continuity of the divergence, for all $\hat{P}_n \in \hat{T}_n(\delta)$ it holds that

$$c_n \equiv D(\hat{P}_n||P_{X_1X_2Y}) < r - 2\gamma, \tag{A.6}$$

provided that we take $\gamma > 0$ and $\delta > 0$ sufficiently small. Consequently, according to Lemma 2.2, we have

$$|\mathcal{A}_n(\hat{P}_n)| \geq (1 - (n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\gamma))|S_0(\hat{P}_n)| \tag{A.7}$$

for all $\hat{P}_n \in \hat{T}_n(\delta)$. Now we define the set

$$T_n(\delta) = \{(x_1^n, x_2^n, y^n) \in \mathcal{X}_1^n \times \mathcal{X}_2^n \times \mathcal{Y}^n : X_1^{(n)}X_2^{(n)}Y^{(n)} \in \hat{T}_n(\delta)\} \tag{A.8}$$

and consider an i.i.d. random sequence of length $n$ generated according to the probability distribution $P_{X_{10}X_{20}Y_0}$. Then, from (A.7), we have

$$\begin{aligned}
&P^n_{X_{10}X_{20}Y_0}(\mathcal{A}_n) \geq P^n_{X_{10}X_{20}Y_0}(\mathcal{A}_n \cap T_n(\delta)) \\
&= \sum_{\hat{P}_n \in \hat{T}_n(\delta)} P^n_{X_{10}X_{20}Y_0}(\mathcal{A}_n \cap S_0(\hat{P}_n)) \\
&= \sum_{\hat{P}_n \in \hat{T}_n(\delta)} P^n_{X_{10}X_{20}Y_0}(\mathcal{A}_n(\hat{P}_n)) \\
&= \sum_{\hat{P}_n \in \hat{T}_n(\delta)} \sum_{\substack{tp(x_{10}^n, x_{20}^n, y_0^n) = \hat{P}_n \\ (x_{10}^n, x_{20}^n, y_0^n) \in \mathcal{A}_n}} P^n_{X_{10}X_{20}Y_0}(X_{10}^n = x_{10}^n, X_{20}^n = x_{20}^n, Y_0^n = y_0^n) \\
&\overset{(a)}{=} \sum_{\hat{P}_n \in \hat{T}_n(\delta)} \sum_{\substack{tp(x_{10}^n, x_{20}^n, y_0^n) = \hat{P}_n \\ (x_{10}^n, x_{20}^n, y_0^n) \in \mathcal{A}_n}} \exp\left(-n\left(H\left(X_{10}^{(n)}X_{20}^{(n)}Y_0^{(n)}\right) + D\left(X_{10}^{(n)}X_{20}^{(n)}Y_0^{(n)}||X_{10}X_{20}Y_0\right)\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\hat{P}_n \in \hat{T}_n(\delta)} |\mathcal{A}_n(\hat{P}_n)| \exp\left(-n\left(H\left(X_{10}^{(n)}X_{20}^{(n)}Y_0^{(n)}\right) + D\left(X_{10}^{(n)}X_{20}^{(n)}Y_0^{(n)}||X_{10}X_{20}Y_0\right)\right)\right) \\
&\geq \sum_{\hat{P}_n \in \hat{T}_n(\delta)} (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\gamma))|S_0(\hat{P}_n)| \\
&\qquad \exp\left(-n\left(H\left(X_{10}^{(n)}X_{20}^{(n)}Y_0^{(n)}\right) + D\left(X_{10}^{(n)}X_{20}^{(n)}Y_0^{(n)}||X_{10}X_{20}Y_0\right)\right)\right) \\
&\geq (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\gamma)) \sum_{\hat{P}_n \in \hat{T}_n(\delta)} P^n_{X_{10}X_{20}Y_0}\left(S_0(\hat{P}_n)\right) \\
&= (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\gamma)) P_{X_{10}X_{20}Y_0}\left(\hat{T}_n(\delta)\right) \\
&\geq (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\gamma)) \times (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\delta)), \qquad \text{(A.9)}
\end{aligned}
$$

where $(a)$ is true due to (2.12), and the last step is true due to (2.16).

Now consider the zero-rate $(R_1 = 0, R_2 = 0, R \geq 0)$ hypothesis testing problem with

$$
H_0 : P_{X_{10}X_{20}Y_0} \quad \text{vs} \quad H_1 : Q_{X_1X_2Y}. \qquad \text{(A.10)}
$$

Then, for this hypothesis testing problem, if we use the same acceptance region $\mathcal{A}_n$ as above, the type 1 error probability

$$
\begin{aligned}
\alpha_n^{(0)} &= 1 - P^n_{X_{10}X_{20}Y_0}(\mathcal{A}_n) \\
&\leq 1 - (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\gamma)) \times (1-(n+1)^{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}\exp(-n\delta)) \\
&\leq \epsilon,
\end{aligned}
$$

where $\epsilon$ is the constant-type constraint on the type 1 error probability.

Hence, for the hypothesis testing problem (A.10), the acceptance region $\mathcal{A}_n$ satisfies the constant-type type 1 error probability constraint.

From [34], we know that the type 2 error exponent

$$
\theta(0,0,\epsilon) \leq \min_{\tilde{P}_{X_1X_2Y} \in \mathcal{L}_0} D\left(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right), \qquad \text{(A.11)}
$$

where

$$\mathcal{L}_0 = \left\{ \tilde{P}_{X_1X_2Y} : \tilde{P}_{X_1} = P_{X_{10}}, \tilde{P}_{X_2} = P_{X_{20}}, \tilde{P}_Y = P_{Y_0} \right\}.$$

On the other hand, we note that $P_{X_{10}X_{20}Y_0}$ was arbitrary as far as condition (A.4) is satisfied. Therefore, in the light of the definition of $\mathcal{H}_r$, we see that the infimum of the right-hand side in (A.11) over all possible internal points $P_{X_{10}X_{20}Y_0}$ satisfying (A.4) coincides with

$$\min_{\tilde{P}_{X_1X_2Y} \in \mathcal{H}_r} D\left( \tilde{P}_{X_1X_2Y} || Q_{X_1X_2Y} \right).$$

Thus (A.11) reduces to

$$\sigma(0,0,r) \leq \min_{\tilde{P}_{X_1X_2Y} \in \mathcal{H}_r} D\left( \tilde{P}_{X_1X_2Y} || Q_{X_1X_2Y} \right).$$

