Coding for Data Analytics: New Information Distances

by

Ahmad Sajedi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Distance plays a vital role in many applications of data analytics. For instance, in image retrieval, organization, and management, one needs a proper similarity distance to measure the perceptual similarity between any two images X and Y. Once such a similarity distance is defined for any two images, it could be used to retrieve images in a database that are perceptually similar to a query image according to the similarity distance in image retrieval, and to organize images into different groups according to their mutual similarity in image management. Likewise, in data clustering and bioinformatics, the notion of distance also plays a dominant role.

The concept of distance between any two data objects X and Y (continuous or discrete) is addressed from the perspective of Shannon information theory. Consider a coding paradigm where X and Y are encoded into a sequence of coded bits specifying a codeword (or method) which would, in turn, convert Y into $\hat{X}$, and X into $\hat{Y}$ such that both the distortion between X and $\hat{X}$ and the distortion between Y and $\hat{Y}$ are less than or equal to a prescribed threshold D. To have a universality to some extent, we consider a class $\mathcal{C}$ of coding schemes within the coding paradigm. Given a class $\mathcal{C}$, the information distance $R_{\mathcal{C}}(X, Y, D)$ between X and Y at the distortion level D is defined as the smallest number of coded bits afforded by coding schemes from $\mathcal{C}$. $R_{\mathcal{C}}(X, Y, D)$ is shown to be indeed a pseudo distance in some sense; it is further characterized or bounded. New information distance is defined for all stationary sources with discrete alphabets and Independent and Identically Distributed (IID) sources with continuous alphabets such that the distortion level is small. For example, a pseudo distance for memoryless jointly Gaussian sources is presented when the distortion level is small or the distortion level is less than or equal to a special term which is a function of statistical properties of the sources, such as variances.

When $\mathcal{C}$ is the class of so-called separately precoded broadcast codes, it is shown that for any Discrete Memoryless Source (DMS) X and Y, $R_{\mathcal{C}}(X, Y, D)$ is equal to the maximum of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information. In the general case where $\mathcal{C}$ consists of all codes within the coding paradigm, upper and lower bounds to $R_{\mathcal{C}}(X, Y, D)$ are established, and are further shown to be tight when X and Y are IID pair source, for example memoryless jointly Gaussian or memoryless Doubly Symmetric Binary Source (DSBS)s. The distance $R_{\mathcal{C}}(X, Y, D)$ generalizes the notion of information distance defined within the framework of Kolmogorov complexity. In contrast to other information distances in the literature, this information distance is also applicable to both discrete and continuous-valued data.

Acknowledgements

I would like to thank my supervisor, Prof. En-hui Yang, for his guidance and support. He is a very good exemplar of the researcher and always inspires me to think more logically and precisely. I have learned a lot from him and knowledge in the field. Professor Yang always took the extra step to show his care to my current and future success.

I would like to thank Prof. Mohamed Oussama Damen and Prof. Amir Khandani for being the readers of this thesis and giving me valuable comments and suggestions. I wish to thank Prof. Young Han Kim from the University of California, San Diego, and Prof. Amin Aminzade Gohari from the Sharif University of Technology for their great advice and help in the project.

I would like to especially thank my friends and colleagues who have helped and inspired me to enjoy being in Waterloo. My deepest thanks go to Dr. Hossam Amer, Mr. Ahmed Hussein Salamah, Mr. Masoud Kavian, and Mr. Mahdi Abbasi Azad.

Finally, and the most important, I would like to express my deep appreciation to my wife, Mrs. Haniye Abdi, and my parents, Mr. Nematallah Sajedi and Mrs. Ashraf Afshari for their endless support, patience, understanding and their true love.

Dedication

to my mother Ashraf and my father Nematallah

&

my dear love Haniye

Table of Contents

List of Figures

List of Abbreviations

AEP Asymptotic Equipartition Property 5, 9

DMS Discrete Memoryless Source iii, 11, 14, 19, 47, 50, 57

DSBS Doubly Symmetric Binary Source iii, 3, 33–35, 57, 59

IID Independent and Identically Distributed iii, 3–5, 8, 22, 33–35, 38, 41, 43–46, 57–59

LLN Law of Large Number 8, 9

PMF Probability Mass Function 14, 16, 19, 29, 51, 53

SMID Set Mapping Induced similarity Distance 2, 58

Chapter 1

Introduction

1.1 Research Motivations and Problem Description

Distance plays an important role in many applications of data analytics. For example, in image retrieval, organization, and management, it is so crucial to define a suitable similarity distance such that the perceptual similarity between any two images X and Y can be measured. Once such a similarity distance is defined for any two images, it could be used to retrieve images in a database which are perceptually similar to a query image according to the similarity distance in image retrieval [2], and to organize images into different groups according to their mutual similarity in image management [3]. Likewise, in data clustering and bioinformatics, the notion of distance also plays a dominant role [4].

In the literature of image retrieval [2], a typical approach to determining a perceptual distance between two images is to first extract features from each image, then derive a signature of each image from its respectively extracted features, and finally determine the perceptual distance based on their respective signatures. Euclidean distance, Hausdorff distance, Kullback-Leibler divergence, etc. have all been used as a distance between signatures [2]. The variation in feature extraction, signature derivation, and distance between signatures leads to many different image perceptual distances. In general, however, as one moves from original images to features to signatures, the notion of distance becomes less intuitive and is increasingly disconnected from the original images.

To reduce this issue, a different approach was taken recently in [5]. The paper [5] first expanded each image X conceptually into a set $\phi(X)$ of images, which may contain images perceptually similar to X, and then defined the perceptual distance between X

and Y as the smallest average distortion per pixel between any pair of images, one from $\phi(X)$ and the other from $\phi(Y)$. The resulting distance is dubbed Set Mapping Induced similarity Distance (SMID) and denoted by $d_\phi(X, Y)$. It was demonstrated in [5] that when compared with other standard perceptual distances reported in the literature [6, 7, 8, 9, 10, 11, 12, 13], SMID indeed shows better discriminating power on image similarity. An interesting property relevant to our discussion in this thesis is that the optimization solution in SMID $d_\phi(X, Y)$ gives a method which converts X to $\hat{Y}$, and Y to $\hat{X}$ such that $d_\phi(X, Y)$ is equal to the average distortion per pixel between X and $\hat{X}$, i.e., $d(X, \hat{X})$, and between Y and $\hat{Y}$, i.e., $d(Y, \hat{Y})$. Nonetheless, the descriptive complexity of the conversion method is completely ignored in SMID $d_\phi(X, Y)$.

Based on descriptive complexity, particularly Kolmogorov complexity [14], [15], the notion of information distance was proposed in [16] for discrete data objects such as strings over a finite alphabet. Given any two finite strings x and y, their information distance $E_0(x, y)$ was defined in [16] to be the length of the shortest program which, when running on a universal computer (i.e., Turing machine), will convert x into y when x is the input, and convert y into x when y is the input. An inspiring property of $E_0(x, y)$ is its universality [16], which says in theory $E_0(x, y)$ captures all patterns and regularities that can be utilized computationally and are shared by x and y, and hence is the best cognitive distance one could hope for to a certain extent. In [4] and references therein, this notion was successfully applied to bioinformatics, music clustering, and machine translation.

However, the information distance as defined in [16] has two major issues. First, since it is based on Kolmogorov Complexity, it is uncomputable. Second, more importantly, it does not apply to continuous-valued data, such as images and videos. Therefore, it is desirable to develop a notion of distance which could combine the best of both worlds: the universality from the information distance as defined in [16], and the computability and applicability to both discrete and continuous-valued data as in SMID $d_\phi(X, Y)$.

1.2 Research Contributions

In this thesis, we present the concept of distance between any two data objects X and Y (abstract alphabets) from the view of Shannon information theory. We bring distortion into the information distance $E_0(x, y)$, and descriptive complexity into SMID $d_\phi(X, Y)$. For that reason, we formulate a new coding paradigm where X and Y are encoded into a sequence of coded bits specifying a codeword (or method) which would, in turn, convert Y into $\hat{X}$, and X into $\hat{Y}$ such that both the distortion between X and $\hat{X}$ and the distortion

between Y and $\hat{Y}$ are less than or equal to a prescribed threshold D. As we need universality to some extent, we consider a class $\mathcal{C}$ of coding schemes within the coding paradigm. Given $\mathcal{C}$, the information distance $R_{\mathcal{C}}(X, Y, D)$ between X and Y at the distortion level D is then defined as the smallest number of coded bits afforded by coding schemes from $\mathcal{C}$. We then characterize and analyze the information distance $R_{\mathcal{C}}(X, Y, D)$ for some classes $\mathcal{C}$. In addition to characterizing $R_{\mathcal{C}}(X, Y, D)$, we are also interested in its relationship among different sources X, Y, Z, etc. as a notion of distance.

1.3 Thesis Organization

The rest of the thesis will be organized as follows:

In chapter 2, some background knowledge will be given. First, the brief overview of typicality and method of types, which are useful in achievability proof for the information theory problems, will be presented. Then we will introduce the lossy compression when side information is available at the encoder, the decoder, or both. In the last part, we will explain the rate-distortion function when side information is available only at the decoder (Wyner-Ziv coding) with more details.

In chapter 3, we formally formulate the new coding paradigm and define the information distance $R_{\mathcal{C}}(X, Y, D)$. Then we will analyze the distance property of $R_{\mathcal{C}}(X, Y, D)$ when $\mathcal{C}$ consists of all coding schemes allowed in the coding paradigm, and establish upper and lower bounds to $R_{\mathcal{C}}(X, Y, D)$, which are further shown to be tight when X and Y are jointly Gaussian or DSBS. For the jointly Gaussian sources, we will introduce the new pseudo distance when the distortion level is small or the prescribed threshold D is less than or equal to a special term, which is a function of statistical properties of the sources such as variances. Furthermore, in the last section, the pseudo distance over the set of all real-values and IID sources will be proposed while the distortion measure is quadratic and the distortion level D is small.

In chapter 4, we will impose some constraints on $\mathcal{C}$ and formulate the coding paradigm. Then, $R_{\mathcal{C}}(X, Y, D)$ will be characterized in terms of the Wyner-Ziv Coding rate of X with Y as side information and the Wyner-Ziv Coding rate of Y with X as side information when $\mathcal{C}$ consists only of all so-called separately precoded broadcast codes within the coding paradigm. At the end, its distance property among different sources X, Y, Z, etc. will be presented.

Finally, in the last chapter, we will make a summary of the thesis and discuss several potential paths for future works related to these interesting topics.

Chapter 2

Background

2.1 Overview

In this chapter, we will go over some background materials and topics related to this research thesis. Before describing the detailed framework of new information distance in Chapters 3 and 4, we will talk about some of the required background knowledge on different typicality and method of types, which are important tools in proving coding theorems, in sections 2.2 and 2.3, respectively. In section 2.4, we review and describe the source coding with side information for the lossy case when side information or helper is available at the encoder or the decoder.

2.2 Overview of Typicality

Typicality is an important tool to prove coding theorems in information theory. In this section we review the definition of typicality and some basic properties ([1, 14, 17]) needed in the latter proofs. All the logarithms are in the base 2 unless otherwise specified. We did not go through the details of typicality proofs. More details can be found easily in [1, Ch.3], [14, Ch.2], and [17, Ch.5, 6].

2.2.1 Weak Typicality

We consider an information source $\{X_i\}_{i=1}^n$ where X_i are IID with distribution $p(x)$. We use X to denote the generic random variable and $H(X)$ to denote the common entropy for

all X_i, where $H(X) < \infty$. Since X_i are IID, then

$$p(X_1, X_2, \cdots, X_n) = p(X_1)p(X_2)\cdots p(X_n) \tag{2.1}$$

We now review an asymptotic property of $p(X_1, X_2, \cdots, X_n)$ called the weak Asymptotic Equipartition Property (AEP).

Theorem 1. *If $X_1, X_2, \cdots, X_n$ are IID $\sim p(x)$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) \longrightarrow H(X) \tag{2.2}$$

in probability as $n \to \infty$, i.e., for any $\epsilon > 0$, for n sufficiently large,

$$\Pr\{|-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) - H(X)| \leq \epsilon\} > 1 - \epsilon.[17] \tag{2.3}$$

Then we have the following definitions and theorems [1, 14, 17].

Definition 2.1. *The weakly typical set $\mathcal{A}_\epsilon^{(n)}(X)$ with respect to $p(x)$ is the set of sequences $x^n = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ such that*

$$|-\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) - H(X)| \leq \epsilon \tag{2.4}$$

or equivalently,

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{2.5}$$

where ϵ is an arbitrarily small positive real number. The sequences in $\mathcal{A}_\epsilon^{(n)}(X)$ are called weakly ϵ-typical sequences.

Theorem 2. *The following hold for any $\epsilon > 0$:*

1. ***Uniformity:*** *If $(x_1, x_2, \cdots, x_n) \in \mathcal{A}_\epsilon^{(n)}(X)$, then*

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) \leq H(X) + \epsilon.$$

2. ***Unit probability:*** *For n sufficiently large,*

$$\Pr\{X^n \in \mathcal{A}_\epsilon^{(n)}(X)\} \geq 1 - \epsilon.$$

3. ***Bounded size:*** *For n sufficiently large,*

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^{(n)}(X)| \leq 2^{n(H(X)+\epsilon)}.$$

Remark 1. *Weak typicality can be handy when we are working with abstract alphabets, stationary, and ergodic sources because it is related to the Shannon– McMillan–Breiman theorem. When source* $X = \{X_i\}_{i=1}^{\infty}$ *is a discrete stationary ergodic, we have*

$$-\frac{1}{n}\log p(X^n) \to \overline{H}(X)$$

where $\overline{H}(X)$ *is the entropy rate of the source. At the same time, provided that the source is continuous stationary ergodic, we will have*

$$-\frac{1}{n}\log p(X^n) \to \overline{h}(X)$$

where $\overline{h}(X)$ *is the differential entropy rate of the source [1].*

2.2.2 Strong Typicality

In this subsection, a stronger notion of typicality is defined such that the empirical probability of each possible outcome is near to the real corresponding probability. Strong typicality is more powerful than weak typicality as a tool for theorem proving in stronger results, such as rate-distortion theory and universal coding [14]. However, strong typicality can be used only for discrete-valued data[17].

Definition 2.2. *The strongly typical set* $\mathcal{A}_{\epsilon}^{*(n)}(X)$ *with respect to* $p(x)$ *is the set of sequences* $x^n = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ *such that*

1. *If* $x \in \mathcal{X}$ *with* $p(x) > 0$*, we have*

$$|N(x|x^n) - p(x)| \leq \frac{\epsilon}{|\mathcal{X}|} \tag{2.6}$$

2. *For all* $x \in \mathcal{X}$ *with* $p(x) = 0$*, then* $N(x|x^n) = 0$

where ϵ *is an arbitrarily small positive real number and* $N(x|x^n)$ *is the number of occurrences of* x *in the sequence* x^n*. The sequences in* $\mathcal{A}_{\epsilon}^{*(n)}(X)$ *are called strongly* ϵ*-typical sequences [17].*

In parallel with Theorem 2, We have The following Theorem [17], [1].

Theorem 3. *There exists $\eta > 0$ such that $\eta \to 0$ as $\epsilon \to 0$, and then the following hold:*

1. ***Uniformity:*** *If $(x_1, x_2, \cdots, x_n) \in \mathcal{A}_\epsilon^{*(n)}(X)$, then*

$$2^{-n(H(X)+\eta)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\eta)}.$$

2. ***Unit probability:*** *For n sufficiently large,*

$$\Pr\{X^n \in \mathcal{A}_\epsilon^{*(n)}(X)\} \geq 1 - \epsilon.$$

3. ***Bounded size:*** *For n sufficiently large,*

$$(1-\epsilon)2^{n(H(X)-\eta)} \leq |\mathcal{A}_\epsilon^{*(n)}(X)| \leq 2^{n(H(X)+\eta)}.$$

2.2.3 Robust Typicality

Robust typicality has certain advantages over its strong counterpart. The bounds have more natural expressions, many proofs are simplified, and a single definition applies to arbitrary collections of random variables. The crucial property distinguishing robust from strong typicality is: If x^n is robustly ϵ-typicality, then $p(x) = 0$ implies $N(x|x^n) = 0$ [18].

Definition 2.3. *The robustly typical set $\mathcal{T}_\epsilon^{(n)}(X)$ with respect to $p(x)$ is the set of sequences $x^n = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ such that*

$$|N(x|x^n) - p(x)| \leq \epsilon p(x), \qquad for\ all\ x\ \in \mathcal{X} \tag{2.7}$$

where ϵ is an arbitrarily small positive real number and $N(x|x^n)$ is the number of occurrences of x in the sequence x^n. The sequences in $\mathcal{T}_\epsilon^{(n)}(X)$ are called robustly ϵ-typical sequences [18].

The following simple fact is a direct consequence of the definition of the robustly ϵ-typical sequences[1].

Lemma 1. ***Typical Average Lemma.*** *Let $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$. Then for any non-negative function $f(x)$ on $\mathcal{X}$,*

$$(1-\epsilon)\mathbf{E}[f(X)] \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i) \leq (1+\epsilon)\mathbf{E}[f(X)]. \tag{2.8}$$

In the rest of this subsection, we give robust typicality equivalents of standard strong typicality results. The proof of all lemmas and theorems can be found in [1, Ch.2] and [18].

Theorem 4. *There exists $\delta(\epsilon) = \epsilon H(X) > 0$ such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and then the following hold [1]:*

1. ***Uniformity:*** *If $p(x^n) = \prod_{i=1}^n p_X(x_i)$ and $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$, then*

$$2^{-n(H(X)+\delta(\epsilon))} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\delta(\epsilon))}.$$

2. ***Unit probability:*** *If $X_1, X_2, \cdots, X_n$ are IID with $X_i \sim p_X(x_i)$, then by the Law of Large Number (LLN),*

$$\lim_{n\to\infty} \Pr\{(X^n) \in \mathcal{T}_\epsilon^{(n)}(X)\} = 1.$$

3. ***Bounded size:*** *For n sufficiently large,*

$$(1-\epsilon)2^{n(H(X)-\delta(\epsilon))} \leq |\mathcal{T}_\epsilon^{(n)}(X)| \leq 2^{n(H(X)+\delta(\epsilon))}.$$

Theorem 4 is illustrated in Figure 2.1.

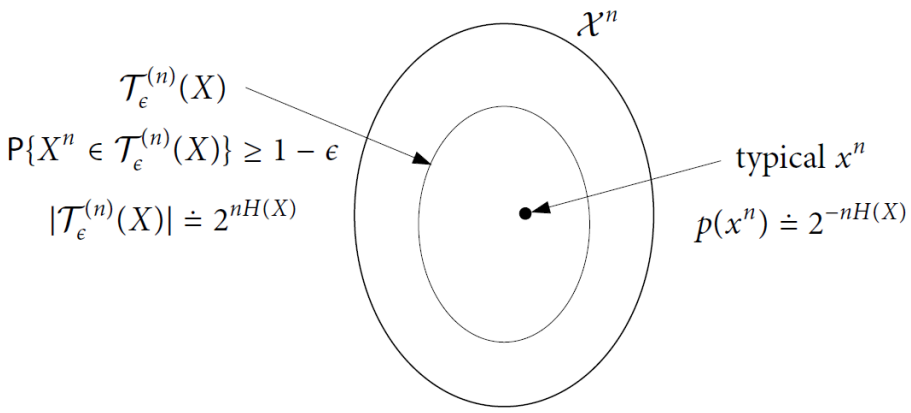

Figure 2.1: Properties of typical sequences. Here $X^n \sim \prod_{i=1}^n p_X(x_i)$[1].

The notion of the robust typicality can be easily extended to multiple random variables.

Definition 2.4. *The robustly jointly typical set $\mathcal{T}_\epsilon^{(n)}(X,Y)$ with respect to $p(x,y)$ is the set of sequences $(x^n,y^n)\in\mathcal{X}^n\times\mathcal{Y}^n$ such that*

$$|N(x,y|x^n,x^n)-p(x,y)|\leq\epsilon p(x,y) \tag{2.9}$$

where ϵ is an arbitrarily small positive real number and $N(x,y|x^n,x^n)$ is the number of occurrences of (x,y) in the pair of sequences (x^n,y^n). The sequences in $\mathcal{T}_\epsilon^{(n)}(X,Y)$ are called robustly ϵ-jointly typical sequences. [18]

Remark 2. *If $(x^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)$, then $x^n\in\mathcal{T}_\epsilon^{(n)}(X)$ and $y^n\in\mathcal{T}_\epsilon^{(n)}(Y)$.*

Lemma 2. ***Conditionally Typicality Lemma.*** *Let $(X,Y)\sim p(x,y)$. Suppose that $x^n\in\mathcal{T}_{\epsilon'}^{(n)}(X)$ and $Y^n\sim p(y^n|x^n)=\prod_{i=1}^n p_{(Y|X)}(y_i|x_i)$. Then for every $\epsilon>\epsilon'$, with the use of LLN, we have [1]*

$$\lim_{n\to\infty}\Pr\{(x^n,Y^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)\}=1.$$

The following lemma is one of the most important lemma which is very useful in the achievability proofs of coding problems. (see Appendixes A.1, A.2)

Lemma 3. ***Joint Typicality Lemma.*** *Let $(X,Y)\sim p(x,y)$ and $\epsilon>\epsilon'$. Then there exists $\delta(\epsilon)>0$ which tends to zero as $\epsilon\to 0$ such that the following statements hold [1]:*

1. *If $\tilde{x}^n$ is an arbitrary sequence and $\tilde{Y}^n\sim\prod_{i=n}^n p_Y(\tilde{y}_i)$, then*

$$\Pr\{(\tilde{x}^n,\tilde{Y}^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)\}\leq 2^{-n(I(X;Y)-\delta(\epsilon))}.$$

2. *If $x^n\in\mathcal{T}_{\epsilon'}^{(n)}(X)$ and $\tilde{Y}^n\sim\prod_{i=n}^n p_Y(\tilde{y}_i)$, then for n sufficiently large,*

$$\Pr\{(x^n,\tilde{Y}^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)\}\geq 2^{-n(I(X;Y)+\delta(\epsilon))}.$$

2.3 Overview of Types

The AEP in Theorem 1 focuses our attention on a small subset of typical sequences. The method of types is an even more powerful procedure in which we consider sequences that have the same empirical distribution [14]. The method of types evolved from notions of strong typicality and this method was fully developed by Csiszar and Korner [19], who derived the main theorems of information theory from this viewpoint.

Let $x^n=(x_1,x_2,\cdots,x_n)$ be a sequence from alphabet $\mathcal{X}$. Suppose that like the previous section, $N(x|x^n)$ denotes the number of occurrences of x in the sequence x^n. Then we have the following definitions and theorems [14, 19].

Definition 2.5. ***Type.*** *The type P_{x^n} (empirical probability distribution) of a sequence x^n is the relative proportion of occurrences of $\mathcal{X}$, i.e.,*

$$P_{x^n}(x) = \frac{N(x|x^n)}{n}, \qquad \forall\, x \in \mathcal{X}. \tag{2.10}$$

where $N(x|x^n)$ is the number of occurrences of x in the sequence x^n.

Example 1. *Let $\mathcal{X} = \{0, 1, 2\}$, $n = 6$, and $x^6 = (1, 1, 2, 2, 2, 0)$. Then $N(0|x^6) = 1$, $N(1|x^6) = 2$, and $N(2|x^6) = 3$. Therefore, $P_{x^n} = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$.*

Definition 2.6. ***All possible types.*** *Let $\mathcal{P}_n(\mathcal{X})$ be the collection of all possible types of sequences of length n on $\mathcal{X}$*

For example, if $\mathcal{X} = \{0, 1\}$, the set of possible types with the sequence of length n is

$$\mathcal{P}_n(\mathcal{X}) = \{(P(0), P(1)) : (\frac{0}{n}, \frac{n}{n}), (\frac{1}{n}, \frac{n-1}{n}), \cdots, (\frac{n}{n}, \frac{0}{n})\}.$$

Lemma 4. *An upper bound for $|\mathcal{P}_n(\mathcal{X})|$:*

$$|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}. \tag{2.11}$$

Definition 2.7. *A $P \in \mathcal{P}_n(\mathcal{X})$ is said to be an n-type if for any $x \in \mathcal{X}$, $P(x) \in \{0, \frac{1}{n}, \frac{2}{n}, \cdots 1\}$.*

Definition 2.8. ***Type class.*** *Let $P \in \mathcal{P}_n(\mathcal{X})$, the set of sequences of length n with type P is called type class of P, denoted $T^n_{\mathcal{X}}(P)$:*

$$T^n_{\mathcal{X}}(P) = \{x^n \in \mathcal{X}^n : P_{x^n} = P\}. \tag{2.12}$$

Theorem 5. ***Size of a type class.*** *For any type $P \in \mathcal{P}_n(\mathcal{X})$,*

$$\frac{2^{nH(P)}}{(n+1)^{|\mathcal{X}|}} \leq |T^n_{\mathcal{X}}(P)| \leq 2^{nH(P)}. \tag{2.13}$$

2.4 Lossy Compression with Side information

When the information source is continuous valued data, it is not possible to encode the source information symbols using finitely many bits. Therefore some information has to be lost when digital communications are used. We say that the source coding is lossy in this situation. It is important to mention we can also apply lossy source coding for discrete information sources, if some information loss is tolerable. Theoretically, lossy source coding can be studied using rate distortion theory [14].

First, a distortion measure is defined as follows:

Definition 2.9. *A distortion measure is a mapping*

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathcal{R}^+ \tag{2.14}$$

from the set of source alphabet- reproduction alphabet pairs into the set of non-negative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by the symbol $\hat{x}$[14].

The distortion measure is defined on a symbol-to-symbol basis. We extend the definition to sequences by using the additive distortion measure:

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i). \tag{2.15}$$

Formally, a $(2^{nR}, n)$ lossy source code consists of

1. An encoder which assigns an index $m(x^n) \in [1 : 2^{nR})$ to each sequence $x^n \in \mathcal{X}^n$, and

2. A decoder that allocate an estimate $\hat{X}^n(m) \in \hat{\mathcal{X}}^n$ to each index $m \in [1 : 2^{nR})$.

The set $\mathcal{C} = \{\hat{x}^n(1), \cdots, \hat{x}^n(\lfloor 2^{nR} \rfloor)\}$ constitutes the codebook.

The expected distortion associated with a $(2^{nR}, n)$ lossy source code is defined as

$$\mathbf{E}[d(X^n, \hat{X}^n)] = \sum_{x^n} p(x^n) d(x^n, \hat{x}^n(m(x^n))).$$

A rate distortion pair (R, D) is said to be achievable if there exists a sequence of $(2^{nR}, n)$-rate distortion codes with

$$\limsup_{n \to \infty} \mathbf{E}[d(X^n, \hat{X}^n)] \leq D. \tag{2.16}$$

The rate distortion function $R(D)$ is the infimum of rates R such that (R, D) is achievable [1, 14].

2.4.1 Three Simple Lossy Source Coding Cases

In this subsection, we only mention three simple lossy source coding based on the availability of side information at the encoder or the decoder. Let $(X, Y) = \{(X_I)\}$ be a 2-DMS and $d(x, \hat{x})$ be a distortion measure function. Then we have these simple cases:

A. (Rate distortion) No side information at either the encoder or the decoder

By the lossy source coding theorem in [1, Ch.3], the rate distortion function with no side information at either the encoder or the decoder is

$$R(D) = \min_{p(\hat{x}|x):\mathbf{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}). \tag{2.17}$$

This case is illustrated in Figure 2.2.

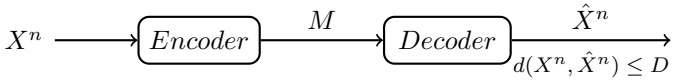

Figure 2.2: Lossy compression system with no side information.

B. Side information available only at the encoder

One can easily show that when side information is available only at the encoder, the rate distortion function is exactly similar to the Case A [1], i.e.,

$$R_{SI-E}(D) = \min_{p(\hat{x}|x):\mathbf{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}) = R(D). \tag{2.18}$$

This case is shown in Figure 2.3.

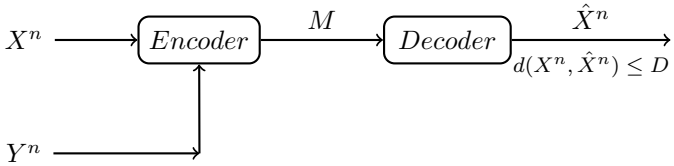

Figure 2.3: Lossy compression system with side information available only at the encoder.

C. (Conditional Rate Distortion) Side information available at both the encoder and the decoder

In this case, side information Y^n is available at the encoder and the decoder. By simple extension of the proof of the lossy source coding theorem [1, Ch.3], the rate distortion function will be a conditional version of case with no side information (Case A), i.e.,

$$R_{SI-ED}(D) = R_{X|Y}(D) = \min_{p(\hat{x}|x,y):\mathbf{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}|Y). \tag{2.19}$$

This case also is depicted in Figure 2.4.

But what if the encoder does not have side information? In that case, the result is different, and is the subject of the next subsection.

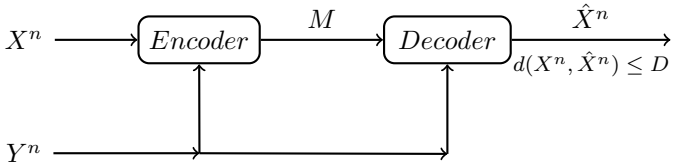

Figure 2.4: Lossy compression system with side information available at both the encoder and the decoder.

2.4.2 Wyner-Ziv Coding

We did not yet consider another special case of side information availability. Suppose that the side information sequence is available only at the decoder. This case was studied by Wyner and Ziv in their fundamental paper from 1976 [20].

As we shown in Figure 2.5, a $(2^{nR}, n)$ lossy source code with side information available at the decoder consists of

1. An encoder which maps an index $m(x^n) \in [1 : 2^{nR})$ to each sequence $x^n \in \mathcal{X}^n$, and

2. A decoder that assigns an estimate $\hat{X}^n(m, y^n)$ to each received index m and side information sequence y^n.

The rate distortion function with side information available at the decoder ($R^{WZ}_{X|Y}$) is the infimum of rates R such that there exists a sequence of $(2^{nR}, n)$-rate distortion codes with

$$\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)] \leq D.$$

Wyner-Ziv Coding is illustrated in Figure 2.5.

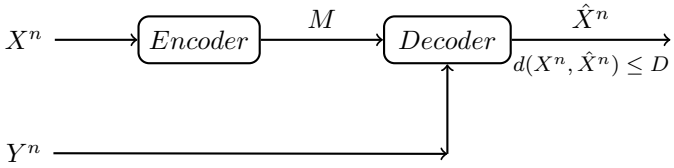

Figure 2.5: Wyner-Ziv coding system.

Since the compressed-binning idea of the proof of Wyner-Ziv coding is helpful for us in Chapter 4, we will show the proof of Wyner-Ziv coding with details. The following theorem gives the rate distortion function for this specific case.

Theorem 6. ***(Wyner-Ziv Theorem)*** *[20, 1]*

Let (X, Y) be a 2-DMS and $d(x, \hat{x})$ be a distortion measure. If the side information Y is available only to the decoder, the rate-distortion function for X should be

$$R_{X|Y}^{WZ}(D) = \min_{p(u|x), \hat{x}(u,y)} (I(X;U) - I(Y;U)) = \min_{p(u|x), \hat{x}(u,y)} I(X;U|Y) \tag{2.20}$$

where the minimum is over all conditional Probability Mass Function (PMF) $p(u|x)$ with random variable U such that $|\mathcal{U}| \leq |\mathcal{X}| + 1$, Y — X — U and X —(U, Y)— $\hat{X}$ form Markov chains, and over all reconstruction functions $\hat{x}(u, y)$ such that $E(d(X, \hat{X})) \leq D$.

Proof. To prove Wyner-Ziv theorem, we need to verify both Achievability and Converse proofs:

Achievability part: We want to show that if $R > R_{X|Y}^{WZ}(D)$, then there exists a sequence of $(2^{nR}, n)$ codes such that

$$\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)] \leq D.$$

The Wyner-Ziv coding scheme uses the **compress-bin** idea illustrated in Figure 2.6. To describe X by U, We use joint typicality encoding. Since U has a correlation with Y, binning method can reduce their description rate. The bin index of U is transferred to the decoder. The receiver uses joint typicality decoding with Y to recover U and then reconstructs $\hat{X}$ from U and Y. We now provide the details.

Codebook generation: Fix the conditional PMFs $p(u|x)$ and compute $p(u)$ according to $p(u) = \sum_{x\in\mathcal{X}} p(x)p(u|x)$; also fix the reconstruction function $\hat{x}(u, y)$ such that $\mathbf{E}[d(X, \hat{X})] \leq \frac{D}{1+\epsilon}$, where D is the desired distortion. Randomly and independently produce $2^{n\tilde{R}}$ sequences $u^n(l)$, $l \in [1 : 2^{n\tilde{R}}]$, each according to $\prod_{i=1}^n p_U(u_i)$. Now, partition the set of indices l into equal-size subsets referred to as bins $\mathcal{B}(m)$, as follows:

$$\mathcal{B}(m) = [(m-1)2^{n(\tilde{R}-R)} + 1 : m2^{n(\tilde{R}-R)}], \qquad m \in [1 : 2^{nR}]. \tag{2.21}$$

The codebook generation is revealed to the encoder and the decoder.

Encoding: Given a source sequence x^n, the encoder looks for a codeword $u^n(l)$ such that $(x^n, u^n(l)) \in \mathcal{T}_{\epsilon'}^{(n)}(X, U)$. If there is more than one such index, the encoder uses one of them uniformly at random. If there is no such index, it sets randomly an index from $[1 : 2^{n\tilde{R}}]$ uniformly. The encoder will send the bin index m such that $l \in \mathcal{B}(m)$.