## A.2 Proof of Theorem 2.6

Now we simplify the upper bound in (2.45) in the following steps. First consider

$$\begin{aligned}
nR_1 &\geq H(M_1) \\
&\geq I(M_1; X_1^n) \\
&= \sum_{i=1}^{n} I(M_1; X_{1i} | X_1^{i-1}) \\
&= \sum_{i=1}^{n} I(M_1 X_1^{i-1}; X_{1i}) \\
&\overset{(a)}{=} \sum_{i=1}^{n} I(M_1 X_1^{i-1} X_2^{i-1}; X_{1i}) \\
&\overset{(b)}{=} \sum_{i=1}^{n} I(U_{1i}; X_{1i}),
\end{aligned}$$

where $(a)$ follows since $X_{1i} \leftrightarrow (M_1, X_1^{i-1}) \leftrightarrow X_2^{i-1}$ forms a Markov chain, which can be derived by the following:

$$\begin{aligned}
&(X_1^n, X_{1i}) \leftrightarrow X_1^{i-1} \leftrightarrow X_2^{i-1} \\
&\overset{(c)}{\Rightarrow} (M_1, X_{1i}) \leftrightarrow X_1^{i-1} \leftrightarrow X_2^{i-1} \\
&\overset{(d)}{\Rightarrow} X_{1i} \leftrightarrow (M_1, X_1^{i-1}) \leftrightarrow X_2^{i-1},
\end{aligned} \tag{A.12}$$

$(c)$ is true as $M_1$ is a function of $X_1^n$ and $(d)$ is true due to the weak union property of Markov chain [59]. $(b)$ is true by identifying $U_{1i} = (M_1, X_1^{i-1}, X_2^{i-1})$ and noting that $U_{1i} \leftrightarrow X_{1i} \leftrightarrow (X_{2i}, Y_i)$ forms a Markov chain as

$$\begin{aligned}
&(X_1^n, X_1^{i-1}, X_2^{i-1}) \leftrightarrow X_{1i} \leftrightarrow (X_{2i}, Y_i) \\
&\Rightarrow (M_1, X_1^{i-1}, X_2^{i-1}) \leftrightarrow X_{1i} \leftrightarrow (X_{2i}, Y_i).
\end{aligned}$$

Following similar steps as above, we have

$$nR_2 \overset{(e)}{\geq} \sum_{i=1}^{n} I(M_2 X_2^{i-1} Y^{i-1}; X_{2i}) \overset{(f)}{=} \sum_{i=1}^{n} I(U_{2i}; X_{2i}),$$

where $(e)$ follows since $Y^{i-1} \leftrightarrow (M_2, X_2^{i-1}) \leftrightarrow X_{2i}$; $(f)$ is true by identifying $U_{2i} = (M_2, X_2^{i-1}, Y^{i-1})$ and noting that $U_{2i} \leftrightarrow X_{2i} \leftrightarrow (X_{1i}, Y_i)$.

Finally, we consider

$$\begin{aligned}
H(Y^n|M_1M_2) &= \sum_{i=1}^{n} H(Y_i|M_1M_2Y^{i-1}) \\
&\geq \sum_{i=1}^{n} H(Y_i|M_1M_2Y^{i-1}X_1^{i-1}X_2^{i-1}) \\
&= \sum_{i=1}^{n} H(Y_i|U_{1i}U_{2i}).
\end{aligned}$$

Define the time-sharing random variable $Q$ to be the uniformly distributed over $[1:n]$ and

independent of $(M_1, M_2, X_1^n, X_2^n, Y^n)$, and identify $U_1 = (U_{1Q}, Q),\ U_2 = (U_{2Q}, Q, )\ X_1 = X_{1Q},\ X_2 = X_{2Q}$, and $Y = Y_Q$. Clearly, we have $U_1 \leftrightarrow X_1 \leftrightarrow (X_2, Y)$ and $U_2 \leftrightarrow X_2 \leftrightarrow (X_1, Y)$ forms three Markov chains. Hence we have shown

$$
\begin{aligned}
R_1 &\geq I(U_1; X_1), \\
R_2 &\geq I(U_2; X_2), \\
\lim_{\epsilon \to 0} \theta(R_1, R_2, \epsilon) &\leq H(Y) - H(Y|U_1U_2) = I(Y; U_1U_2),
\end{aligned}
$$

for some conditional PMF $P_{U_1|X_1}$ and $P_{U_2|X_2}$.

# Appendix B

# Appendix of Chapter 3

## B.1 Proof of the Markov chain $U_{2i} \leftrightarrow (U_{1i}, X_{2i}) \leftrightarrow (X_{1i}, Y_i)$

First, we need the following lemma introduced and proved in [19, Lemma 1].

**Lemma B.1.** [19] Let $A_1, A_2, B_1, B_2$ be the random variables with joint PMF $P_{A_1A_2B_1B_2} = P_{A_1B_1}P_{A_2B_2}$ and assume that $\{f^i\}_{i=1}^k, \{g^i\}_{i=1}^k$ are any collection of $P$-measurable mappings with domain structure given by:

$$\begin{aligned} &f^1(A_1, A_2); f^2(A_1, A_2, g^1); \cdots; f^k(A_1, A_2, g^1, \cdots, g^{k-1}), \\ &g^1(B_1, B_2, f^1); \cdots; g^k(B_1, B_2, f^1, \cdots, f^k). \end{aligned} \tag{B.1}$$

Then,

$$I(A_2; B_1 | A_1, B_2, f^1, f^2, \cdots, f^k, g^1, g^2, \cdots, g^k) = 0. \tag{B.2}$$

To prove the Markov chain $U_{2i} \leftrightarrow (U_{1i}, X_{2i}) \leftrightarrow (X_{1i}, Y_i)$, first we set

$$\begin{cases} A_1 := X_1^{i-1}, & B_1 := X_2^{i-1} \\ A_2 := X_{1i}^n, & B_2 := X_{2i}^n \end{cases} \tag{B.3}$$

Then according to Lemma B.1, we have

$$I(X_{1i}^n; X_2^{i-1}|X_1^{i-1}, X_{2i}^n, M_1) = 0, \tag{B.4}$$

where $M_1 = f^1(A_1, A_2)$. Thus, we have the following Markov chain,

$$(X_{1i}, X_{1(i+1)}^n) \leftrightarrow (X_1^{i-1}, X_{2i}^n, M_1) \leftrightarrow X_2^{i-1}. \tag{B.5}$$

As $M_2 = g^1(B_1, B_2, M_1)$, we have

$$X_{1i} \leftrightarrow (X_1^{i-1}, X_{2i}^n, M_1) \leftrightarrow M_2. \tag{B.6}$$

Since $Y_i \leftrightarrow (X_{1i}, X_{2i}) \leftrightarrow (X_1^{i-1}, X_{2(i+1)}^n, M_1, M_2)$, we can have

$$(X_{1i}, Y_i) \leftrightarrow (X_1^{i-1}, X_{2i}, X_{2(i+1)}^n, M_1) \leftrightarrow (M_2, Y^{i-1}), \tag{B.7}$$

i.e.

$$(X_{1i}, Y_i) \leftrightarrow (X_{2i}, U_{1i}) \leftrightarrow U_{2i}. \tag{B.8}$$

## B.2 Proof sketch of Theorem 3.4

In this appendix, we provide a proof sketch of Theorem 3.4.