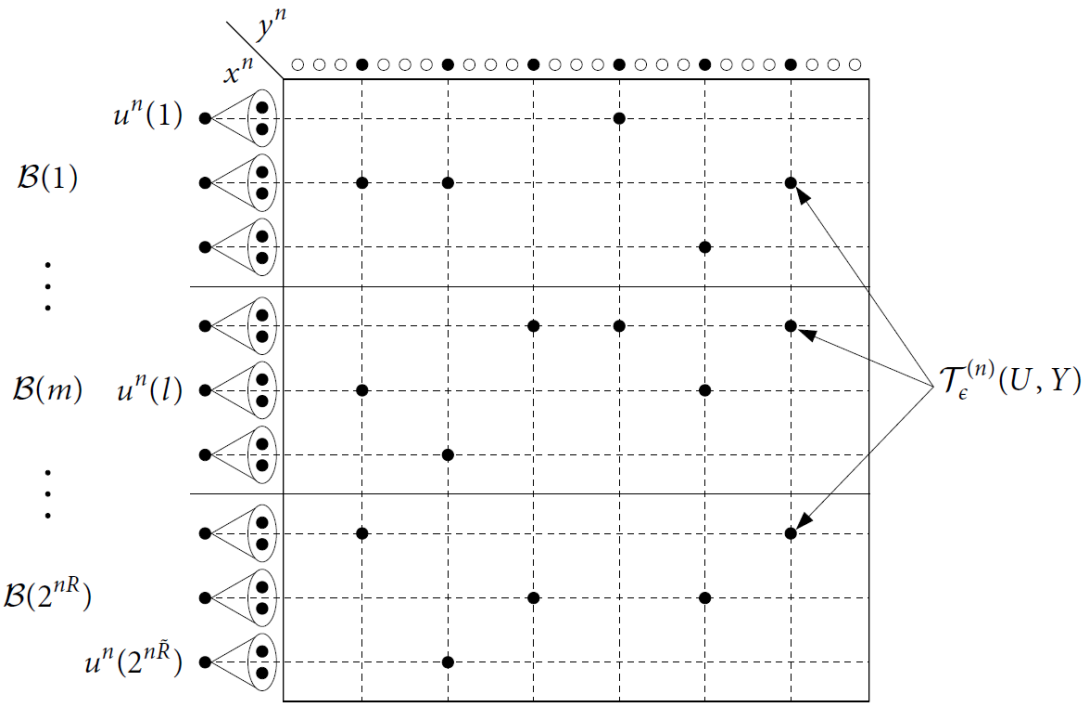

Figure 2.6: Wyner-Ziv coding scheme. Each bin $\mathcal{B}(m)$, $m \in [1 : 2^{nR}]$, consists of $2^{n(\tilde{R}-R)}$ indices [1].

Decoding: Let $\epsilon > \epsilon'$. Upon receiving m, the decoder finds the index $\hat{l} \in \mathcal{B}(m)$ such that $(y^n, u^n(\hat{l})) \in \mathcal{T}_\epsilon^{(n)}(Y, U)$. If there is a unique such codeword, the decoder outputs the reconstruction sequence as $\hat{x}_i = \hat{x}(u_i(\hat{l}), y_i)$ for $i \in [1 : n]$; otherwise it sets $\hat{l} = 1$ and then compute the reconstruction sequence.

Analysis of expected distortion: Let (L, M) denote the chosen indices at the encoder and $\hat{L}$ be the index estimate at the receiver. We define the " Error " event as

$$\mathcal{E} = \{(U^n(\hat{L}), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(U, X, Y)\} \tag{2.22}$$

Now, consider the events

$$\begin{aligned}
\mathcal{E}_1 &= \{(U^n(l), X^n) \notin \mathcal{T}_{\epsilon'}^{(n)}(U, X) \text{ for all } l \in [1 : 2^{nR_1})\}, \\
\mathcal{E}_2 &= \{(U^n(L), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(U, X, Y)\}, \\
\mathcal{E}_3 &= \{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U, Y) \text{ for some } \tilde{l} \in \mathcal{B}(M), \tilde{l} \neq L\},
\end{aligned}$$

Since the "Error" event happens when $(U^n(L), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(U, X, Y)$ or $\hat{L} \neq L$, by the union of events bound, we have

$$P(\mathcal{E}) \leq P(\mathcal{E}_1) + P(\mathcal{E}_1^{\mathrm{c}} \cap \mathcal{E}_2) + P(\mathcal{E}_3)$$

We now bound each term. By the covering lemma (see Appendix A.1), $P(\mathcal{E}_1)$ tends to zero as $n \to \infty$ if

$$\tilde{R} > I(X; U) + \delta(\epsilon'). \tag{2.23}$$

Since $\epsilon > \epsilon'$, $\mathcal{E}_1^{\mathrm{c}} = \{(U^n(L), X^n) \in \mathcal{T}_{\epsilon'}^{(n)}(U, X)\}$, and $Y^n \mid \{U^n(L) = u^n, X^n = x^n\} \sim \prod_{i=1}^n P_{Y|U,X}(y_i|u_i, x_i) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)$, by the conditional typicality lemma (Lemma 2), $P(\mathcal{E}_1^{\mathrm{c}} \cap \mathcal{E}_2) \to 0$ as $n \to \infty$.
To bound $P(\mathcal{E}_3)$, we first use lemma 11.1 of [1] (see Appendix A.3) to find an upper bound as follows:

$$P(\mathcal{E}_3) \leq \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U, Y) \text{ for some } \tilde{l} \in \mathcal{B}(1)\}$$

For each $\tilde{l} \in \mathcal{B}(1)$, the sequence $U^n(\tilde{l}) \sim \prod_{i=1}^n P_U(u_i)$ is independent of Y^n. Now, by the packing lemma (see Appendix A.2), $\Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U, Y) \text{ for some } \tilde{l} \in \mathcal{B}(1)\}$ tends to zero as $n \to \infty$ if $\tilde{R} - R < I(Y; U) - \delta(\epsilon)$. Therefore, by combining the bounds, we have shown that $P(\mathcal{E})$ tends to zero as $n \to \infty$ if

$$R > \tilde{R} - I(Y; U) + \delta(\epsilon) > I(X; U) - I(Y; U) + \delta(\epsilon) + \delta(\epsilon') = I(X; U|Y) + \delta'(\epsilon) \tag{2.24}$$

Since in the codebook generation part, we fixed the the conditional PMFs $p(u|x)$ as well as functions $\hat{x}(u, y)$ such that $\mathbf{E}[d(X, \hat{X})] \leq \frac{D}{(1+\epsilon)}$, we can achieve tighter lower bound by minimizing over $p(u|x)$ and $\hat{x}(u, y)$. Therefore, we have

$$R > \min_{p(u|x), \hat{x}(u,y):\, \mathbf{E}[d(X,\hat{X})] \leq \frac{D}{1+\epsilon}} I(X; U|Y) + \delta'(\epsilon) = R_{X|Y}^{WZ}\left(\frac{D}{1+\epsilon}\right) + \delta'(\epsilon) \tag{2.25}$$

Now, if $R > R_{X|Y}^{WZ}(\frac{D}{1+\epsilon}) + \delta'(\epsilon)$, then there will be no "Error" i.e., $P(\mathcal{E}) \to 0$.

When there is no "Error", $(U^n(L), X^n, Y^n) \in \mathcal{T}_\epsilon^n(U, X, Y)$. Then by the law of total expectation and the typical average lemma (Lemma 1), the asymptotic averaged distortion is upper bounded as

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)] = \limsup_{n\to\infty} (P(\mathcal{E})\mathbf{E}[d(X^n, \hat{X}^n)|\mathcal{E}] + P(\mathcal{E}^{\mathrm{c}})\mathbf{E}[d(X^n, \hat{X}^n)|\mathcal{E}^{\mathrm{c}}]) \\
&\leq \limsup_{n\to\infty} (d_{max} P(\mathcal{E}) + P(\mathcal{E}^{\mathrm{c}})(1+\epsilon)\mathbf{E}[d(X, \hat{X})] \leq (1+\epsilon)\frac{D}{(1+\epsilon)} = D
\end{aligned}$$

where $d_{max} = \max_{(x,\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} d(x,\hat{x})$.

In the last step, from the continuity of $R_{X|Y}^{WZ}(D)$ and $R_{Y|X}^{WZ}(D)$ in D [20], [14], taking $\epsilon \to 0$ shows that any rate distortion pair (R, D) with $R > R_{X|Y}^{WZ}(D)$ is achievable, which completes the proof of achievability.

Converse part: Suppose that R is achievable. Then there exists a sequence of codes $(2^{nR}, n)$ that satisfies 2.16. We want to show that if R is achievable, then $R \geq R_{Y|X}^{WZ}(D)$. M denotes the compressed value of the vector X^n and $\hat{X}^n$ represents the reconstructed vector. For a given block of length n, we have

$$\begin{aligned}
nR &\geq H(M) \\
&\geq H(M|Y^n) \\
&= I(X^n; M|Y^n) \\
&= H(X^n|Y^n) - H(X^n|M, Y^n) \\
&= \sum_{i=1}^{n} H(X_i|Y^n, X^{i-1}) - H(X_i|M, Y^n, X^{i-1}) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|M, Y^{i-1}, Y_i, Y_{i+1}^n, X^{i-1}) \\
&\geq \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|M, Y^{i-1}, Y_i, Y_{i+1}^n)
\end{aligned}$$

where (a) follows since (X_i, Y_i) is independent of $(Y^{i-1}, Y_{i+1}^n, X^{i-1})$. Let $U_i = (M, Y^{i-1}, Y_{i+1}^n)$. This random variable satisfies two interesting properties:

- The triple (U_i, X_i, Y_i) is a Markov chain i.e., U_i — X_i — Y_i, because U_i has information about Y_i only through M which is a function of X_i,
- $\hat{X}_i$ is a function f of $(U_i, Y_i) = (Y^n, M)$,

Then we have

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|M, Y^{i-1}, Y_i, Y_{i+1}^n) \\
&= \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|U_i, Y_i) \\
&= \sum_{i=1}^{n} I(X_i; U_i|Y_i) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} R_{X|Y}^{WZ}(\mathbf{E}[d(X_i, \hat{X}_i)]) \\
&\overset{(c)}{\geq} nR_{X|Y}^{WZ}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[d(X_i, \hat{X}_i)]) \\
&= nR_{X|Y}^{WZ}(\mathbf{E}[d(X^n, \hat{X}^n)])
\end{aligned}
$$

where (b) comes from the definition of $R_{X|Y}^{WZ}(D) = \min I(X;U|Y)$, and (c) follows by the convexity of $R_{X|Y}^{WZ}(D)$ [20], [14]. Since $R_{X|Y}^{WZ}(D)$ is continuous and non-increasing in D, it follows from the bound on distortion in 2.16 that

$$
\begin{aligned}
R &\geq \limsup_{n\to\infty} R_{X|Y}^{WZ}(\mathbf{E}[d(X^n, \hat{X}^n)]) \\
&\geq R_{X|Y}^{WZ}(\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)]) \\
&\geq R_{X|Y}^{WZ}(D) \qquad (2.26)
\end{aligned}
$$

which completes the proof of converse. Also the cardinality bound on $\mathcal{U}$ is proved in [1] by the convex cover method. (Since it is not relevant to our work, We will not go through the details.) This completes the proof of the Wyner-Ziv Theorem. □

Remark 3. *We can also use the random binning instead of deterministic binning. In Theorem 6, we generate a set of random sequences; therefore, we do not need to use a random binning. However, in [14], they used the random binning instead of deterministic binning.*

Remark 4. *In Theorem 6, we used the robust typicality. For the achievability part of Wyner-Ziv Theorem, we can also use strong or weak typicality. For example, [14] uses strong typicality instead of robust typicality inside of the achievability part. It is important to mention that analysis of expected distortion will be a little different while we are using*

a strong typicality. In this case, the distortion is related to the type of the random vectors in the sense that

$$d(X^n, \hat{X}^n) = \frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) = \sum_{(x\times\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} P_{x^n,\hat{x}^n}(x,\hat{x})d(x,\hat{x}) \tag{2.27}$$

where $P_{x^n,\hat{x}^n}(x,\hat{x})$ is the joint type of $(X^n, \hat{X}^n)$. The expression above can be understood as the "mean" or "empirical expectation" of the distortion.

Since there is no "Error" and $(U^n(L), X^n, Y^n) \in \mathcal{A}_\epsilon^{*(n)}(U,X,Y)$, the joint type of $(U^n(L), X^n, Y^n)$ is close to its real PMF, then the above "mean" is close to the real expectation value. More precisely

$$\begin{aligned}
d(X^n, \hat{X}^n) &= \sum_{(x\times\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} P_{x^n,\hat{x}^n}(x,\hat{x})d(x,\hat{x}) \\
&= \sum_{(x\times\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} (P_{x^n,\hat{x}^n}(x,\hat{x}) - p(x,\hat{x}) + p(x,\hat{x}))d(x,\hat{x}) \\
&= \sum_{(x\times\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} (P_{x^n,\hat{x}^n}(x,\hat{x}) - p(x,\hat{x}))d(x,\hat{x}) + \sum_{(x\times\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} p(x,\hat{x})d(x,\hat{x}) \\
&\leq |\mathcal{X}\times\hat{\mathcal{X}}| \max_{x\times\hat{x}}(d(x,\hat{x})|P_{x^n,\hat{x}^n}(x,\hat{x}) - p(x,\hat{x})|) + \mathbf{E}[d(X,\hat{X})]
\end{aligned} \tag{2.28}$$

Finally, by taking the expectation over the random codebook and then taking $n \to \infty$ we have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)] \\
&\leq \limsup_{n\to\infty}(|\mathcal{X}\times\hat{\mathcal{X}}| \max_{x\times\hat{x}}(d(x,\hat{x})|P_{x^n,\hat{x}^n}(x,\hat{x}) - p(x,\hat{x})|) + \mathbf{E}[\mathbf{E}[d(X,\hat{X})]])
\end{aligned} \tag{2.29}$$

The first term tends to zero as $n \to \infty$, and the second one is not more than D. Therefore,

$$\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)] \leq D.$$

We finish this chapter by comparing the Wyner-Ziv coding with the three cases which have been explained in subsection 2.4.1.

Remark 5. *Let (X,Y) be 2-DMS. Then the following holds:*

$$R_{SI-ED}(D) = R_{X|Y}(D) \leq R_{SI-D}(D) = R_{X|Y}^{WZ}(D) \leq R_{SI-E}(D) = R(D). \tag{2.30}$$

2.5 Summary

In this chapter, we gave an overview of the different typicality and method of types as useful tools for proving the coding theorems. Then we introduced the four different lossy source coding based on the availability of side information at the encoder or the decoder. We went to the details of the most interesting case, when the side information is available only at the decoder, and reviewed the compress-binning idea in achievability proof of Wyner-Ziv Theorem.

Chapter 3

Coding for Data Analytics: New Information Distances

3.1 Overview

In this chapter, we will address the notion of distance between any two data objects X and Y from the perspective of Shannon information theory. A general coding paradigm will be introduced where X and Y are encoded into a sequence of coded bits which would, in turn, convert Y into $\hat{X}$, and X into $\hat{Y}$ such that both the distortion between X and $\hat{X}$ and the distortion between Y and $\hat{Y}$ are less than or equal to a prescribed threshold D. To have a universality to some extent, we consider a class $\mathcal{C}$ of coding schemes within the coding paradigm. Given $\mathcal{C}$, the information distance $R_{\mathcal{C}}(X, Y, D)$ between X and Y at the distortion level D is then defined as the smallest number of coded bits afforded by coding schemes from $\mathcal{C}$. We then characterize and analyze the information distance $R_{\mathcal{C}}(X, Y, D)$ for some classes $\mathcal{C}$.

In Section 3.2, we formally formulate the new coding paradigm and define the information distance $R_{\mathcal{C}}(X, Y, D)$. Section 3.3 defines how to analyze the distance property of $R_{\mathcal{C}}(X, Y, D)$ when $\mathcal{C}$ consists of all coding schemes allowed in the coding paradigm, and establish upper and lower bounds to $R_{\mathcal{C}}(X, Y, D)$, which are further shown to be tight when X and Y are jointly Gaussian. In Section 3.4, we will introduce the new pseudo distance over the set of all memoryless jointly Gaussian sources when the distortion level is small or the prescribed threshold D is less than or equal to a special term, which is a function of the statistical properties of jointly Gaussian sources, such as variance. Finally,

Section 3.5 defines the pseudo distance for any real-valued and IID source such that the distortion level D has a small value.

3.2 Formal Definitions: Codes and New Information Distances

Let $\mathbf{A}$ and $\hat{\mathbf{A}}$ be two abstract alphabets. They could be either continuous or discrete. The sets $\mathbf{A}$ and $\hat{\mathbf{A}}$ will serve as our source alphabet and reproduction alphabet, respectively. Let $\mathcal{A}$ be a σ-field of subsets of $\mathbf{A}$, and let $\hat{\mathcal{A}}$ be a σ-field of subsets of $\hat{\mathbf{A}}$. (Here we implicitly assume that any element of $\hat{\mathbf{A}}$ belongs to the σ-field $\hat{\mathcal{A}}$.) Let the measurable space

$$(\mathbf{A}^\infty, \mathcal{A}^\infty) = \prod_{k=1}^{\infty} (\mathbf{A}_k, \mathcal{A}_k)$$

be the infinite Cartesian product of exemplars $(\mathbf{A}_k, \mathcal{A}_k)$ of the measurable space $(\mathbf{A}, \mathcal{A})$. The measurable space $(\hat{\mathbf{A}}^\infty, \hat{\mathcal{A}}^\infty)$ is defined similarly. If $x = (x_i)$ is a finite or infinite sequence of symbols from $\mathbf{A}$ or $\hat{\mathbf{A}}$, let $x_m^n = (x_m, x_{m+1}, \cdots, x_n)$ and, for simplicity, write x_1^n as x^n. The same conventions apply to sequences of random variables taking their values in these sets as well. We denote the set of all n-tuples drawn from $\mathbf{A}$ ($\hat{\mathbf{A}}$) by $\mathbf{A}^n$ ($\hat{\mathbf{A}}^n$).

Without loss of generality, we assume that each of data objects X, Y, Z, etc. is a sequence of symbols from $\mathbf{A}$, and its lossy version is a sequence of symbols of the same length from $\hat{\mathbf{A}}$. (Discussions and results below can be easily extended to data objects from different alphabets.) In most cases, we model each data object as a stationary source taking values in $\mathbf{A}$. For example, $X = \{X_i\}_{i=1}^\infty$ will be a stationary source with each X_i being a random variable taking values in $\mathbf{A}$, and its lossy version $\hat{X} = \{\hat{X}_i\}_{i=1}^\infty$ will be a sequence of random variables taking values in $\hat{\mathbf{A}}$. Like the Chapter 2, let $d : \mathbf{A} \times \hat{\mathbf{A}} \to [0, \infty)$ be a measurable function. Let $\{d_n\}_{n=1}^\infty$ be the single-letter fidelity criterion generated by d, by which we mean that for each n, $d_n : \mathbf{A}^n \times \hat{\mathbf{A}}^n \to [0, \infty)$ is the map in which $d_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, y_i)$ for any $x^n \in \mathbf{A}^n$ and $y^n \in \hat{\mathbf{A}}^n$. The distortion between X^n and $\hat{X}^n$ is measured by $d(X^n, \hat{X}^n)$.

Graphically, our coding paradigm for data analytics is illustrated in Figure 3.1, where X^n and Y^n are encoded jointly into a sequence of coded bits at rate R in bits per symbol. The coded bits specify a codeword (or method) which would, in turn, convert X^n into $\hat{Y}^n$ at Decoder 1, and Y^n into $\hat{X}^n$ at Decoder 2 such that for all sufficiently large n, both $d(X^n, \hat{X}^n)$ and $d(Y^n, \hat{Y}^n)$ are less than or equal to a prescribed threshold D.

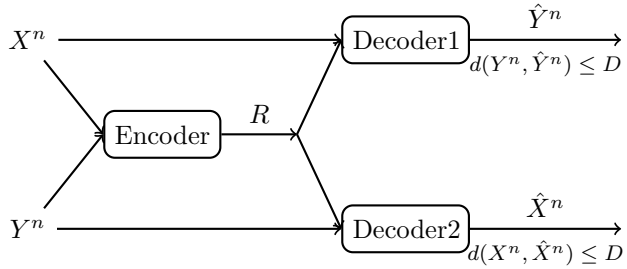

Figure 3.1: Coding for data analytics.

For any $R > 0$ and n, let

$$\Omega(n, R) = \{1, 2, \cdots, \lfloor 2^{nR} \rfloor\}.$$

Formally, we have the following definition.

Definition 3.1. *A block code C_n of order n and rate R consists of one encoding mapping*

$$f : \mathbf{A}^n \times \mathbf{A}^n \to \Omega(n, R)$$

and two decoding mappings

$$g_1 : \Omega(n, R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n$$

and

$$g_2 : \Omega(n, R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n.$$

For any two data objects x^n and y^n, the encoder f encodes x^n and y^n into $f(x^n, y^n)$ of nR bits. On the decoding side, the encoded message $f(x^n, y^n)$ then converts x^n into $\hat{y}^n = g_1(f(x^n, y^n), x^n)$, and y^n into $\hat{x}^n = g_2(f(x^n, y^n), y^n)$.

Impose no other constraints on C_n, and let $\mathcal{C}$ consist of all possible block codes of order n for all n. We want to seek the best trade-off between R and the maximum distortion $\max\{d(X^n, \hat{X}^n), d(Y^n, \hat{Y}^n)\}$ attainable by $\mathcal{C}$ for any stationary sources $X = \{X_i\}_{i=1}^{\infty}$ and $Y = \{Y_i\}_{i=1}^{\infty}$ and sufficiently large n.

Definition 3.2. *Given stationary sources $X = \{X_i\}_{i=1}^{\infty}$ and $Y = \{Y_i\}_{i=1}^{\infty}$, a rate distortion pair (R, D) is said to be achievable for (X, Y) if for any $\epsilon > 0$, there exists, for all sufficiently large n, a block code $C_n = (f, g_1, g_2)$ of order n and rate $R + \epsilon$ such that*

$$\Pr\{d(X^n, \hat{X}^n) > D\} < \epsilon \tag{3.1}$$

and

$$\Pr\{d(Y^n, \hat{Y}^n) > D\} < \epsilon \tag{3.2}$$

where $\hat{X}^n = g_2(f(X^n, Y^n), Y^n)$, *and* $\hat{Y}^n = g_1(f(X^n, Y^n), X^n)$.

Let $\mathcal{R}(X, Y)$ denote the set of all achievable (R, D) pairs for (X, Y). It can be verified that $\mathcal{R}(X, Y)$ is closed. Given $D \geq 0$, we define the information distance between X and Y at the distortion level D as

$$R(X, Y, D) \triangleq \min\{R : (R, D) \in \mathcal{R}(X, Y)\}. \tag{3.3}$$

One of our purposes in this chapter is to characterize $R(X, Y, D)$, and analyze its relationship among different sources X, Y, Z, etc. as a notion of distance.

Remark 1. *The diagram shown in Figure 3.1 resembles the butterfly network in network coding [17]. As such, the coding diagram illustrated Figure 3.1 may be regarded as lossy network coding in the context of transmission. Also related works are Wyner-Ziv coding [20] and coding with multiple decoders with different side information considered by Kaspi [21] and Heegard & Berger [22].*

3.3 Distance Property and Bounds of R(X, Y, D)

Unless otherwise specified, in this section X, Y, Z, etc. denote arbitrary stationary sources. We begin with the distance property of $R(X, Y, D)$ among finite alphabets sources for a fixed $D \geq 0$.

3.3.1 Finite Alphabets

Suppose that both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are finite, and

$$\max_{x \in \mathbf{A}} \min_{\hat{x} \in \hat{\mathbf{A}}} d(x, \hat{x}) = 0. \tag{3.4}$$

For any $D \geq 0$, define

$$H(D) \triangleq \max H(U|\hat{U}) \tag{3.5}$$

where the maximum is taken over all random variables U and $\hat{U}$ taking values in $\mathbf{A}$ and $\hat{\mathbf{A}}$, respectively, such that $\mathbf{E}[d(U, \hat{U})] \leq D$, and $H(U|\hat{U})$ denotes the conditional entropy of U given $\hat{U}$. (All information quantities in this thesis are expressed in bits, and the function log is to base 2.)

Corollary 1. *It is easy to see that in the case where* $\mathbf{A} = \hat{\mathbf{A}}$ *and* d *is the Hamming distance measure on* $\mathbf{A}$*,*

$$H(D) = h(D) + D\log(|\mathbf{A}| - 1) \tag{3.6}$$

for any $0 \leq D \leq 1/2$*, where* $h(D) = -D\log D - (1-D)\log(1-D)$*, and* $|\mathbf{A}|$ *denotes the cardinality of* $\mathbf{A}$ *if* $\mathbf{A}$ *is a finite set.*

Proof. The Hamming distance measure is given by

$$d(u, \hat{u}) = \begin{cases} 0 & \text{if } u = \hat{u} \\ 1 & \text{if } u \neq \hat{u}, \end{cases} \tag{3.7}$$

which results in a probability of error distortion, since $\mathbf{E}[d(U, \hat{U})] = \Pr(U \neq \hat{U})$ [14]. Define an error random variable,

$$E \triangleq d(U, \hat{U}) = \begin{cases} 0 & \text{if } U = \hat{U} \\ 1 & \text{if } U \neq \hat{U}. \end{cases}$$

Then, using the chain rule for entropies to expand $H(U, H|\hat{U})$ in two different ways, we have

$$H(U, E|\hat{U}) = H(U|\hat{U}) + H(E|U, \hat{U}) \tag{3.8}$$

$$= H(E|\hat{U}) + H(U|E, \hat{U}) \tag{3.9}$$

Since E is a function of U and $\hat{U}$, the conditional entropy $H(E|U, \hat{U})$ is equal to zero. $H(U|E, \hat{U})$ can be wrote as follows:

$$\begin{aligned} H(U|E, \hat{U}) &= \Pr(E = 0)H(U|E = 0, \hat{U}) + \Pr(E = 1)H(U|E = 1, \hat{U}) \\ &\overset{(a)}{=} \Pr(U \neq \hat{U})H(U|U \neq \hat{U}, \hat{U}) \end{aligned}$$

where (a) follows since given $E = 0$, $U = \hat{U}$ and $H(U|E = 0, \hat{U}) = 0$. Now, by combining these results, we obtain

$$\begin{aligned} H(D) = \max H(U|\hat{U}) &= \max[H(E|\hat{U}) + \Pr(U \neq \hat{U})H(U|U \neq \hat{U}, \hat{U})] \\ &= h(D) + D\log(|\mathbf{A}| - \mathbf{1}) \end{aligned}$$

where the maximum is taken over all random variables U and $\hat{U}$ taking values in $\mathbf{A}$, such that $\mathbf{E}[d(U, \hat{U})] = \Pr(U \neq \hat{U}) \leq D \leq 1/2$. □

Theorem 1. *Fix $D \geq 0$. Let*

$$R^*(X,Y,D) = \begin{cases} R(X,Y,D) & \text{if } X = Y \\ R(X,Y,D) + H(D) & \text{otherwise.} \end{cases} \tag{3.10}$$

Then $R^*(X,Y,D)$ is a pseudo distance over the set of all stationary sources, i.e., satisfying the following three properties:

(1) $R^*(X,Y,D) = 0$ *if* $X = Y$.

(2) $R^*(X,Y,D) = R^*(Y,X,D)$.

(3) $R^*(X,Z,D) \leq R^*(X,Y,D) + R^*(Y,Z,D)$.

Proof. Properties (1) and (2) above follow immediately from the definition of $R(X,Y,D)$ along with Definitions 3.1 and 3.2. To prove Property (3), i.e., the triangle inequality, we first show that random encoding mappings in the definition of block codes C_n will not decrease the rate distortion function for any (X,Y). Given any X, Y, and Z, we then show that $(R(X,Y,D) + R(Y,Z,D) + H(D), D)$ is achievable by block codes with random encoding mappings for (X,Z). Thus,

$$R(X,Z,D) \leq R(X,Y,D) + R(Y,Z,D) + H(D)$$

from which Property (3) follows.
By using the union bound along with Definition 3.2, we have

$$\begin{aligned} \Pr\{d(X^n,\hat{X}^n) > D \textbf{ or } d(Y^n,\hat{Y}^n) > D\} &\leq \Pr\{d(X^n,\hat{X}^n) > D\} + \Pr\{d(Y^n,\hat{Y}^n) > D\} \\ &< \epsilon + \epsilon = 2\epsilon \end{aligned} \tag{3.11}$$

which is equivalent to

$$\Pr\{d(X^n,\hat{X}^n) \leq D \textbf{ and } d(Y^n,\hat{Y}^n) \leq D\} \geq 1 - 2\epsilon \tag{3.12}$$

Now, let W be an arbitrary random variable independent of the sources X and Y which denote the random codebook we use at the encoder and both decoders. By the law of total probability, one can easily rewrite 3.11 as

$$\begin{aligned} &\Pr\{d(X^n,\hat{X}^n) > D \textbf{ or } d(Y^n,\hat{Y}^n) > D\} \\ &= \sum_{w \in W} \Pr\{d(X^n,\hat{X}^n) > D \textbf{ or } d(Y^n,\hat{Y}^n) > D \mid W = w\}.\Pr\{W = w\} < 2\epsilon \end{aligned} \tag{3.13}$$

We know that if the weighted average of a series of numbers tends to zero, surely one or some of these numbers will tend to zero. The summation in 3.13 is the weighted average of a series of probability which tends to zero as $\epsilon \to 0$. Then there exists the specific realization of W such that

$$\{w^* \mid \Pr\{d(X^n, \hat{X}^n) > D \quad \textbf{or} \quad d(Y^n, \hat{Y}^n) > D \mid W = w^*\} \to 0\} \neq \emptyset \tag{3.14}$$

or equivalently,

$$\{w^* \mid \Pr\{d(X^n, \hat{X}^n) \leq D \textbf{ and } d(Y^n, \hat{Y}^n) \leq D \mid W = w^*\} \to 1\} \neq \emptyset. \tag{3.15}$$

Therefore, the randomization in the definition of block codes C_n will not decrease the rate distortion function.

To prove the second part, we first assumed that $R(X, Y, D)$ and $R(Y, Z, D)$ are achievable by block codes with two deterministic mappings for (X, Y) and (Y, Z), respectively. (Figures 3.2 and 3.3)

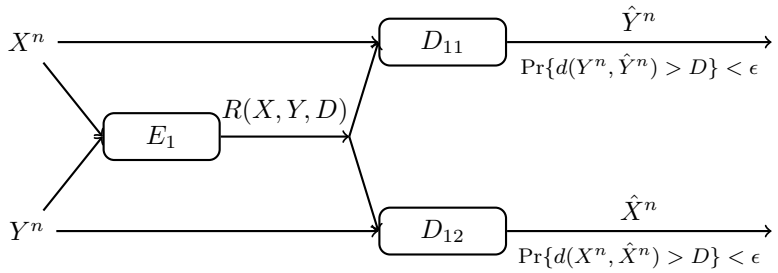

Figure 3.2: Deterministic mappings for achievablity of $R(X, Y, D)$.

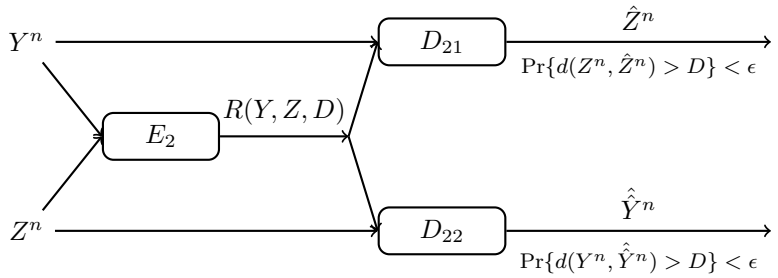

Figure 3.3: Deterministic mappings for achievablity of $R(Y, Z, D)$.

Given any stationary sources X, Y, and Z, we then show that the pair $(R(X,Y,D) + R(Y,Z,D) + H(D), D)$ is achievable by block codes with random encoding mappings for (X,Z). To achieve the aforementioned rate, we construct the random coding based on the proposed coding diagram which is depicted in Figure 3.4.

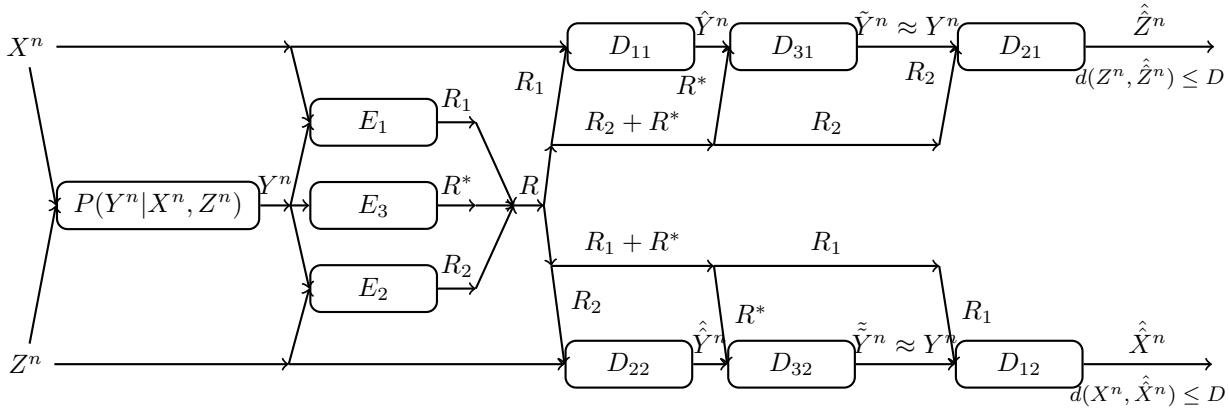

Figure 3.4: Proposed random coding mappings for achievability of $(R(X,Y,D) + R(Y,Z,D) + H(D), D)$ respect to (X,Z).

First, we create artificial Y^n given two realization X^n and Z^n based on $P(Y^n|X^n,Z^n)$. Then the early deterministic encoder E_1 (E_2, resp.), encodes (X^n,Y^n) ((Y^n,Z^n), resp.) to the codewords R_1 (R_2, resp.) which is roughly equal to $R(X,Y,D)$ ($R(Y,Z,D)$, resp.). The early decoder D_{11} (D_{22}, resp.) uses the rate R_1 (R_2, resp.) to create $\hat{Y}^n$ ($\hat{\hat{Y}}^n$, resp.) in such a way that $\hat{Y}^n$ ($\hat{\hat{Y}}^n$, resp.) is in the D-ball [1] of Y^n. As well, we need to construct the random encoder E_3 for the purpose of encoding Y^n to the random codeword R^* which would ,in turn, convert $\hat{Y}^n$ into $\tilde{Y}^n$ via decoder D_{31}, and $\hat{\hat{Y}}^n$ into $\hat{\tilde{\tilde{Y}}}^n$ via decoder D_{32} such that $\tilde{Y}^n$ and $\hat{\tilde{\tilde{Y}}}^n$ are equal to the Y^n with high probability. In the end, the early decoder D_{21} (D_{12}, resp.) along with the rates R_2 (R_1, resp.) build up $\hat{\hat{Z}}^n$ ($\hat{\hat{X}}^n$, resp.) such that the distortion between Z and $\hat{\hat{Z}}^n$ (X and $\hat{\hat{X}}^n$, resp.) is less than or equal to a prescribed threshold D.
To fully understand and characterize this important and rather surprising random coding, we are led to the following questions:

[1] A distortion ball $B(y^n, D)$ centered on $y^n \in \mathbf{A}^n$ with radius $D \ge 0$ is defined by $B(y^n, D) = \{x^n \in \mathbf{A}^n : d(x^n, y^n) \le D\}$ [23].