In the following, $\eta > \eta' > \eta'' > \eta'''$ are given small numbers.

**Codebook generation.** Fix a joint distribution attaining the maximum in (3.11), which satisfies

$$P_{U_1 \cdots U_L | X_1 \cdots X_L Y} = P_{U_1|X_1} \prod_{l=2}^{L} P_{U_l|U_1 \cdots U_{l-1} X_l}.$$

Let

$$P_{U_1}(u_1) = \sum_{x_1} P_{X_1}(x_1) P_{U_1|X_1}(u_1|x_1),$$

and

$$\begin{aligned}
&P_{U_l|U_{l-1}\cdots U_1}(u_l|u_{l-1}\cdots u_1) \\
&= \sum_{x_l} P_{X_l|U_1\cdots U_{l-1}}(x_l|u_1,\cdots,u_{l-1}) \cdot P_{U_l|U_{l-1}\cdots U_1 X_l}(u_l|u_{l-1},\cdots,u_1,x_l)
\end{aligned}$$

for $l = 2, \cdots, L$.

Randomly and independently generate $\lfloor 2^{nR_1} \rfloor$ sequences $u_1^n(m_1)$, $m_1 \in \{1, \cdots, \lfloor 2^{nR_1} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_1}(u_{1i})$. For each $(u_1^n(m_1), \cdots, u_{l-1}^n(m_{l-1}))$, randomly and independently generate $\lfloor 2^{nR_l} \rfloor$ sequences $u_l^n(m_l)$, $m_l \in \{1, \cdots, \lfloor 2^{nR_l} \rfloor\}$ each according to $\prod_{i=1}^n P_{U_l|U_{l-1}\cdots U_1}(u_{li}|u_{(l-1)i}\cdots u_{1i})$ for $l = 2, \cdots, L$. These sequences constitute the codebook $c$, which is revealed to all terminals. We use $\mathcal{C}$ to denote the set of all possible codebooks.

**Encoding for terminal $\mathcal{X}_1$.** Given a sequence $x_1^n$, terminal $\mathcal{X}_1$ finds a $u_1^n(m_1)$ such that $(x_1^n, u_1^n(m_1)) \in T_{\eta'''}^{(n)}(X_1U_1)$, then it sends the index $m_1$ to both terminal $\mathcal{X}_2$ and $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends $0$.

**Encoding for terminal $\mathcal{X}_l$, $l = 2, \cdots, L$.** If at least one $0$ is received from terminals $\mathcal{X}_1, \cdots, \mathcal{X}_{l-1}$, terminal $\mathcal{X}_l$ sends $m_l = 0$ to terminal $\mathcal{Y}$. If $m_1 \neq 0, \cdots, m_{l-1} \neq 0$ are received from terminals $\mathcal{X}_1, \cdots, \mathcal{X}_{l-1}$, given $x_l^n$ and $(m_1 \cdots m_{l-1})$, terminal $\mathcal{X}_l$ finds a $u_l^n(m_l)$ such that $(u_1^n(m_1), \cdots, u_l^n(m_l), x_l^n) \in T_{\eta''}^{(n)}(U_1 \cdots U_l X_l)$ and sends the index $m_l$ to terminal $\mathcal{Y}$. If there is more than one such index, it sends the smallest one among them. If there is no such index, it sends $0$.

**Testing.** Upon receiving messages from terminal $\mathcal{X}_1, \cdots, \mathcal{X}_L$, terminal $\mathcal{Y}$ sets the accep-

tance region $\mathcal{A}_n$ for $H_0$ to

$$\mathcal{A}_n = \left\{(m_1, \cdots, m_L, y^n) : (u_1^n(m_1), \cdots, u_L^n(m_L), y^n) \in T_\eta^{(n)}(U_1 \cdots U_L Y)\right\}.$$

This implies, terminal $\mathcal{Y}$ decides $\hat{H} = H_0$ if and only if no 0 is received and $(u_1^n(m_1), \cdots, u_L^n(m_L), y^n) \in T_\eta^{(n)}(U_1 \cdots U_L Y)$.

**Analysis of two types of errors.** Terminal $\mathcal{Y}$ chooses $\hat{H} = H_1$ if and only if one or more of the following events occur:

$$\begin{aligned}
\varepsilon_1 &= \left\{(U_1^n(m_1), X_1^n) \notin T_{\eta'''}^{(n)}(U_1 X_1) \text{ for all } m_1 \in [1 : \lfloor 2^{nR_1} \rfloor]\right\}, \\
\varepsilon_l &= \Big\{(U_1^n(M_1), \cdots, U_{l-1}^n(M_{l-1}), U_l^n(m_l), X_l^n) \notin T_{\eta''}^{(n)}(U_1 \cdots U_l X_l) \\
&\qquad \text{for all } m_l \in [1 : \lfloor 2^{nR_l} \rfloor]\Big\},\ l = 2, \cdots, L, \\
\varepsilon_{L+1} &= \left\{(U_1^n(M_1), \cdots, U_L^n(M_L), Y^n) \notin T_\eta^{(n)}(U_1 \cdots U_L Y)\right\}.
\end{aligned}$$

Here, we can see that $\mathcal{A}_n^c = \varepsilon_1 \cup \cdots \cup \varepsilon_{L+1}$.

For any particular codebook $c \in \mathcal{C}$, we use $\alpha_{nc}$ and $\beta_{nc}$ to denote the type 1 and the type 2 error probabilities respectively. In the following, we will first compute the probabilities of two types of errors averaged over all possible codebooks:

$$\begin{aligned}
\mathbb{E}\{\alpha_{nc}\} &= \sum_{c \in \mathcal{C}} \alpha_{nc} \text{Pr}(c), \\
\mathbb{E}\{\beta_{nc}\} &= \sum_{c \in \mathcal{C}} \beta_{nc} \text{Pr}(c).
\end{aligned}$$

We will then argue that there exists a particular codebook $c^*$ that has the desired properties.