Q1. What is the coding schemes within the swapping part i.e., what is the random encoder E_3 and two decoders D_{31} and D_{32}? Is R^* equal to defined $H(D)$ in 3.5?

Q2. Is $\hat{\hat{Z}}^n$ inside of the D-ball of Z^n? There is the same question for another output, $\hat{\hat{X}}^n$.

For the first question, we use random binning to show the achievabiliy of R^* such that the probability of error, $P(\mathcal{E}) = \Pr\{\tilde{Y}^n \neq Y^n \textbf{ or } \tilde{\tilde{Y}}^n \neq Y^n\}$, tends to zero for all sufficiently large n. (Figure 3.5)

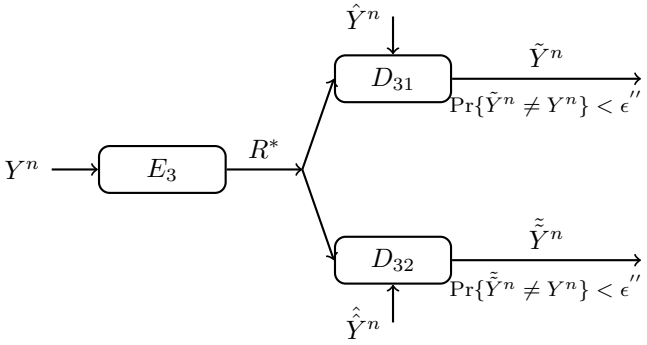

Figure 3.5: Lossless source coding with two decoders and side information.

The following provides the details.

Codebook generation: Randomly and independently set an index $m(y^n) \in [1 : 2^{nR_*}]$ to each sequence $y^n \in \mathbf{A^n}$ according to a uniform PMF over $[1 : 2^{nR_*}]$. We refer to each subset of sequences with the same index m as a bin $\mathcal{B}(m)$, $m \in [1 : 2^{nR_*}]$. The bin assignments are revealed to the encoder and both decoders.

Encoding: Upon receiving $y^n \in \mathcal{B}(m)$, the encoder E_3 sends the bin index m.

Decoding: Given the received index m, the decoder D_{31} (D_{32}, resp.) declares $\tilde{Y}^n$ ($\tilde{\tilde{Y}}^n$, resp.) to be the estimate of the source sequence if it is the unique sequence in $\mathcal{B}(m)$ which has the distance less than or equal to D from $\hat{Y}^n$ ($\hat{\hat{Y}}^n$, resp.); otherwise it declares an error.

Analysis of the probability of error: We bound the probability of error averaged over bin assignments. Let M denote the random bin indices for Y^n i.e., $Y^n \in \mathcal{B}(M)$. It is important to mention that $M \sim Unif[1 : 2^{nR^*}]$ is independent of Y^n. The decoders make

an error if and only if one or more of the following events occur:

$$\mathcal{E}_1 = \{d(Y^n, \hat{Y}^n) > D\},$$
$$\mathcal{E}_2 = \{d(Y^n, \hat{\hat{Y}}^n) > D\},$$
$$\mathcal{E}_3 = \{y'^n \in \mathcal{B}(M) \text{ for some } y'^n \neq Y^n,\ d(y'^n, \hat{Y}^n) \leq D \text{ and } d(Y^n,\ \hat{Y}^n) \leq D\},$$
$$\mathcal{E}_4 = \{y''^n \in \mathcal{B}(M) \text{ for some } y''^n \neq Y^n,\ d(y''^n, \hat{\hat{Y}}^n) \leq D \text{ and } d(Y^n, \hat{\hat{Y}}^n) \leq D\}.$$

Then, by the symmetry of codebook construction and the union of events bound, the average probability of error is upper bounded as:

$$\begin{aligned} P(\mathcal{E}) &\leq P(\mathcal{E}_1) + P(\mathcal{E}_2) + P(\mathcal{E}_3) + P(\mathcal{E}_4) \\ &= P(\mathcal{E}_1) + P(\mathcal{E}_2) + P(\mathcal{E}_3 | Y^n \in \mathcal{B}(1)) + P(\mathcal{E}_4 | Y^n \in \mathcal{B}(1)) \end{aligned} \tag{3.16}$$

We now bound each term. With the assistant of two distortion conditions for the early decoders D_{11} and D_{22} i.e.,

$$\Pr\{\ d(Y^n, \hat{Y}^n) > D\} < \epsilon$$

and

$$\Pr\{\ d(Y^n, \hat{\hat{Y}}^n) > D\} < \epsilon,$$

$P(\mathcal{E}_1)$ and $P(\mathcal{E}_2)$ tends to zero if $\epsilon \to 0$ as $n \to \infty$, respectively. To bound the third term, we have

$$P(\mathcal{E}_3) = P(\mathcal{E}_3 | Y^n \in \mathcal{B}(1)) = \sum_{(y^n, \hat{y}^n)} \Pr\{(Y^n, \hat{Y}^n) = (y^n, \hat{y}^n) \mid Y^n \in \mathcal{B}(1)\}$$

$$. \Pr\{y'^n \in \mathcal{B}(1) \text{ for some } y'^n \neq y^n, d(y'^n\ \hat{y}^n) \leq D \text{ and } d(y^n,\ \hat{y}^n) \leq D \mid y^n \in \mathcal{B}(1), (Y^n, \hat{Y}^n) = (y^n, \hat{y}^n)\} \overset{(a)}{\leq} \sum_{(y^n, \hat{y}^n)} p(y^n, \hat{y}^n) \sum_{\substack{y'^n \in \{y^{\star n} \in \mathbf{A}^n\,:\, d(y^{\star n}, \hat{y}^n) \leq D\} \\ y'^n \neq y^n}} \Pr\{y'^n \in \mathcal{B}(1)\} \leq$$

$$\max_{\hat{y}^n \in \hat{\mathbf{A}}^n} |\{y^{\star n} \in \mathbf{A}^n : d(y^{\star n}, \hat{y}^n) \leq D\}| . 2^{-nR^*} \overset{(b)}{\leq} 2^{nH(D) + O(\log n)} . 2^{-nR^*} = 2^{-n(R^* - H(D)) + O(\log n)} \tag{3.17}$$

where (a) follows since for every $y'^n \neq y^n$, the events $\{y^n \in \mathcal{B}(1)\}$, $\{y'^n \in \mathcal{B}(1)\}$, and $\{(Y^n, \hat{Y}^n) = (y^n, \hat{y}^n)\}$ are mutually independent. In the inequality (b), we find an upper bound on the cardinality of any D-ball with any center. Let s be a joint n-type in $\mathcal{P}_n(\mathbf{A} \times \hat{\mathbf{A}})$, then

$$t(i) = \sum_{j \in \hat{\mathbf{A}}} s(i, j)$$

defines an n-type in $\mathcal{P}_n(\mathbf{A})$ which is called the marginal n-type of s on $\mathbf{A}$. Similarly, we can define the marginal n-type r of s on $\hat{\mathbf{A}}$[23]. Yang *et al.* [23] showed that when $\hat{y}^n \in T^n_{\hat{\mathbf{A}}}(r)$ the cardinality of the restricted D-ball [2] is upper bounded as:

$$|B(\hat{y}^n, t, D)| \leq 2^{n(\max_{s\in\mathcal{S}(t,r,D)} H(s)-H(r))+O(\log n)} \tag{3.18}$$

where $H(s)$ and $H(r)$ denotes the entropy of the distribution s and the distribution r, respectively

$$\begin{aligned} H(s) &= -\sum_{i\in\mathbf{A}, j\in\hat{\mathbf{A}}} s(i,j)\log s(i,j), \\ H(r) &= -\sum_{j\in\hat{\mathbf{A}}} r(j)\log r(j), \end{aligned} \tag{3.19}$$

and

$$\mathcal{S}(t,r,D) = \{s \in \mathcal{P}_n(\mathbf{A}\times\hat{\mathbf{A}}) : t \text{ and } r \text{ are the two marginals of} s \text{ and } \mathbf{E}[d(Y,\hat{Y})] \leq D\}.$$

It is easy to see that in 3.18, the cardinality of $B(\hat{y}^n, t, D)$ depends on $\hat{y}^n$ only through the type of $\hat{y}^n$ i.e., $\hat{y}^n \in T^n_{\hat{\mathbf{A}}}(r)$. In our problem, D-ball is not restricted and the center $\hat{y}^n$ can be any sequence. Thus, simply summarizing the cardinalities of restricted D-balls with a given center over all types would give us

$$\begin{aligned} |B(\hat{y}^n, D)| &= \sum_{t\in\mathcal{P}_n(\mathbf{A})} |B(\hat{y}^n, t, D)| \leq (n+1)^{|\mathbf{A}|} \cdot 2^{n(\max_{s\in\mathcal{S}(r,D)} H(s)-H(r))+O(\log n)} \\ &= 2^{n(\max_{s\in\mathcal{S}(r,D)} H(s)-H(r))+|A|\log(n+1)+O(\log n)} \\ &= 2^{n(\max_{s\in\mathcal{S}(r,D)} H(s)-H(r))+O(\log n)} \end{aligned} \tag{3.20}$$

where

$$\mathcal{S}(r,D) = \{s : r \in \mathcal{P}_n(\hat{\mathbf{A}}) \text{ is the marginal of } s \text{ and, } \mathbf{E}[d(Y,\hat{Y})] \leq D\}.$$

We found the first order upper bound for a given center. Finally, the upper bound for any center (not only $\hat{y}^n \in T^n_{\hat{\mathbf{A}}}(r)$) and any D-ball can be written as

$$\begin{aligned} |B(\hat{Y}^n, D)| &\leq \max_{\hat{y}^n\in\hat{\mathbf{A}}^n} |\{y^{\star n} \in \mathbf{A}^n : d(y^{\star n}, \hat{y}^n) \leq D\}| \\ &\leq 2^{n(\max[H(U,\hat{U})-H(\hat{U})])+O(\log n)} = 2^{n(\max H(U|\hat{U}))+O(\log n)} \\ &= 2^{nH(D)+O(\log n)} \end{aligned} \tag{3.21}$$

[2] For $t \in \mathcal{P}_n(\mathbf{A})$, we define the restricted D-ball as $B(\hat{y}^n, t, D) = B(\hat{y}^n, D) \cap T^n_{\mathbf{A}}(t)$.

where the maximum in 3.21 is taken over all random variables U and $\hat{U}$ taking values in $\mathbf{A}$ and $\hat{\mathbf{A}}$, respectively, such that $\mathbf{E}[d(U,\hat{U})] \leq D$. This completes the proof of the inequality (b) in 3.17.
Careful examination on 3.17 reveals that $P(\mathcal{E}_3)$ tends to zero as $n \to \infty$ if $R^* > H(D)$. Similarly, $P(\mathcal{E}_4)$ tends to zero if $R^* > H(D)$ for all sufficiently large n. Thus, the probability of error averaged over bin assignments tends to zero as $n \to \infty$ if $R^* > H(D)$. Therefore, there exists a sequence of bin assignments such that $\lim_{n\to\infty} P(\mathcal{E}) = 0$. This completes the achievability proof of the lossless source coding with two decoders and side information. (Figure 3.5.)

To answer the second question, with the help of the deterministic coding for (Y,Z) (Figure 3.3) and the random coding for Y^n (Figure 3.5), we have

$$\Pr\{d(Z^n,\hat{Z}^n) > D\} < \epsilon$$

and

$$\Pr\{Y^n \neq \tilde{Y}^n\} < \epsilon'',$$

respectively. Given the rate R_2 and $\tilde{Y}^n$ as inputs for the decoder D_{21}, by the union bound and the above inequalities, we can write

$$\Pr\{d(Z^n,\hat{Z}^n) > D \textbf{ or } Y^n \neq \tilde{Y}^n\} \leq \Pr\{d(Z^n,\hat{Z}^n) > D\} + \Pr\{Y^n \neq \tilde{Y}^n\} < \epsilon + \epsilon'' \quad (3.22)$$

or equivalently

$$\Pr\{d(Z^n,\hat{Z}^n) \leq D \textbf{ and } Y^n = \tilde{Y}^n\} \geq 1 - (\epsilon + \epsilon'') \quad (3.23)$$

Now, if both $\epsilon, \epsilon'' \to 0$, then the output of the decoder D_{21} in proposed coding diagram (Figure 3.4) will be inside of the D-ball of Z^n which mathematically is equivalent to

$$\Pr\{d(Z^n,\hat{\tilde{Z}}^n) > D\} < \epsilon + \epsilon'' \stackrel{\Delta}{=} \epsilon'$$

Similarly, the distortion between X^n and $\hat{\tilde{X}}^n$ is less than or equal to a prescribed threshold D with the high probability. This completes the proof of Theorem 1. □

3.3.2 Abstract Alphabets

In this subsection, both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are abstract. As usual, however, we assume that the distortion measure d and stationary sources $X = \{X_i\}_{i=1}^{\infty}$ satisfy the following condition:

$$\mathbf{E}[d(X_1,\hat{x})] < \infty \quad (3.24)$$

for some $\hat{x} \in \hat{\mathbf{A}}$ which means for any symbol $x \in \mathbf{A}$, there exists a reconstruction symbol $\hat{x} \in \hat{\mathbf{A}}$ such that $d(x,\hat{x})$ is bounded. Then, we have the following result.

Theorem 2. *Let $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ be a stationary, ergodic pair with each of X and Y satisfying 3.24. Let $R_{X|Y}(D)$ ($R_{Y|X}(D)$, resp.) denote the conditional rate distortion function of X (Y, resp.) given Y (X, resp.). Then the following holds:*

$$\max\{R_{X|Y}(D), R_{Y|X}(D)\} \leq R(X,Y,D) \tag{3.25}$$

$$\leq \inf\{\max\{I(Y_1;U|X_1), I(X_1;U|Y_1)\} + I(X_1;\hat{X}_1|Y_1U) + I(Y_1;\hat{Y}_1|X_1U) : U, \hat{X}_1, \hat{Y}_1\} \tag{3.26}$$

where the infimum is taken over all random variables U, $\hat{X}_1$, and $\hat{Y}_1$ such that $\mathbf{E}[d(X_1,\hat{X}_1)] \leq D$ and $\mathbf{E}[d(Y_1,\hat{Y}_1)] \leq D$, and I denotes the mutual information and conditional mutual information, as the case may be.

Proof. $R(X,Y,D)$ should be at least as large as the smallest rate needed to encode X^n at the Encoder for decoding by the Decoder 2, while ignoring the distortion condition for Y^n i.e, 3.2 at the Decoder 1. The smallest such rate is given by the conditional rate distortion function of 2.19 [24]. Inequality 3.25 is followed by combining the above cut-set lower bound with the converse for lossy source coding with side information at the encoder and the decoder [25]. The proof of the upper bound for $R(X,Y,D)$ follows directly from Heegard and Berger's bound [22, Th.2] and is omitted. □

Remark 2. *Lower bound in 3.25 is the tightest lower bound for $R(X,Y,D)$ in the literature. Heegard and Berger's upper bound is less than Kimura-Uyematso's bound [26] and is the tightest upper bound in the literature, too. When the sources are IID, we have the closed form expression for it based on the [26].*

The following definition describes a binary source for which the conditional rate distortion is tight for all distortions.

Definition 3.3. *The source (X_1,Y_1) is said to be a DSBS with cross-over probability p if $\mathbf{A}$ and $\hat{\mathbf{A}}$ are binary alphabets, $0 \leq p < 1/2$, and for each $(x,y) \in (\mathbf{A} \times \mathbf{A})$ we have*

$$p_{X_1Y_1}(x,y) = \begin{cases} 0.5(1-p) & \text{if } x = y \\ 0.5p & \text{otherwise.} \end{cases} \tag{3.27}$$

Equivalently, (X_1,Y_1) is a DSBS if

$$X_1 = Y_1 \oplus N_1$$

where $\oplus$ denotes module-two (binary) addition, Y_1 is uniform on $\{0,1\}$, and N_1 is independent of Y_1 with $\Pr\{N_1 = 1\} = p < 1/2$ and $\Pr\{N_1 = 0\} = 1-p$.

Here, we show that for a DSBS, the cut-set lower bound in 3.25 is tight.

Corollary 2. *Let source (X_1, Y_1) be a DSBS and d be the Hamming distortion measure over $\{0,1\}$. Suppose that $(X,Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$ is IID. In this case, it follows from 2 and [24] that*

$$R(X,Y,D) = \max\{R_{X|Y}(D), R_{Y|X}(D)\} = R_{X|Y}(D) \tag{3.28}$$

$$= \begin{cases} h(p) - h(D) & \text{if } 0 \le D < p \\ 0 & \text{otherwise.} \end{cases} \tag{3.29}$$

where p is cross-over probability and $h(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ is the binary entropy function (set $h(0) = 0$). Note that by symmetry, $R_{X|Y}(D) = R_{Y|X}(D)$.

Proof. We now explain a coding scheme that achieves the cut-set lower bound [24]. If $D \ge p$, then each decoder in Figure 3.1 can form its reconstruction directly from its side information i.e., Decoder 1 decodes $\hat{X}^n = Y^n$ and Decoder 2 decodes $\hat{Y}^n = X^n$. Therefore, we have

$$R(X,Y,D) = \max\{R_{X|Y}(D), R_{Y|X}(D)\} = 0, \qquad D \ge p.$$

Suppose that $0 \le D < p$. The encoder adds (module two) the symbols of X^n and Y^n:

$$N_i = X_i \oplus Y_i, \qquad i = 1, 2, \cdots, n.$$

where $N_1, N_2, \cdots, N_n$ are IID with $\Pr\{N = 1\} = p < 1/2$ and $\Pr\{N = 0\} = 1 - p$. The encoder sends a compression version $\hat{N}^n$ of N^n to both decoders with a Hamming distortion of D. This compression can be achieved with a rate arbitrarily close to the rate distortion function of N, which is given by [14, Th.10.3.1]

$$R_N(D) = h(p) - h(D), \qquad 0 \le D < p.$$

Decoder 1 reconstructs $\hat{Y}^n$ from $\hat{N}^n$ and X^n by setting

$$\hat{Y}_i = \hat{N}_i \oplus X_i, \qquad i = 1, 2, \cdots, n.$$

Similarly, Decoder 2 decodes $\hat{X}^n$ from $\hat{N}^n$ and Y^n by setting

$$\hat{X}_i = \hat{N}_i \oplus Y_i, \qquad i = 1, 2, \cdots, n.$$

Both constructions have the distortion D. So, the rate distortion function for a DSBS with Hamming distance measure is

$$R(X,Y,D) = \begin{cases} h(p) - h(D) & \text{if } 0 \le D < p \\ 0 & \text{otherwise.} \end{cases}$$

Note that, both conditional rate distortion functions $R_{X|Y}(D)$ and $R_{Y|X}(D)$ for a DSBS was shown in [25] which is equal to 3.29. This completes the proof of Corollary 2. □

The next corollary checks the information distance (Theorem 1) for DSBS with Hamming distortion measure.

Corollary 3. *Suppose that* (X_1, Y_1), (Y_1, Z_1), *and* (X_1, Z_1) *are DSBSs and* $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$, $(Y, Z) = \{(Y_i, Z_i)\}_{i=1}^{\infty}$, *and* $(X, Z) = \{(X_i, Z_i)\}_{i=1}^{\infty}$ *are IID. Then with Hamming distortion function over* $\{0, 1\}$, $R^*(X, Y, D)$ *is a pseudo distance over the set of all DSBSs.*

Proof. To show that $R^*(X, Y, D)$ is a pseudo distance, as usual, we have to satisfy all three distance properties. The following provides the details.

(I) If $X = Y$, then immediately from the definition of $R(X, Y, D)$ along with Definitions 3.1 and 3.2, we have

$$R^*(X, Y, D) = R(X, X, D) = 0$$

which satisfy the identity of indiscernibles property. In the other word, each decoder can form its reconstruction directly from its side information.

(II) For the symmetry property, if $X = Y$,

$$R^*(X, Y, D) = R^*(Y, X, D) = 0$$

Otherwise,

$$\begin{aligned} R^*(X, Y, D) &= R(X, Y, D) + H(D) \\ &\overset{(a)}{=} R_{X|Y}(D) + H(D) \\ &\overset{(b)}{=} R_{Y|X}(D) + H(D) \\ &\overset{(a)}{=} R(Y, X, D) + H(D) \\ &= R^*(Y, X, D) \end{aligned}$$

where (a) and (b) are followed from equation 3.28 and the symmetry property of $R(X|Y)$ for DSBS, respectively.

(III) The triangle inequality: According to Definition 3.3, since (X_1, Y_1) and (Y_1, Z_1) are DSBSs, and both $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$ and $(Y, Z) = \{(Y_i, Z_i)\}_{i=1}^{\infty}$ are IID, we have

$$X_1 = Y_1 \oplus N_1$$

$$Y_1 = Z_1 \oplus N_1'.$$

where N_1 and N_2 are independent of each other. One can immediately write

$$X_1 = (Z_1 \oplus N_1') \oplus N_1 = Z_1 \oplus (N_1 \oplus N_1') = Z_1 \oplus N_1''$$

where $N_1'' = N_1 \oplus N_1'$. Now, we can analyze the triangle inequality. Consider five different cases: (I) $X = Y = Z$, (II) $X = Y \neq Z$, (III) $X \neq Y = Z$, (IV) $X = Z \neq Y$ and (V) $X \neq Y \neq Z$. In Case (I), the triangle inequality happens with the equality. (Pseudo distance over all sources will be 0.) In Cases (II), and (III), equality is also satisfied but the pseudo distance is zero only over the sources (X, Y) and (Y, Z), respectively. Careful examination on Theorem 1 reveals that R^* always is a non-negative function. Therefore, the triangle inequality occurs in Case (IV), too. Finally, in the last case, we must show that

$$R^*(X,Z,D) \leq R^*(X,Y,D) + R^*(Y,Z,D) \Rightarrow R(X,Z,D) \leq R(X,Y,D) + R(Y,Z,D) + H(D)$$

is valid. Since the distortion measure is Hamming distance and $\mathbf{A}$ is a binary alphabet, 3.6 is simplified to

$$H(D) = h(D) + D\log(|\mathbf{A}| - 1) = h(D).$$

Without loss of generality, we assume that $Pr\{N_1 = 1\} = p_1 \leq Pr\{N_1' = 1\} = p_2 < 1/2$. By law of total probability and $N_1'' = N_1 \oplus N_1'$, we have

$$\begin{aligned} Pr\{N_1'' = 1\} &= Pr\{N_1'' = 1 | N_1' = 1\} Pr\{N_1' = 1\} + Pr\{N_1'' = 1 | N_1' = 0\} Pr\{N_1' = 0\} \\ &= Pr\{N_1 = 0\} Pr\{N_1' = 1\} + Pr\{N_1 = 1\} Pr\{N_1' = 0\} \\ &= (1 - p_1)p_2 + p_1(1 - p_2) = p_1 * p_2 \end{aligned}$$

and consequently $p_1 \leq p_2 \leq (p_1 * p_2) < 1/2$. Equation 3.29 gives us this fact which we need to analyze four separate intervals for D:
1. When $0 \leq D < \min\{p_1, p_2, (p_1 * p_2)\} = p_1$,

$$\begin{aligned} R(X,Z,D) &\leq R(X,Y,D) + R(Y,Z,D) + H(D) \\ h(p_1 * p_2) - h(D) &\leq h(p_1) - h(D) + h(p_2) - h(D) + h(D) \\ h(p_1 * p_2) &\leq h(p_1) + h(p_2) \\ 0 &\leq h(p_1) + h(p_2) - h(p_1 * p_2) \end{aligned}$$

The Figure 3.6 shows that the function $f(p_1, p_2) \stackrel{\Delta}{=} h(p_1) + h(p_2) - h(p_1 * p_2)$ is always non-negative, hence the triangle inequality will hold in this interval.

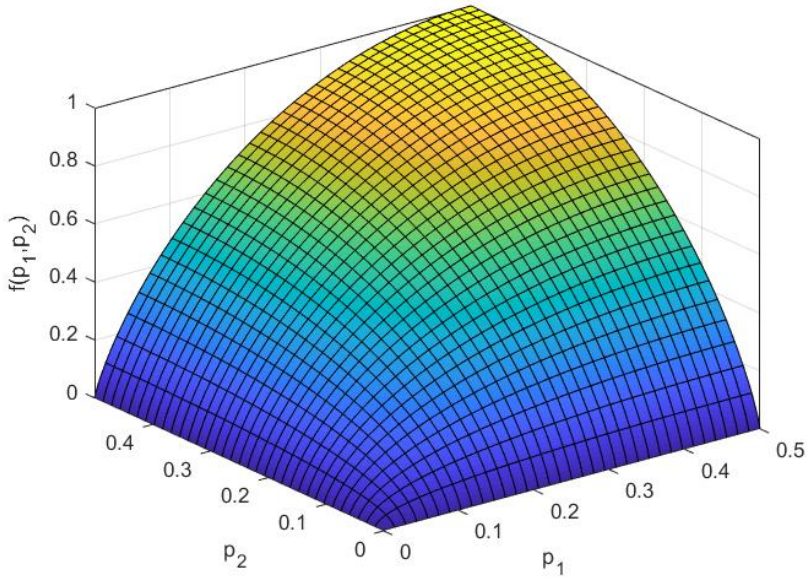

Figure 3.6: $f(p_1, p_2)$ for all p_1, $p_2 < 1/2$.

2. If $p_1 \leq D < p_2$, then $R(X, Y, D) = 0$ and

$$\begin{aligned} R(X, Z, D) &\leq R(X, Y, D) + R(Y, Z, D) + H(D) \\ h(p_1 * p_2) - h(D) &\leq h(p_2) - h(D) + h(D) \\ 0 &\leq h(p_2) + h(D) - h(p_1 * p_2) \end{aligned}$$

Since $p_1 \leq D < p_2 < 1/2$, we have $h(p_1) \leq h(D)$ and then

$$0 \leq f(p_1, p_2) = h(p_1) + h(p_2) - h(p_1 * p_2) \leq h(D) + h(p_2) - h(p_1 * p_2)$$

Therefore, the triangle inequality is satisfied when $p_1 \leq D < p_2$.

3. When $p_2 \leq D < (p_1 * p_2) < 1/2$, then $R(X, Y, D) = R(Y, Z, D) = 0$ and

$$\begin{aligned} R(X, Z, D) &\leq R(X, Y, D) + R(Y, Z, D) + H(D) \\ h(p_1 * p_2) - h(D) &\leq h(D) \\ 0 &\leq h(D) + h(D) - h(p_1 * p_2) \end{aligned}$$

like the previous interval, $h(p_1) \leq h(p_2) \leq h(D)$ and we can write

$$0 \leq f(p_1, p_2) = h(p_1) + h(p_2) - h(p_1 * p_2) \leq h(D) + h(D) - h(p_1 * p_2)$$

4. In the last interval i.e., $(p_1 * p_2) \leq D < 1/2$, all of the rate distortions will be zero and the triangle inequality is converted to $0 \leq h(D)$, which is always true. The proof of any permutation of $R^*(.,.,D)$ in triangle inequality is the same as the aforementioned method. This completes the proof of Corollary 3. □

Theorem 2 yields the next corollary for jointly Gaussian sources with quadratic distortion measure.

Corollary 4. *Suppose that X_1 and Y_1 are jointly Gaussian, and $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ is IID. Then with $d(x,\hat{x}) = (x-\hat{x})^2$, both the lower bound 3.25 and upper bound 3.26 are tight, and*

$$R(X,Y,D) = \begin{cases} \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D} & \text{if } 0 \leq D < (1-\rho^2)\sigma^2 \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

where ρ is the correlation between X_1 and Y_1, and σ^2 is the maximum of the variances of X_1 and Y_1.

Proof. Without loss of generality, assume that $\mathbf{E}X_1 = \mathbf{E}Y_1 = 0$, and $\sigma^2 = \sigma_X^2 \geq \sigma_Y^2$, where σ_X^2 and σ_Y^2 are the variances of X_1 and Y_1, respectively. Consider three cases: (1) $D \geq (1-\rho^2)\sigma^2$, (2) $D < (1-\rho^2)\sigma_Y^2$, and (3) $(1-\rho^2)\sigma_Y^2 \leq D < (1-\rho^2)\sigma^2$. In Case (1), X and Y can be estimated directly from each other. Specifically, let $\hat{X}_1 = \rho\frac{\sigma}{\sigma_Y}Y_1$ and $\hat{Y}_1 = \rho\frac{\sigma}{\sigma_X}X_1$. The resulting $\hat{X}$ and $\hat{Y}$ satisfy the distortion requirement, and we do not need any information from the encoder to decoders. Hence, $R(X,Y,D) = 0$.

In Case (2), let

$$a = \frac{(1-\rho^2)\sigma^2 - D}{(1-\rho^2)\sigma^2} \text{ and } b = \frac{(1-\rho^2)\sigma_Y^2 - D}{(1-\rho^2)\sigma_Y^2}.$$

Define

$$V = a(X_1 + N_1) \text{ and } W = b(Y_1 + N_2) \tag{3.31}$$

where N_1 and N_2 are zero mean Gaussian random variables with variances

$$\frac{D(1-\rho^2)\sigma^2}{(1-\rho^2)\sigma^2 - D} \text{ and } \frac{D(1-\rho^2)\sigma_Y^2}{(1-\rho^2)\sigma_Y^2 - D}$$

respectively. Furthermore, N_1 and N_2 are independent of each other and of both X_1 and Y_1. Let

$$\hat{X}_1 = V + (1-a)\rho\frac{\sigma}{\sigma_Y}Y_1 \text{ and } \hat{Y}_1 = W + (1-b)\rho\frac{\sigma_Y}{\sigma}X_1. \tag{3.32}$$

One can verify that

$$\mathbf{E}[X_1 - \hat{X}_1]^2 = D \text{ and } \mathbf{E}[Y_1 - \hat{Y}_1]^2 = D. \tag{3.33}$$

Let $U = (V, W)$. Observe that if we are given the values of X_1 and U (Y_1 and U, resp.), then we can calculate $\hat{Y}_1$ ($\hat{X}_1$, resp.). Given the values of (X_1, U) ((Y_1, U), resp.), the random variable $\hat{Y}_1$ ($\hat{X}_1$) will be a constant. Plugging U, $\hat{X}_1$, and $\hat{Y}_1$ into the respective information quantities in 3.26, we have

$$I(Y_1; \hat{Y}_1 | X_1 U) = 0 \tag{3.34}$$

$$I(X_1; \hat{X}_1 | Y_1 U) = 0 \tag{3.35}$$

and

$$\begin{aligned}
I(X_1; U|Y_1) &= I(X_1; VW|Y_1) \\
&= I(X_1; V|Y_1) + I(X_1; W|Y_1 V) \\
&= I(X_1; V|Y_1) && (3.36) \\
&= H(V|Y_1) - H(V|Y_1 X_1) \\
&= H(V|Y_1) - H(V|X_1) && (3.37) \\
&= H(V - a\rho\frac{\sigma}{\sigma_Y}Y_1|Y_1) - H(aN_1) \\
&= H(V - a\rho\frac{\sigma}{\sigma_Y}Y_1) - H(aN_1) && (3.38) \\
&= \frac{1}{2}\log 2\pi e a^2 \frac{[(1-\rho^2)\sigma^2]^2}{(1-\rho^2)\sigma^2 - D} - H(aN_1) && (3.39) \\
&= \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D} && (3.40)
\end{aligned}$$

where 3.36 and 3.37 are due to 3.31 which implies the conditional independence of W and (X_1, V) given Y_1, and the conditional independence of V and Y_1 given X_1 [3]; and 3.38 follows from the fact that under the joint Gaussian assumption, $V - a\rho\frac{\sigma}{\sigma_Y}Y_1$ is independent

[3]The long chain V —X_1 —Y_1 —W is equivalent to the three chains V —(X_1, Y_1) —W, V —X_1 —Y_1, and X_1 —Y_1 —W. Since $I(W; X_1 V|Y_1) = I(W; X_1|Y_1) + I(W; V|X_1 Y_1) = I(W; V|Y_1) + I(W; X_1|Y_1 V)$, we can easily conclude that $I(W; X_1|Y_1 V) = 0$.

of Y_1 [4]. In parallel with 3.40, we have

$$\begin{aligned} I(Y_1;U|X_1) &= I(Y_1;VW|X_1) \\ &= I(Y_1;W|X_1) + I(Y_1;V|X_1W) \\ &= I(Y_1;W|X_1) \\ &= H(W|X_1) - H(W|Y_1X_1) \\ &= H(W|X_1) - H(W|Y_1) \\ &= H(W - b\rho\frac{\sigma_Y}{\sigma}X_1|X_1) - H(bN_2) \\ &= H(W - b\rho\frac{\sigma_Y}{\sigma}X_1) - H(bN_2) \qquad (3.41) \\ &= \frac{1}{2}\log 2\pi e b^2 \frac{[(1-\rho^2)\sigma_Y^2]^2}{(1-\rho^2)\sigma_Y^2 - D} - H(bN_2) \qquad (3.42) \\ &= \frac{1}{2}\log\frac{(1-\rho^2)\sigma_Y^2}{D}. \qquad (3.43) \end{aligned}$$

By combining the information quantities in 3.43, 3.40, 3.35, and 3.34 together, it follows from 3.26 that

$$R(X,Y,D) \le \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D}.$$

This, together with 3.25 and

$$R_{X|Y}(D) = \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D}$$

implies 3.30 in Case (2).

In Case (3), Y can be estimated directly from X. Specifically, let $\hat{Y}_1 = \rho\frac{\sigma_Y}{\sigma}X_1$. With V and $\hat{X}_1$ defined as in 3.31 and 3.32, respectively, we now let $U = V$. Plug U, $\hat{X}_1$, and $\hat{Y}_1$ into the respective information quantities in 3.26. Again, 3.30 follows from a similar argument to the above. This completes the proof of Corollary 4. □

Careful examination reveals that the equal sign = in 3.38, 3.39, 3.41, and 3.42 can be replaced by $\le$ when X_1 and Y_1 are not necessarily jointly Gaussian. Therefore, the above argument also shows that the right side of 3.30 is actually an upper bound to $R(X,Y,D)$ for any real-valued sources X and Y satisfying 3.24 with $d(x,\hat{x}) = (x-\hat{x})^2$, which is stated as a corollary below.