Following similar analysis as in Section 3.3.1, we can show that the error exponent stated in the theorem is achievable using this scheme.

## B.3 Proof of Theorem 3.8

In this appendix, to facilitate the presentation, we show a detailed proof for $L = 2$. The proof for the general $L$ is similar. Our proof follows a similar strategy as that in [34] and employs the "blowing-up" lemma [2].

First, we define

$$C_{m_1} = \{x_1^n \in \mathcal{X}_1^n : f_1(x_1^n) = m_1\},$$
$$D_{m_2|m_1} = \{x_2^n \in \mathcal{X}_2^n : f_2(x_2^n, m_1) = m_2\},$$
$$F_{m_1,m_2} = \{y^n \in \mathcal{Y}^n : \psi(m_1, m_2, y^n) = H_0\},$$

then we can write

$$\mathcal{A}_n = \bigcup_{m_1=1}^{||f_1||} \bigcup_{m_2=1}^{||f_2||} C_{m_1} \times D_{m_2|m_1} \times F_{m_1,m_2}. \tag{B.9}$$

And we can see that $C_{m_1}$s are pairwise disjoint and for fixed $m_1$, $D_{m_2|m_1}$s are pairwise disjoint for different $m_2$.

We have $P^n_{X_1X_2Y}(\mathcal{A}_n) \geq 1 - \epsilon$, then there exists an index $(m_{10}, m_{20})$ such that

$$P^n_{X_1X_2Y}(C_{m_{10}} \times D_{m_{20}|m_{10}} \times F_{m_{10},m_{20}}) \geq \frac{1-\epsilon}{||f_1|| \cdot ||f_2||}.$$

To simplify the notations, we let $C = C_{m_{10}}$, $D_{m_{20}|m_{10}} = D$ and $F_{m_{10},m_{20}} = F$. We can rewrite the equation above as

$$P^n_{X_1X_2Y}(C \times D \times F) \geq \exp(-n\delta_n) \tag{B.10}$$

where $\delta_n = -\frac{1}{n}\log(1-\epsilon) + \frac{1}{n}\log(||f_1|| \cdot ||f_2||)$ and $\delta_n \to 0$ by (2.17) and (2.18). (B.10) implies that

$$P^n_{X_1}(C) \geq \exp(-n\delta_n),\ P^n_{X_2}(D) \geq \exp(-n\delta_n),$$

$$P_Y^n(F) \geq \exp(-n\delta_n).$$

Define the Hamming k-neighborhood $\Gamma^k C$ of $C$ by

$$\Gamma^k C = \{z^n \in \mathcal{X}_1^n : \exists x_1^n \in C, \text{ s.t. } d(x_1^n, u^n) \leq k\}.$$

Using Blowing-up lemma [10], there exists sequences $k_n$ and $\gamma_n$ satisfying $k_n/n \to 0$ and $\gamma_n \to 0$, and such that

$$P_{X_1}^n(\Gamma^{k_n} C) \geq 1 - \gamma_n, \tag{B.11}$$

$$P_{X_2}^n(\Gamma^{k_n} D) \geq 1 - \gamma_n, \tag{B.12}$$

$$P_Y^n(\Gamma^{k_n} F) \geq 1 - \gamma_n. \tag{B.13}$$

Furthermore, $k_n$ and $\gamma_n$ depend only on $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{Y}$ and $\gamma_n$. In the following, we will use $k$ instead of $k_n$. (B.11), (B.12) and (B.13) hold true if we replace $P$ by $\tilde{P}$ where $\tilde{P}_{X_1X_2Y}$ satisfies the marginal constraints $\tilde{P}_{X_1} = P_{X_1}, \tilde{P}_{X_2} = P_{X_2}$, and $\tilde{P}_Y = P_Y$. Moreover, via simple derivations we have

$$\tilde{P}_{X_1X_2Y}^n(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \geq 1 - 3\gamma_n. \tag{B.14}$$

As $\tilde{T}_\eta^{(n)}(X_1X_2Y)$ is the set of $(\tilde{P}_{X_1X_2Y}, \eta)$-typical sequences, then $\tilde{P}_{X_1X_2Y}^n(\tilde{T}_\eta^{(n)}(X_1X_2Y)) \geq 1 - \eta_n$, where $\eta_n$ is a small number such that $\eta_n/n \to 0$ as $n \to \infty$. Hence, for all sufficiently large $n$, we obtain

$$\tilde{P}_{X_1X_2Y}^n\big((\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \tilde{T}_\eta^{(n)}(X_1X_2Y)\big) \geq \frac{1}{2}. \tag{B.15}$$

By the definition of $\tilde{T}_{\eta}^{(n)}(X_1X_2Y)$, we have the following decomposition:

$$\tilde{T}_{\eta}^{(n)}(X_1X_2Y) = \bigcup_{\substack{\hat{P}_{X_1X_2Y} \in \mathcal{P}^n(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}) \\ |\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y}}} \hat{T}^{(n)}(X_1X_2Y).$$

Given the fact of equiprobable elements of a given $\hat{T}^{(n)}(X_1X_2Y)$, (B.15) can be rewritten as

$$\sum_{\substack{\hat{P}_{X_1X_2Y} \in \mathcal{P}^n(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}) \\ |\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y}}} \tilde{P}^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y)) \frac{|(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)|}{|\hat{T}^{(n)}(X_1X_2Y)|} \geq \frac{1}{2}.$$

Hence, there exists a type $\hat{P}_{X_1X_2Y} \in \mathcal{P}^n(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y})$ satisfying $|\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}| \leq \eta \tilde{P}_{X_1X_2Y}$ and such that

$$\frac{|(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)|}{|\hat{T}^{(n)}(X_1X_2Y)|} \geq \frac{1}{2}.$$

Since pairs $(x_1^n, x_2^n, y^n)$ of the same type are also equiprobable under $Q^n_{X_1X_2Y}$, we conclude that for the previous type $\hat{P}_{X_1X_2Y}$,

$$\begin{aligned}
&Q^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \\
&\geq Q^n_{X_1X_2Y}((\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)) \\
&= Q^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y)) \frac{|(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \cap \hat{T}^{(n)}(X_1X_2Y)|}{|\hat{T}^{(n)}(X_1X_2Y)|} \\
&\geq \frac{1}{2} Q^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y)).
\end{aligned} \tag{B.16}$$