[4] If two random variables X_1 and Y_1 are jointly Gaussian and are uncorrelated, then they are independent [27, Ch.6]. In this case $\mathbf{E}[(V - a\rho\frac{\sigma}{\sigma_Y}Y_1)Y_1] = \mathbf{E}[V - a\rho\frac{\sigma}{\sigma_Y}Y_1]\mathbf{E}[Y_1] = 0$

Corollary 5. *Let $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ be a real-valued and IID source pair with each satisfying 3.24 with $d(x,\hat{x}) = (x-\hat{x})^2$. Then*

$$R(X,Y,D) \leq \begin{cases} \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D} & \text{if } 0 \leq D < (1-\rho^2)\sigma^2 \\ 0 & \text{otherwise} \end{cases} \tag{3.44}$$

where ρ is the correlation between X_1 and Y_1, and σ^2 is the maximum of the variances of X_1 and Y_1.

We conclude this section by pointing out that the single-letter characterization of $R(X,Y,D)$ remains open in general even when $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ is IID.

3.4 New Information Distance for Jointly Gaussian Sources

In this section, we define $R(X,Y,D)$ as a pseudo distance over the set of all jointly Gaussian sources when the distortion level is small. (For example high resolution in images and videos.) Also, we extend our distance to the case which D is less than or equal to a special term which is a function of the statistical properties of jointly Gaussian sources. We have the following results.

Theorem 3. *Let (X_1,Y_1), (Y_1,Z_1), and (X_1,Z_1) be jointly Gaussian, and $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$, $(Y,Z) = \{(Y_i,Z_i)\}_{i=1}^{\infty}$, and $(X,Z) = \{(X_i,Z_i)\}_{i=1}^{\infty}$ are IID. Fix $D \geq 0$ and very small. Then with quadratic distortion measure, $R(X,Y,D)$ is a pseudo distance over the set of all jointly Gaussian sources i.e., satisfying the three properties of distance:*

(1) *$R(X,Y,D) = 0$ if $X = Y$.*

(2) *$R(X,Y,D) = R(Y,X,D)$.*

(3) *$R(X,Z,D) \leq R(X,Y,D) + R(Y,Z,D)$.*

Proof. Identity of Indiscernibles and symmetry properties i.e., properties (1) and (2), respectively, follow immediately from the definition of $R(X,Y,D)$ along with Definitions 3.1 and 3.2. Without loss of generality, we can assume that $\mathbf{E}X_1 = \mathbf{E}Y_1 = \mathbf{E}Z_1 = 0$, and

$\sigma_X^2 \geq \sigma_Y^2 \geq \sigma_Z^2$, where σ_X^2, σ_Y^2 and σ_Z^2 are the variances of X_1, Y_1 and Z_1, respectively. To prove property (3), we have to show that

$$\begin{aligned} R(X,Z,D) &\leq R(X,Y,D) + R(Y,Z,D) \\ \frac{1}{2}\log\frac{(1-\rho_{XZ}^2)\sigma_X^2}{D} &\leq \frac{1}{2}\log\frac{(1-\rho_{XY}^2)\sigma_X^2}{D} + \frac{1}{2}\log\frac{(1-\rho_{YZ}^2)\sigma_Y^2}{D} \\ 0 &\leq \frac{1}{2}\log\frac{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)\sigma_Y^2}{D(1-\rho_{XZ}^2)} \end{aligned} \tag{3.45}$$

are valid. Provided that D is very small in comparison to the other terms of the inequality 3.45, we can say the value inside of the $\log(.)$ will be greater than 1, and then the triangle inequality will be satisfied. More precisely, if

$$D \leq \frac{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)\sigma_Y^2}{(1-\rho_{XZ}^2)} \triangleq D_1$$

inequality 3.45 will be valid. For the other permutation of $R(.,.,D)$, the process is as same as above such that we need the following conditions

$$\begin{aligned} D &\leq \frac{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)\sigma_Y^2}{(1-\rho_{XY}^2)} \triangleq D_2 \\ D &\leq \frac{(1-\rho_{XY}^2)(1-\rho_{XZ}^2)\sigma_X^2\sigma_X^2}{(1-\rho_{YZ}^2)\sigma_Y^2} \triangleq D_3 \end{aligned}$$

which are corresponding to

$$\begin{aligned} R(X,Y,D) &\leq R(X,Z,D) + R(Y,Z,D) \\ R(Y,Z,D) &\leq R(X,Y,D) + R(X,Z,D), \end{aligned}$$

Therefore, if $D \leq \min\{D_1, D_2, D_3\}$, we can say the triangle inequality condition is satisfied. This completes the proof of Theorem 3.

□

We can go further and find the information distance when D is less than or equal to a special expression in which is a function of the statistical properties of jointly Gaussian sources, such as variance. Then we have the following theorem.

Theorem 4. *Suppose that* (X_1, Y_1), (Y_1, Z_1), *and* (X_1, Z_1) *are jointly Gaussian, and* $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$, $(Y, Z) = \{(Y_i, Z_i)\}_{i=1}^{\infty}$ *and* $(X, Z) = \{(X_i, Z_i)\}_{i=1}^{\infty}$ *are IID. Fix* $0 \leq D \leq \min\{\max\{\sigma_X^2, \sigma_Y^2\}, \max\{\sigma_Y^2, \sigma_Z^2\}, \max\{\sigma_X^2, \sigma_Z^2\}\}$. *Let*

$$R^{**}(X, Y, D) = \begin{cases} R(X, Y, D) & \text{if } X = Y \\ R(X, Y, D) + I(X_1, Y_1) & \text{otherwise.} \end{cases} \tag{3.46}$$

Then $R^{**}(X, Y, D)$ *is a pseudo distance over the set of all jointly Gaussian sources with* $d(x, \hat{x}) = (x - \hat{x})^2$, *i.e., satisfying the following properties:*

(1) $R^{**}(X, Y, D) = 0$ *if* $X = Y$.

(2) $R^{**}(X, Y, D) = R^{**}(Y, X, D)$.

(3) $R^{**}(X, Z, D) \leq R^{**}(X, Y, D) + R^{**}(Y, Z, D)$.

Proof. Properties (1) and (2) follow from the definition of $R(X, Y, D)$, Definitions 3.1 and 3.2 along with the symmetry property of the mutual information. Without loss of generality, we assume that $\mathbf{E}X = \mathbf{E}Y = \mathbf{E}Z = 0$, and $\sigma_X^2 \geq \sigma_Y^2 \geq \sigma_Z^2$, where σ_X^2, σ_Y^2 and σ_Z^2 are the variances of X_1, Y_1 and Z_1, respectively. To prove property (3), Consider five different cases: (I) $X = Y = Z$, (II) $X = Y \neq Z$, (III) $X \neq Y = Z$, (IV) $X = Z \neq Y$ and (V) $X \neq Y \neq Z$. In cases (I), (II), and (III), the triangle inequality is satisfied with the equality. At the same time, non-negativity of R^{**} reveals that the triangle inequality occurs in Case (IV). Finally, in the last case, we have to show that

$$R^{**}(X, Z, D) \leq R^{**}(X, Y, D) + R^{**}(Y, Z, D)$$

$$R(X, Z, D) + I(X_1; Z_1) \leq R(X, Y, D) + I(X_1; Y_1) + R(Y, Z, D) + I(Y_1; Z_1)$$

are valid. By Corollary 4 and mutual information for jointly Gaussian sources [14, Ch.8], the above inequality can be simplified to

$$\begin{aligned} \frac{1}{2}\log\frac{(1-\rho_{XZ}^2)\sigma_X^2}{D} + \frac{1}{2}\log\frac{1}{(1-\rho_{XZ}^2)} \leq \frac{1}{2}\log\frac{(1-\rho_{XY}^2)\sigma_X^2}{D} + \frac{1}{2}\log\frac{1}{(1-\rho_{XY}^2)} \\ + \frac{1}{2}\log\frac{(1-\rho_{YZ}^2)\sigma_Y^2}{D} + \frac{1}{2}\log\frac{1}{(1-\rho_{YZ}^2)} \end{aligned}$$

Then we have

$$\frac{1}{2}\log\frac{\sigma_X^2}{D} \leq \frac{1}{2}\log\frac{\sigma_X^2}{D} + \frac{1}{2}\log\frac{\sigma_Y^2}{D} \Rightarrow \frac{1}{2}\log\frac{\sigma_Y^2}{D} \overset{(?)}{\geq} 0$$

The last expression is valid since

$$D \leq \min\{\max\{\sigma_X^2, \sigma_Y^2\}, \max\{\sigma_Y^2, \sigma_Z^2\}, \max\{\sigma_X^2, \sigma_Z^2\}\} \leq \min\{\sigma_Y^2, \sigma_X^2\} = \sigma_Y^2. \tag{3.47}$$

For the other permutation of $R^{**}(.,.,D)$, we have two following cases:

(1) $R^{**}(X,Y,D) \leq R^{**}(X,Z,D) + R^{**}(Y,Z,D)$

(2) $R^{**}(Y,Z,D) \leq R^{**}(X,Y,D) + R^{**}(X,Z,D)$

Again, the first inequality follows from a similar argument as the above method. For the second one, the process is explained as follows

$$R^{**}(Y,Z,D) \leq R^{**}(X,Y,D) + R^{**}(X,Z,D)$$

$$R(Y,Z,D) + I(Y_1;Z_1) \leq R(X,Y,D) + I(X_1;Y_1) + R(X,Z,D) + I(X_1;Z_1)$$

For jointly Gaussian sources, we have

$$\frac{1}{2}\log\frac{(1-\rho_{YZ}^2)\sigma_Y^2}{D} + \frac{1}{2}\log\frac{1}{(1-\rho_{YZ}^2)} \leq \frac{1}{2}\log\frac{(1-\rho_{XY}^2)\sigma_X^2}{D} + \frac{1}{2}\log\frac{1}{(1-\rho_{XY}^2)} + \frac{1}{2}\log\frac{(1-\rho_{XZ}^2)\sigma_Y^2}{D} + \frac{1}{2}\log\frac{1}{(1-\rho_{XZ}^2)}$$

In the last step,

$$\frac{1}{2}\log\frac{\sigma_Y^2}{D} \leq \frac{1}{2}\log\frac{\sigma_X^2}{D} + \frac{1}{2}\log\frac{\sigma_X^2}{D} \Rightarrow \frac{1}{2}\log\frac{\sigma_X^2}{D} + \frac{1}{2}\log\frac{\sigma_X^2}{\sigma_Y^2} \overset{(?)}{\geq} 0 \tag{3.48}$$

The first term is non-negative because of 3.47. By our assumption ($\sigma_X^2 \geq \sigma_Y^2$), the second term will be greater than or equal to zero. Thus, the triangle inequality is satisfied in this case. Furthermore, if we change the assumption ($\sigma_X^2 \geq \sigma_Y^2 \geq \sigma_Z^2$), the results will be remained. This completes the proof of Theorem 4. □

3.5 New Information Distance for IID Sources with Small Distortion

In this section, we define the pseudo distance for any real-valued and IID source such that the distortion level D is small. More precisely, we analyze the distance property of $R(X,Y,D)$ when prescribed D is less than or equal to a special term which is a function of conditional entropy and statistical properties of given sources, such as correlation coefficients and variances.

Theorem 5. *Suppose that* $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$, $(Y,Z) = \{(Y_i,Z_i)\}_{i=1}^{\infty}$ *and* $(X,Z) = \{(X_i,Z_i)\}_{i=1}^{\infty}$ *are real-valued and IID pairs with each of* X, Y, *and* Z *satisfying 3.24 with* $d(x,\hat{x}) = (x-\hat{x})^2$. *Then* $R(X,Y,D)$ *will be a pseudo distance over the set of all IID sources when* D *is small or the distortion level is less than or equal to a special term which is a function of conditional entropy and statistical properties of given sources.* $R(X,Y,D)$ *should satisfy the following three properties:*

(1) $R(X,Y,D) = 0$ *if* $X = Y$.

(2) $R(X,Y,D) = R(Y,X,D)$.

(3) $R(X,Z,D) \leq R(X,Y,D) + R(Y,Z,D)$.

Proof. The properties of symmetry and identity of indiscernibles are obvious. The only property which needs to analyze precisely is the triangle inequality. Since we do not have a closed-form expression for $R(X,Y,D)$ in general, we need to use the appropriate upper and lower bounds to check the triangle inequality. The cut-set lower bound in 3.25 and conditional Shannon lower bound [28] are two useful lower bounds for $R(X,Y,D)$ specially when D has a small value. To proof the triangle inequality, we use the following lower bounds for $R(X,Y,D)$ and $R(Y,Z,D)$:

$$R(X,Y,D) \overset{(a)}{\geq} \max\{R_{X|Y}(D), R_{Y|X}(D)\} \overset{(b)}{\geq} \max\{h(X_1|Y_1), h(Y_1|X_1)\} - \frac{1}{2}\log 2\pi eD \quad (3.49)$$

$$R(Y,Z,D) \overset{(a)}{\geq} \max\{R_{Y|Z}(D), R_{Z|Y}(D)\} \overset{(b)}{\geq} \max\{h(Y_1|Z_1), h(Z_1|Y_1)\} - \frac{1}{2}\log 2\pi eD \quad (3.50)$$

where (a) follows from the cut-set lower bound for $R(X,Y,D)$ and $R(Y,Z,D)$ (Theorem 2) and (b) comes from the conditional Shannon lower bound [28] [5]. Inasmuch as the distortion function is quadratic, Corollary 5 gives us a suitable upper bound for $R(X,Z,D)$. If we show that the sum of two conditional lower bounds for $R(X,Y,D)$ and $R(Y,Z,D)$ is greater than or equal to the upper bound for $R(X,Z,D)$, then we can conclude that the triangle inequality will be satisfied. Without loss of generality, we can assume that $\sigma_X^2 \geq \sigma_Y^2 \geq \sigma_Z^2$. Now we have to show that

$$\begin{aligned} \max\{h(X_1|Y_1), h(Y_1|X_1)\} - \frac{1}{2}\log 2\pi eD + \max\{h(Y_1|Z_1), h(Z_1|Y_1)\} - \frac{1}{2}\log 2\pi eD \\ \overset{(?)}{\geq} \frac{1}{2}\log\frac{(1-\rho_{XZ}^2)\sigma_X^2}{D} \end{aligned} \quad (3.51)$$

[5] Shannon lower bounds are usually used to prove small distortion results; for example, see [24, 29, 30, 31]

is valid. The inequality 3.51 can be written in the following form:

$$\max\{h(X_1|Y_1), h(Y_1|X_1)\} + \max\{h(Y_1|Z_1), h(Z_1|Y_1)\} - \frac{1}{2}\log(2\pi e)^2 D(1-\rho_{XZ}^2)\sigma_X^2 \overset{(?)}{\geq} 0 \tag{3.52}$$

Since D is small, the first and second terms in 3.52 are negligible in comparison to the term which has D. More precisely, if

$$D \leq \frac{1}{(2\pi e)^2(1-\rho_{XZ}^2)\sigma_X^2} 2^{2(\max\{h(X_1|Y_1),h(Y_1|X_1)\}+\max\{h(Y_1|Z_1),h(Z_1|Y_1)\})} \triangleq D_1 \tag{3.53}$$

inequality 3.52 will be valid. It is so important to analyze other permutation of $R(.,.,D)$. Same as the above procedure, we need these two conditions for satisfying the triangle inequality

$$D \leq \frac{1}{(2\pi e)^2(1-\rho_{XY}^2)\sigma_X^2} 2^{2(\max\{h(X_1|Z_1),h(Z_1|X_1)\}+\max\{h(Y_1|Z_1),h(Z_1|Y_1)\})} \triangleq D_2 \tag{3.54}$$

$$D \leq \frac{1}{(2\pi e)^2(1-\rho_{YZ}^2)\sigma_Y^2} 2^{2(\max\{h(X_1|Y_1),h(Y_1|X_1)\}+\max\{h(X_1|Z_1),h(Z_1|X_1)\})} \triangleq D_3 \tag{3.55}$$

which are corresponding to

$$R(X,Y,D) \leq R(X,Z,D) + R(Y,Z,D)$$
$$R(Y,Z,D) \leq R(X,Y,D) + R(X,Z,D),$$

respectively. Hence, if $D \leq \min\{D_1, D_2, D_3\}$, then we can say the triangle inequality condition is met. This completes the proof of Theorem 5. □

Remark 3. *Precise examination on Theorem 5 shows that we can also define $R^{**}(X,Y,D)$ as a pseudo distance over the set of all real-valued and IID sources.*

3.6 Summary

In this chapter, we proposed a new information distance between two data objects X and Y as a smallest number of coded bits would convert X into $\hat{Y}$, and Y into $\hat{X}$ such that the distortion between X and $\hat{X}$ and the distortion between Y and $\hat{Y}$ are less than or equal to the distortion level D. This distance is totally defined when the alphabet is discrete. We then characterized and analyzed the information distance for some popular sources. Finally, the new information distance over the set of all real-valued and IID sources is discussed when the distortion level has a small value.