Consider an arbitrary element $(z^n, v^n, w^n)$ of $\Gamma^k C \times \Gamma^k D \times \Gamma^k F$. By definition of $\Gamma^k$, there exists at least one element $(x_1^n, x_2^n, y^n) \in C \times D \times F$ such that $(x_{1i}, x_{2i}, y_i)$ differs from

$(z_i, v_i, w_i)$ for at most $3k$ values of $i$. We thus have

$$\begin{aligned} Q^n_{X_1X_2Y}(z^n, v^n, w^n) &= \prod_{i=1}^{n} Q_{X_1X_2Y}(z_i, v_i, w_i) \\ &\leq \rho^{-3k} \prod_{i=1}^{n} Q_{X_1X_2Y}(x_{1i}, x_{2i}, y_i) \\ &= \rho^{-3k} Q_{X_1X_2Y}(x_1^n, x_2^n, y^n), \end{aligned} \tag{B.17}$$

where $\rho = \min_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, y \in \mathcal{Y}} Q_{X_1X_2Y}(x_1, x_2, y) > 0$. As $(z^n, v^n, w^n)$ ranges over $\Gamma^k C \times \Gamma^k D \times \Gamma^k F$, each element $(x_1^n, x_2^n, y^n)$ of $C \times D \times F$ will be selected at most $|\Gamma^k(x_1^n)| \cdot |\Gamma^k(x_2^n)| \cdot |\Gamma^k(y^n)|$ times. By virtue of this, (B.17) yields

$$\begin{aligned} &Q^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \\ &\leq \rho^{-3k} |\Gamma^k C(x_1^n)| \cdot |\Gamma^k(x_2^n)| \cdot |\Gamma^k(y^n)| Q^n_{X_1X_2Y}(C \times D \times F). \end{aligned}$$

From [10], we have the upper bound

$$|\Gamma^k(x_1^n)| \leq \exp\left[ \left( n \left( H\left(\frac{k}{n}\right) + \frac{k}{n} \log |\mathcal{X}_1| \right) \right) \right.$$

Thus, we can write

$$Q^n_{X_1X_2Y}(\Gamma^k C \times \Gamma^k D \times \Gamma^k F) \leq \exp(n\xi_n) Q^n_{X_1X_2Y}(C \times D \times F), \tag{B.18}$$

where

$$\xi_n = 3H\left(\frac{k}{n}\right) + \frac{k}{n} \log(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|) - \frac{3k}{n} \log \rho \to 0.$$

Finally, combining (B.16) and (B.18) with the upper bound on $Q^n_{X_1X_2Y}(\hat{T}^{(n)}(X_1X_2Y))$, we have

$$Q^n_{X_1X_2Y}(C \times D \times F)$$

$$
\begin{aligned}
&\geq \frac{1}{2}\exp(-n\xi_n)Q^n_{X_1X_2Y}\left(\hat{T}^{(n)}(X_1X_2Y)\right)\\
&\geq \frac{(n+1)^{-|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|}}{2}\exp\left(-n\left(D\left(\hat{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right)+\xi_n\right)\right)\\
&\geq \exp\left(-n\left(D\left(\hat{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right)+\varsigma_n\right)\right),
\end{aligned}
$$

where $\varsigma_n \to 0$.

$$\varsigma_n = \varsigma_n(\rho, \eta, M_1, M_2, |\mathcal{X}_1|, |\mathcal{X}_2|, |\mathcal{Y}|) \to 0.$$

Since $D(\tilde{P}_{X_1X_2Y}||\tilde{Q}_{X_1X_2Y})$ is uniformly continuous, we can find a sequence $\mu_n = \mu_n(\rho, |\mathcal{X}_1|, |\mathcal{X}_2|, |\mathcal{Y}|) \to 0$ such that

$$
\begin{aligned}
&\left|\hat{P}_{X_1X_2Y} - \tilde{P}_{X_1X_2Y}\right| \leq \eta\tilde{P}_{X_1X_2Y}\\
&\quad\Rightarrow \left|D\left(\hat{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right) - D\left(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right)\right| \leq \mu_n.
\end{aligned}
$$

Hence,

$$Q^n_{X_1X_2Y}(C\times D\times F) \geq \exp\left(-n\left(D\left(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right)+\varsigma_n+\mu_n\right)\right), \quad \text{(B.19)}$$

and consequently

$$\theta(+0,+0,\epsilon) \leq \min_{\tilde{P}_{X_1X_2Y}\in\mathcal{L}_0} D\left(\tilde{P}_{X_1X_2Y}||Q_{X_1X_2Y}\right).$$

Since $\tilde{P}_{X_1X_2Y}$ satisfies the appropriate marginal constraints, the proof is complete.

# Appendix C

# Appendix of Chapter 4

## C.1 Proof of Lemma 4.1

In this appendix, we provide the proof for Lemma 4.1. To simplify the presentation, we show that Lemma 4.1 exists when $\mathcal{Y} = \phi$ and $\mathcal{X} = \{a, b\}$.

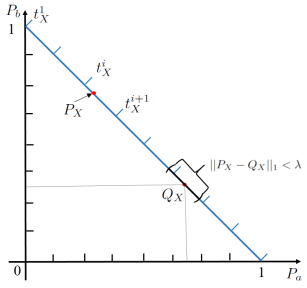



Figure C.1: Example

*Proof.* (1) First, we notice that for two different types $t_{X_1}$, $t_{X_2}$,

$$||t_{X_1} - t_{X_2}||_1 \geq \frac{2}{n}. \tag{C.1}$$

Hence, we can number all the types $t_X \in \Lambda_n(\mathcal{X})$ as $t_X^i$, $i \in \{1, \cdots, (n+1)^{|\mathcal{X}|}\}$ with $||t_X^i - t_X^{i+1}||_1 = \frac{2}{n}$. Then, the space $\mathcal{P}_X$ is cut into cells by $t_X^i$, label each cell by $C_{t_X^i}^{t_X^{i+1}}$.

(2) Then, we show that for each $P_X \in \mathcal{P}_X \cap \Pi$, we have $P_X \in \bigcup_{t_X \in \Lambda_n^{\Pi}(\mathcal{X})} \mathcal{N}_{t_X}$.

- If $\exists\, t_X \in \Lambda_n^{\Pi}(\mathcal{X})$, such that $P_X = t_X$, it is easy to see $P_X \in \bigcup_{t_X \in \Lambda_n^{\Pi}(\mathcal{X})} \mathcal{N}_{t_X}$.
- If $P_X \neq t_X$, $\forall t_X \in \Lambda_n^{\Pi}(\mathcal{X})$, we will show that there exists a $t_X \in \Lambda_n^{\Pi}(\mathcal{X})$ such that $P_X \in \mathcal{N}_{t_X}$.

  Suppose $\forall\, t_X \in \Lambda_n^{\Pi}(\mathcal{X}),\ P_X \notin \mathcal{N}_{t_X}$, i.e.

$$||t_X - P_X||_1 > \frac{1}{n},\ \forall t_X \in \Lambda_n^{\Pi}(\mathcal{X}). \tag{C.2}$$

  As $\exists\, i \in \{1, \cdots, (n+1)^{|\mathcal{X}|}\}$ such that $P_X \in C_{t_X^i}^{t_X^{i+1}}$, where at least one of $\{t_X^i, t_X^{i+1}\}$ belongs to $\Lambda_n^{\Pi}(\mathcal{X})$, then

$$\begin{aligned}
||t_X^i - t_X^{i+1}||_1 &= ||t_X^i - P_X + P_X - t_X^{i+1}||_1 \\
&= \sum_{x_j \in \mathcal{X}} |t_X^i(x_j) - P_X(x_j) + P_X(x_j) - t_X^{i+1}(x_j)| \\
&= \sum_{x_j \in \mathcal{X}} |t_X^i(x_j) - P_X(x_j)| + |P_X(x_j) - t_X^{i+1}(x_j)| \\
&= ||t_X^i - P_X||_1 + ||P_X - t_X^{i+1}||_1 \\
&> \frac{1}{n} + \frac{1}{n} = \frac{2}{n},
\end{aligned}$$

  which contradicts with $||t_X^i - t_X^{i+1}||_1 = \frac{2}{n}$. Hence there must exist a $t_X \in \Lambda_n^{\Pi}(\mathcal{X})$ such that $P_X \in \mathcal{N}_{t_X}$.

□

## C.2 Proof of (4.17)

Given $t_{XY} \in \Lambda_n^{\Pi}(\mathcal{X}\mathcal{Y})$ and $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$, then for any $\tilde{P}_{XY} \in \mathcal{P}_{XY}$, we have

$$f_{t_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right) \leq f_{\tilde{t}_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right) + \delta(\zeta). \tag{C.3}$$

*Proof.* $\tilde{t}_{XY} \in \mathcal{N}_{t_{XY}}$ means that $||t_{XY} - \tilde{t}_{XY}||_1 \leq \zeta$, i.e. $t_{XY}(x,y) - \zeta \leq \tilde{t}_{XY}(x,y) \leq$

$t_{XY}(x,y)+\zeta$ for each $(x,y)\in\mathcal{X}\times\mathcal{Y}$.

$$
\begin{aligned}
f_{t_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right) = & \min_{\substack{\hat{P}_{XY}\\ \hat{P}_X=\check{X}^{(n)}(x^n)\\ \hat{P}_Y=\check{Y}^{(n)}(y^n)}} D\left(\hat{P}_{XY}||t_{XY}\right)\\
= & \min_{\substack{\hat{P}_{XY}\\ \hat{P}_X=\check{X}^{(n)}(x^n)\\ \hat{P}_Y=\check{Y}^{(n)}(y^n)}} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}_{XY}(x,y)\log\frac{\hat{P}_{XY}(x,y)}{\tilde{t}_{XY}(x,y)}\\
& + \min_{\substack{\hat{P}_{XY}\\ \hat{P}_X=\check{X}^{(n)}(x^n)\\ \hat{P}_Y=\check{Y}^{(n)}(y^n)}} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}_{XY}(x,y)\log\frac{\tilde{t}_{XY}(x,y)}{t_{XY}(x,y)}\\
= & f_{\tilde{t}_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right)\\
& + \min_{\substack{\hat{P}_{XY}\\ \hat{P}_X=\check{X}^{(n)}(x^n)\\ \hat{P}_Y=\check{Y}^{(n)}(y^n)}} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}_{XY}(x,y)\log\frac{\tilde{t}_{XY}(x,y)}{t_{XY}(x,y)}\\
\leq & f_{\tilde{t}_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right)\\
& + \min_{\substack{\hat{P}_{XY}\\ \hat{P}_X=\check{X}^{(n)}(x^n)\\ \hat{P}_Y=\check{Y}^{(n)}(y^n)}} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}_{XY}(x,y)\log\frac{t_{XY}(x,y)+\zeta}{t_{XY}(x,y)}\\
\overset{(a)}{\leq} & f_{\tilde{t}_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right)\\
& + \min_{\substack{\hat{P}_{XY}\\ \hat{P}_X=\check{X}^{(n)}(x^n)\\ \hat{P}_Y=\check{Y}^{(n)}(y^n)}} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}_{XY}(x,y)\frac{\zeta}{t_{XY}(x,y)}\\
\overset{(b)}{\leq} & f_{\tilde{t}_{XY}}\left(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)\right)+\delta(\zeta),
\end{aligned}
$$

where $(a)$ is true due to the inequality $\log(1+c\cdot x)\leq c\cdot x,$ for $x>0,\ c>0$; and $(b)$ is true as $\frac{\hat{P}_{XY}}{t_{XY}}<\infty$: if $\exists(x_0,y_0)\in\mathcal{X}\times\mathcal{Y}$, such that $t_{XY}(x_0,y_0)=0$, then $\tilde{t}_{XY}(x_0,y_0)<\frac{1}{n}$

and we can find $\hat{P}_{XY}(x_0, y_0) = 0$.

Hence, we can get $f_{t_{XY}}(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)) \leq f_{\tilde{t}_{XY}}(\check{X}^{(n)}(x^n)\check{Y}^{(n)}(y^n)) + \delta(\zeta)$ and $\delta(\zeta)$ is a function of $\zeta$ and $\delta(\zeta) \sim O(\frac{1}{n})$.

□

## C.3 Proof of (4.40)

From the distribution we draw $U_{S_1}^n$ and $V_{S_2}^n$, we have the Markov chain

$$V_{S_2}^n \leftrightarrow Y^n \leftrightarrow X^n \leftrightarrow U_{S_1}^n.$$

As $(v^n, x^n, y^n) \in T_{\epsilon''}^n$ and from the Markov chain we know that

$$\Pr\{U_{S_1}^n = u^n | V_{S_2}^n = v^n, X^n = x^n, Y^n = y^n\} = \Pr\{U_{S_1}^n = u^n | x^n\}.$$

By the covering lemma, $\Pr\{(x^n, U^n) \in T_{\epsilon''}^n\}$ converges to 1 as $n \to \infty$, that is $\Pr\{U_{S_1}^n = u^n | x^n\}$ satisfies the first condition in the Markov lemma. Then we show that it also satisfies the second condition in the Markov lemma.

For all $u^n \in T_{\epsilon''}^n(U|x^n)$,

$$\begin{aligned}
&\Pr\{U_{S_1}^n = u^n | X^n = x^n\} \\
&= \Pr\{U_{S_1}^n = u^n, U_{S_1}^n \in T_{\epsilon''}^n(U|x^n) | X^n = x^n\} \\
&= \Pr\{U_{S_1}^n \in T_{\epsilon''}^n(U|x^n) | X^n = x^n\} \\
&\quad \times \Pr\{U_{S_1}^n = u^n | U_{S_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n\} \\
&\leq \Pr\{U_{S_1}^n = u^n | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n\} \\
&= \sum_{m_1} \Pr\{U_{S_1}^n = u^n, S_1 = s_1 | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n\} \\
&= \sum_{m_1} \Pr\{U_{S_1}^n = u^n | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n, S_1 = s_1\}
\end{aligned}$$

$$
\begin{aligned}
&\quad \times \Pr\{S_1 = s_1 | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n\} \\
&\overset{(a)}{=} \sum_{m_1} \Pr\{U_{s_1}^n = u^n | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n)\} \\
&\quad \times \Pr\{S_1 = s_1 | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n\} \\
&\overset{(b)}{\leq} \sum_{m_1} \Pr\{S_1 = s_1 | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n\} \times 2^{-n(H(U|X)-\delta(\epsilon''))} \\
&= 2^{-n(H(U|X)-\delta(\epsilon''))},
\end{aligned}
$$

where (a) follows since

$$
\begin{aligned}
&\Pr\{U_{S_1}^n = u^n | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n), X^n = x^n, S_1 = s_1\} \\
&= \Pr\{U_{S_1}^n = u^n | U_{s_1}^n \in T_{\epsilon''}^n(U|X^n = x^n), X^n = x^n, S_1 = s_1\} \\
&= \Pr\{U_{S_1}^n = u^n | U_{s_1}^n \in T_{\epsilon''}^n(U|x^n)\}.
\end{aligned}
$$

(b) follows from properties of typical sequences. Similarly, we can also prove that for every $u^n \in T_{\epsilon''}^n(U|x^n)$ and $n$ sufficiently large,

$$
\Pr\{U_{s_1}^n = u^n | X^n = x^n\} \geq (1-\epsilon'')2^{-n(H(U|X)+\delta(\epsilon''))}.
$$

Hence, this satisfies the second condition in the Markov Lemma. By the Markov lemma, we have $(U_{s_1}^n, V_{s_1}^n, X^n, Y^n) \in T_{\epsilon'}^n$.

# Bibliography


[1] R. Ahlswede and I. Csiszár. Hypothesis testing with communication constraints. *IEEE Trans. Inform. Theory*, 32:533–542, July 1986.

[2] R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 34(2):157–177, 1976.

[3] R. Ahlswede and J. Körner. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inform. Theory*, 21:629–637, Nov. 1975.

[4] E. Akofor and B. Chen. Interactive distributed detection: Architecture and performance analysis. *IEEE Trans. Inform. Theory*, 60(10):6456–6473, Oct. 2014.

[5] T. Berger. Decentralized estimation and decision theory. In *Proc. IEEE Inform. Theory Workshop*, Mt. Kisco, NY, Sept. 1979.

[6] F. S. Cattivelli and A. H. Sayed. Distributed detection over adaptive networks using diffusion adaptation. *IEEE Trans. Signal Processing*, 59(5):1917–1932, May 2011.

[7] J-F. Chamberland and V. V. Veeravalli. Decentralized detection in sensor networks. *IEEE Trans. Signal Processing*, 51(2):407–416, 2003.

[8] H. Chen, B. Chen, and P. K. Varshney. A new framework for distributed detection with conditionally dependent observations. *IEEE Trans. Signal Processing*, 60(3):1409–1419, March 2012.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theroy*. Wiley, New York, 2005.

[10] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1982.

[11] A. El Gamal and Y. Kim. *Network Information Theory*. Cambridge Unversity Press, Cambridge, UK, 2011.

[12] M. Falahatgar, A. Jafarpour, A. Orlitsky, V. Pichapathi, and A. T. Suresh. Faster algorithms for testing under conditional sampling. *ArXiv e-prints*, Apr. 2015.

[13] T. S. Han. Hypothesis testing with multiterminal data compression. *IEEE Trans. Inform. Theory*, 33:759–772, Nov. 1987.

[14] T. S. Han and S. Amari. Parameter estimation with multiterminal data compression. *IEEE Trans. Inform. Theory*, 41(6):1802–1833, Nov. 1995.

[15] T. S. Han and S. Amari. Statistical inference under multiterminal data compression. *IEEE Trans. Inform. Theory*, 44(6):2300–2324, Oct. 1998.

[16] T. S. Han and K. Kobayashi. Exponential-type error probabilities for multiterminal hypothesis testing. *IEEE Trans. Inform. Theory*, 35(1):2–14, Jan. 1989.

[17] Z. R. Hesabi, Z. Tari, A. Goscinski, A. Fahad, I. Khalil, and C. Queiroz. *Data Summarization Techniques for Big Data—A Survey*, pages 1109–1152. Springer New York, New York, NY, 2015.

[18] B. Kailkhura, S. Brahma, Y. S Han, and P. K Varshney. Optimal distributed detection in the presence of Byzantines. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 2925–2929, 2013.

[19] A Kaspi. Two-way source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, 31(6):735–740, 1985.

[20] G. Katz, P. Piantanida, R. Couillet, and M. Debbah. On the necessity of binning for the distributed hypothesis testing problem. In *Proc. IEEE Intl. Symposium on Inform. Theory*, Hongkong, China, Jun. 2015.

[21] G. Katz, P. Piantanida, and M. Debbah. Collaborative distributed hypothesis testing. *arXiv preprint arXiv:1604.01292*, 2016.

[22] J. Konečnỳ. Stochastic, distributed and federated optimization for machine learning. *arXiv preprint arXiv:1707.01155*, 2017.

[23] J. D. Lee, Y. Sun, Q. Liu, and J. E. Taylor. Communication-efficient sparse regression: a one-shot approach. *ArXiv e-prints*, Mar. 2015.

[24] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[25] M. Mhanna and P. Piantanida. On secure distributed hypothesis testing. In *Proc. IEEE Intl. Symposium on Inform. Theory*, Hongkong, China, Jun. 2015.

[26] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrák, and A. Krause. Lazier than lazy greedy. In *AAAI*, pages 1812–1818, 2015.

[27] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[28] R. Niu, B. Chen, and P. K. Varshney. Fusion of decisions transmitted over rayleigh fading channels in wireless sensor networks. *IEEE Trans. Signal Processing*, 54(3):1018–1027, 2006.

[29] F. Penna and S. Stanczak. Decentralized eigenvalue algorithms for distributed signal detection in wireless networks. *IEEE Trans. Signal Processing*, 63(2):427–440, Jan 2015.

[30] Z. Quan, W. Ma, S. Cui, and A.H. Sayed. Optimal linear fusion for distributed detection via semidefinite programming. *Signal Processing, IEEE Transactions on*, 58(4):2431–2436, Apr. 2010.

[31] M. S. Rahman and A. B. Wagner. The optimality of binning for distributed hypothesis testing. *IEEE Trans. Inform. Theory*, 58(10):6282–6303, Oct. 2012.

[32] J. B. Rhim, L. R. Varshney, and V. K. Goyal. Quantization of prior probabilities for collaborative distributed hypothesis testing. *IEEE Trans. Signal Processing*, 60(9):4537–4550, Sep. 2012.

[33] A. K. Sahu, D. Jakovetic, and S. Kar. Communication optimality trade-offs for distributed estimation. *arXiv preprint arXiv:1801.04050*, 2018.

[34] H. M. H. Shalaby and A. Papamarcou. Multiterminal detection with zero-rate data compression. *IEEE Trans. Inform. Theory*, 38(2):254–267, Mar. 1992.

[35] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proc. Intl. Conf. on Machine Learning*, Beijing, China, Jun. 2014.

[36] H. Shimokawa, T. S. Han, and S. Amari. Error bound of hypothesis testing with data compression. In *Proc. IEEE Intl. Symposium on Inform. Theory*, page 29, Trondheim, Norway, Jun. 1994.

[37] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, 19(4):471–480, July 1973.

[38] B. Szabo and H. van Zanten. Adaptive distributed methods under communication constraints. *arXiv preprint arXiv:1804.00864*, 2018.

[39] C. Tian and J. Chen. Successive refinement for hypothesis testing and lossless one-helper problem. *IEEE Trans. Inform. Theory*, 54(10):4666–4681, Oct. 2008.

[40] J. Tsitsiklis. Decentralized detection. In H. V. Poor and J. B. Thomas, editors, *Advances in Statistical Signal Processing, vol. 2—Signal Detection*. JAI, New York, 1990.

[41] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli. Universal and composite hypothesis testing via mismatched divergence. *IEEE Trans. Inform. Theory*, 57(3):1587–1603, Mar. 2011.

[42] J. Unnikrishnan and V. V. Veeravalli. Decentralized detection with correlated observations. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, pages 381–385. IEEE, 2007.

[43] G. Valiant, P. Valiant, A. Bhattacharyya, E. Fischer, and R. Rubinfeld. Instance-by-instance optimal identity testing. In *Electronic Colloquium on Computational Complexity*, volume 20, pages 1–1, 2013.

[44] P. K. Varshney. *Distributd Detection and Data Fusion*. Springer-Verlag, New York, 1996.

[45] R. Viswanathan and P. K. Varshney. Distributed detection with multiple sensors: Part I-fundamentals. *Proceedings of the IEEE*, 85(1):54–63, Jan 1997.

[46] Peter W., Peter F. S., and Rick S. B. The good, bad, and ugly: Distributed detection of a known signal in dependent Gaussian noise. *IEEE Trans. Signal Processing*, 48(12):3266–3279, 2000.

[47] Y. Xiang and Y. Kim. Interactive hypothesis testing with communication constraints. In *Proc. Allerton Conf. on Communication, Control, and Computing*, pages 1065–1072, Montecello, IL, Oct. 2012.

[48] Y. Xiang and Y. Kim. Interactive hypothesis testing against independence. In *Proc. IEEE Intl. Symposium on Inform. Theory*, pages 2840–2844, Istanbul, Turkey, Jul. 2013.

[49] P. Yang, B. Chen, H. Chen, and P. K. Varshney. Tandem distributed detection with conditionally dependent observations. In *Proc. Intl. Conf. on Information Fusion*, pages 1808–1813. IEEE, 2012.

[50] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, pages 2328–2336, Stateline, NV, Dec. 2013.

[51] Y. Zhang, S. Tirthapura, and G. Cormode. Learning graphical models from a distributed stream. *arXiv preprint arXiv:1710.02103*, 2017.

[52] W. Zhao and L. Lai. Distributed identity testing with data compression. *IEEE Trans. Inform. Theory*. Submitted in Aug. 2017.

[53] W. Zhao and L. Lai. Distributed testing with cascaded encoders. *IEEE Trans. Inform. Theory*. to appear, 2018.

[54] W. Zhao and L. Lai. Distributed testing against independence with multiple terminals. In *Proc. Allerton Conf. on Communication, Control, and Computing*, pages 1246–1251, Montecello, IL, Oct. 2014.

[55] W. Zhao and L. Lai. Distributed testing against independence with conferencing encoders. In *Proc. IEEE Inform. Theory Workshop*, Jeju Island, Korea, Oct. 2015.

[56] W. Zhao and L. Lai. Distributed testing with zero-rate compression. In *Proc. IEEE Intl. Symposium on Inform. Theory*, Hongkong, China, Jun. 2015.

[57] W. Zhao and L. Lai. Distributed detection with vector quantizer. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):105–119, Jun. 2016.

[58] W. Zhao and L. Lai. Distributed identity testing with zero-rate compression. In *Proc. IEEE Intl. Symposium on Inform. Theory*, pages 3135–3139. IEEE, 2017.

[59] F. Zhu and B. Chen. On the sum capacity of the discrete memoryless interference channel with one-sided weak interference and mixed interference. In *Proc. IEEE Intl. Symposium on Inform. Theory*, pages 2271–2275, Cambridge, MA, Jul. 2012.

[60] H. Zhu, G. B. Giannakis, and A. Cano. Distributed in-network channel decoding. *IEEE Trans. Signal Processing*, 57(10):3970–3983, Oct. 2009.

[61] Y. Zhu and J. Lafferty. Distributed nonparametric regression under communication constraints. *arXiv preprint arXiv:1803.01302*, 2018.

[62] J. Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Trans. Inform. Theory*, 34(2):278–286, Mar. 1988.