Chapter 4

Separately Precoded Broadcast Coding

4.1 Overview

As discovered previously, it can be found out that Given a class $\mathcal{C}$ of coding schemes within the coding paradigm, the information distance $R_{\mathcal{C}}(X, Y, D)$ between X and Y at the distortion level D is then defined as the smallest number of coded bits afforded by coding schemes from $\mathcal{C}$. When $\mathcal{C}$ is the class of so-called separately precoded broadcast codes, it is shown that for any DMS pair $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$, $R_{\mathcal{C}}(X, Y, D)$ is equal to the maximum of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information.

In Section 4.2, we formally formulate the separately precoded broadcast coding paradigm and define the information distance $R_{sb}(X, Y, D)$. Section 4.3 defines how to analyze $R_{\mathcal{C}}(X, Y, D)$ in terms of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information when $\mathcal{C}$ consists only of all so-called separately precoded broadcast codes within the coding paradigm; its distance property among different sources is also presented.

4.2 Formal Definitions: Codes and New Information Distances

Like the previous chapter, Let $\mathbf{A}$ and $\hat{\mathbf{A}}$ be two abstract alphabets. The sets $\mathbf{A}$ and $\hat{\mathbf{A}}$ will denote our source alphabet and reproduction alphabet, respectively.

Let us now impose some constraints on C_n in Definition 3.1 of Chapter 3. In particular, we split the encoding process into two steps. At Step 1, data objects x^n and y^n are separately precoded into nR_1 and nR_2 bits, respectively. At Step 2, the precoded bits are then jointly encoded into nR bits. The resulting type of code is called a separately precoded broadcast code.

For any $R > 0$ and n, let

$$\Omega(n, R) = \{1, 2, \cdots, \lfloor 2^{nR} \rfloor\}.$$

Formally, we have the following definition.

Definition 4.1. *A separately precoded broadcast code C_n of order n and rate R with precoded rates R_1 and R_2 consists of two separate precoding mappings*

$$f_1 : \mathbf{A}^n \to \Omega(n, R_1)$$

$$f_2 : \mathbf{A}^n \to \Omega(n, R_2)$$

a joint encoding mapping

$$f : \Omega(n, R_1) \times \Omega(n, R_2) \to \Omega(n, R)$$

and two decoding mappings

$$g_1 = (g_{11}, g_{12}) : \Omega(n, R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n \times \Omega(n, R_2)$$

and

$$g_2 = (g_{21}, g_{22}) : \Omega(n, R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n \times \Omega(n, R_1).$$

Figure 4.1 illustrates the encoding and decoding processes of a separately precoded broadcast code $C_n = (f_1, f_2, f, g_1, g_2)$ of order n and rate R with precoded rates R_1 and R_2. X^n and Y^n are first separately precoded into $f_1(X^n)$ of nR_1 bits and $f_2(Y^n)$ of nR_2 bits, and then jointly encoded into $f(f_1(X^n), f_2(Y^n))$ of nR bits. On the decoder side, the jointly encoded message $f(f_1(X^n), f_2(Y^n))$ converts: (1) X^n via Decoder 1 into an estimate $\hat{Y}^n = g_{11}(f(f_1(X^n), f_2(Y^n)), X^n)$ of Y^n and an estimate $\hat{f}_2(Y^n) = g_{12}(f(f_1(X^n), f_2(Y^n)), X^n)$ of $f_2(Y^n)$; and (2) Y^n via Decoder 2 into an estimate $\hat{X}^n = g_{21}(f(f_1(X^n), f_2(Y^n)), Y^n)$ of X^n and an estimate $\hat{f}_1(X^n) = g_{22}(f(f_1(X^n), f_2(Y^n)), Y^n)$ of $f_1(X^n)$.

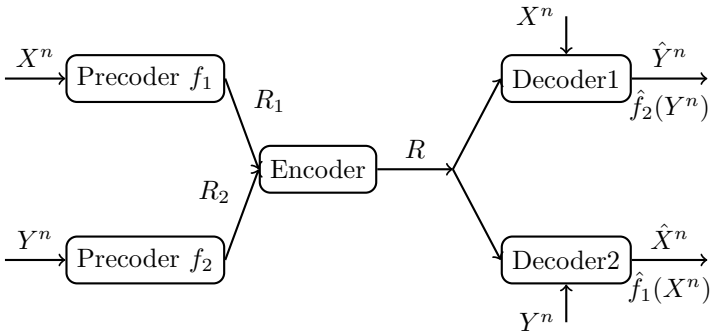

Figure 4.1: Illustration of a separately precoded broadcast code.

Definition 4.2. *Let $\mathcal{C}_{sb}$ consist of all separately precoded broadcast codes. Given stationary sources $X = \{X_i\}_{i=1}^{\infty}$ and $Y = \{Y_i\}_{i=1}^{\infty}$, a rate distortion pair (R, D) is said to be $\mathcal{C}_{sb}$-achievable for (X, Y) if for any $\epsilon > 0$, there exist a finite set $\mathbf{B} \subseteq \hat{\mathbf{A}}$ and, for all sufficiently large n, a separately precoded block code $C_n = (f_1, f_2, f, g_1, g_2)$ of order n and rate $R + \epsilon$ with precoded rates $R_1 + \epsilon$ and $R_2 + \epsilon$ such that*

$$\Pr\{d(X^n, \hat{X}^n) > D\} < \epsilon \tag{4.1}$$

$$\Pr\{d(Y^n, \hat{Y}^n) > D\} < \epsilon \tag{4.2}$$

$$\Pr\{f_1(X^n) \neq \hat{f}_1(X^n)\} < \epsilon \tag{4.3}$$

and

$$\Pr\{f_2(Y^n) \neq \hat{f}_2(Y^n)\} < \epsilon \tag{4.4}$$

where $(\hat{X}^n, \hat{f}_1(X^n)) = g_2(f(f_1(X^n), f_2(Y^n)), Y^n)$, $(\hat{Y}^n, \hat{f}_2(Y^n)) = g_1(f(f_1(X^n), f_2(Y^n)), X^n)$, and both $\hat{X}^n$ and $\hat{Y}^n$ take values in $\mathbf{B}^n$.

Let $\mathcal{R}_{sb}(X, Y)$ denote the set of all $\mathcal{C}_{sb}$-achievable (R, D) pairs for (X, Y). It can be verified that $\mathcal{R}_{sb}(X, Y)$ is closed. Given $D \geq 0$, define the information distance between X and Y at the distortion level D with respect to $\mathcal{C}_{sb}$ as

$$R_{sb}(X, Y, D) \stackrel{\Delta}{=} \min\{R : (R, D) \in \mathcal{R}_{sb}(X, Y)\}. \tag{4.5}$$

As in the case of $R(X, Y, D)$, we also aim to characterize $R_{sb}(X, Y, D)$, and analyze its relationship among different sources X, Y, Z, etc as a notion of distance.

4.3 $R_{sb}(X, Y, D)$: Distance Property and Characterization

In this section, we analyze the distance property of $R_{sb}(X, Y, D)$ over the set of stationary sources X, Y, Z, etc, and characterize it in terms of Wyner-Ziv coding rates for DMS pairs (X, Y). Again, we begin with its distance property.

Suppose that both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are finite, and the condition 3.1 is met. In parallel with Theorem 1 in Chapter 3, we have the following result.

Theorem 1. *Fix $D \geq 0$. Let*

$$R^*_{sb}(X, Y, D) = \begin{cases} R_{sb}(X, Y, D) & \text{if } X = Y \\ R_{sb}(X, Y, D) + H(D) & \text{otherwise.} \end{cases} \tag{4.6}$$

Then $R^*_{sb}(X, Y, D)$ is a pseudo distance over the set of all stationary sources.

Remark 1. *In view of the definitions of $R(X, Y, D)$ and $R_{sb}(X, Y, D)$, it follows that*

$$R(X, Y, D) \leq R_{sb}(X, Y, D). \tag{4.7}$$

However, the above inequality, together with Theorem 1 in Chapter 3, does not imply Theorem 1 directly.

An approach similar to the proof of Theorem 1 of the previous chapter can be used to show Theorem 1. To prove the corresponding triangle inequality, we first show that allowing random joint encoding mappings f in the definition of separately precoded broadcast codes does not decrease the rate distortion function of all $\mathcal{C}_{sb}$-achievable pairs (R, D) for any (X, Y). Given any X, Y, and Z, we then show that $(R_{sb}(X, Y, D)+R_{sb}(Y, Z, D)+H(D), D)$ is achievable by separately precoded broadcast codes with random joint encoding mappings f for (X, Z).

For any DMS pair $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$, let $R^{WZ}_{X|Y}(D)$ ($R^{WZ}_{Y|X}(D)$, resp.) denote the Wyner-Ziv coding rate of X (Y, resp.) with Y (X, resp.) as side information available at the decoder. Then we have the following result.

Theorem 2. *For any DMS pair $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$, we have*

$$R_{sb}(X, Y, D) = \max\{R^{WZ}_{X|Y}(D), R^{WZ}_{Y|X}(D)\}.$$

Proof. To prove this theorem, we need to verify both Achievability and Converse proof **Achievability part:** We want to show that if $R > max\{R_{X|Y}^{WZ}(D), R_{Y|X}^{WZ}(D)\}$, then there exist a separately precoded block code $C_n = (f_1, f_2, f, g_1, g_2)$ of order n such that

$$\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)] \leq D$$

$$\limsup_{n\to\infty} \mathbf{E}[d(Y^n, \hat{Y}^n)] \leq D.$$

We use joint typicality encoding to describe X and Y by U and $\hat{U}$, respectively. Since U ($\hat{U}$, resp.) has a correlation with Y (X, resp.); however, binning can be used to reduce their description rate. The module-2 sum of the bin indexes of U and $\hat{U}$ is sent to the decoders. Then, Decoder 1 recovers the index bin of $\hat{U}$ and uses joint typicality decoding with X to recover $\hat{U}$ and then reconstruct $\hat{Y}$ and $f_2(\hat{Y}^n)$ from $\hat{U}$ and X. At the same time, Decoder 2 reconstructs $\hat{X}$ and $f_1(\hat{X}^n)$ by using the joint typicality with Y. We now provide the details.

Codebook generation: We fix the conditional PMFs $p(u|x)$, $p(\hat{u}|y)$ as well as functions $\hat{x}(u, y)$ and $\hat{y}(\hat{u}, x)$ such that $\mathbf{E}[d(X, \hat{X})] \leq \frac{D}{1+\epsilon}$ and $\mathbf{E}[d(Y, \hat{Y}] \leq \frac{D}{1+\epsilon}$, respectively, where D is the distortion level. Randomly and independently generate 2^{nR_1} sequences $u^n(l)$, $l \in [1 : 2^{nR_1}]$, each according to $\prod_{i=1}^n P_U(u_i)$. Similarly, randomly and independently generate 2^{nR_2} sequences $\hat{u}^n(s)$, $s \in [1 : 2^{nR_2}]$, each according to $\prod_{i=1}^n P_{\hat{U}}(\hat{u}_i)$. Now, partition both set of indices l and s into equal-size subsets referred to as bins $\mathcal{B}(m')$ and $\mathcal{B}(m'')$, as follows

$$\mathcal{B}(m') = [(m'-1)2^{n(R_1-R)} + 1 : m'2^{n(R_1-R)}], \tag{4.8}$$

$$\mathcal{B}(m'') = [(m''-1)2^{n(R_2-R)} + 1 : m''2^{n(R_2-R)}]. \tag{4.9}$$

respectively, where both $m', m'' \in [1 : 2^{nR}]$. The codebook generation is revealed to Encoder and both Decoders 1 and 2.

Precoder f_1 : Upon receiving x^n, Precoder f_1 finds an index l such that $(x^n, u^n(l)) \in \mathcal{T}_{\epsilon'}^{(n)}(X, U)$. If there is more than one such index, Precoder f_1 uses the smallest one. If there is no such index, it selects randomly an index from $[1 : 2^{nR_1}]$ uniformly and then send it to Encoder.

Precoder f_2 : Given y^n, Precoder f_2 finds an index s such that $(y^n, \hat{u}^n(s)) \in \mathcal{T}_{\epsilon'}^{(n)}(Y, \hat{U})$. If there is more than one such index, Precoder f_2 uses the smallest one. If there is no such index, it sets randomly an index from $[1 : 2^{nR_2}]$ uniformly and then send it to Encoder.

Encoder: Given indexes l and s, Encoder finds m' and m'' such that $l \in \mathcal{B}(m')$ and $s \in \mathcal{B}(m'')$. Then it expresses each message (m' and m'') as a binary sequence (nR bits)

and broadcasts the modulo-2 sum of the two sequences ($m = m' \oplus m''$) to both Decoders 1 and 2.

Decoder 1 : Let $\epsilon > \epsilon'$. Upon receiving m, Decoder 1 recovers the massage of the other source (m'') by performing modulo-2 sum on the binary expression of its message (m') and the received sequence (m). Then it finds the unique index $\hat{s} \in \mathcal{B}(m'')$ such that $(x^n, \hat{u}^n(\hat{s})) \in \mathcal{T}_\epsilon^{(n)}(X, \hat{U})$; otherwise it sets $\hat{s} = 1$. Finally, it computes the reconstruction sequence as $\hat{y}_i = \hat{y}(\hat{u}_i(\hat{s}), x_i)$ for $i \in [1 : n]$. Decoder 1 recovers $\hat{y}^n$ and $\hat{s}$ which the binary expression of $\hat{s}$ is equal to $f_2(y^n)$ with the high probability.

Decoder 2 : Let $\epsilon > \epsilon'$. Upon receiving m, Decoder 2 recovers the massage of the other source (m') by performing modulo-2 sum on the binary expression of its message (m'') and the received sequence (m). Then the decoder 2 finds the unique index $\hat{l} \in \mathcal{B}(m')$ such that $(y^n, u^n(\hat{l})) \in \mathcal{T}_\epsilon^{(n)}(Y, U)$; otherwise it sets $\hat{l} = 1$. Finally, it computes the reconstruction sequence as $\hat{x}_i = \hat{x}(u_i(\hat{l}), y_i)$ for $i \in [1 : n]$. Decoder 2 recovers $\hat{x}^n$ and $\hat{l}$ which the binary expression of $\hat{l}$ is equal to $f_1(x^n)$ with the high probability.

Analysis of expected distortion: Let (L, S) denote the chosen indices at Precoders f_1, f_2, respectively, such that $L \in \mathcal{B}(M')$ and $S \in \mathcal{B}(M'')$. At Encoder, M shows the modulo-2 sum of M' and M''. Furthermore, $(\hat{L}, \hat{S})$ be the index estimate at Decoders 2 and 1, respectively. We define the " Error " event as

$$\mathcal{E} = \mathcal{E}' \cup \mathcal{E}'' = \{(U^n(\hat{L}), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(U, X, Y)\} \cup \{(\hat{U}^n(\hat{S}), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(\hat{U}, X, Y)\} \tag{4.10}$$

Now, consider the events

$$\begin{aligned}
\mathcal{E}_1 &= \{(U^n(l), X^n) \notin \mathcal{T}_{\epsilon'}^{(n)}(U, X) \text{ for all } l \in [1 : 2^{nR_1})\},\\
\mathcal{E}_2 &= \{(U^n(L), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(U, X, Y)\},\\
\mathcal{E}_3 &= \{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U, Y) \text{ for some } \tilde{l} \in \mathcal{B}(M'), \tilde{l} \neq L\},\\
\mathcal{E}_4 &= \{(\hat{U}^n(s), Y^n) \notin \mathcal{T}_{\epsilon'}^{(n)}(\hat{U}, Y) \text{ for all } s \in [1 : 2^{nR_2})\},\\
\mathcal{E}_5 &= \{(\hat{U}^n(S), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(\hat{U}, X, Y)\},\\
\mathcal{E}_6 &= \{(\hat{U}^n(\tilde{s}), X^n) \in \mathcal{T}_\epsilon^{(n)}(\hat{U}, X) \text{ for some } \tilde{s} \in \mathcal{B}(M''), \tilde{s} \neq S\}.
\end{aligned}$$

Since the "Error" event occurs only if $(U^n(L), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(U, X, Y)$ or $\hat{L} \neq L$, as well as $(\hat{U}^n(S), X^n, Y^n) \notin \mathcal{T}_\epsilon^{(n)}(\hat{U}, X, Y)$ or $\hat{S} \neq S$, by the union of events bound,

$$P(\mathcal{E}) \leq P(\mathcal{E}') + P(\mathcal{E}'') \leq P(\mathcal{E}_1) + P(\mathcal{E}_1^c \cap \mathcal{E}_2) + P(\mathcal{E}_3) + P(\mathcal{E}_4) + P(\mathcal{E}_4^c \cap \mathcal{E}_5) + P(\mathcal{E}_6).$$

We now bound each term. By the covering lemma (see Appendix A.1), $P(\mathcal{E}_1)$ and $P(\mathcal{E}_4)$ tends to zero as $n \to \infty$ if

$$R_1 > I(X;U) + \delta(\epsilon') \tag{4.11}$$
$$R_2 > I(Y;\hat{U}) + \delta(\epsilon') \tag{4.12}$$

Since $\epsilon > \epsilon'$, $\mathcal{E}_1^c = \{(U^n(L), X^n) \in \mathcal{T}_{\epsilon'}^{(n)}(U,X)\}$, and $Y^n \mid \{U^n(L) = u^n, X^n = x^n\} \sim \prod_{i=1}^n P_{Y|U,X}(y_i|u_i,x_i) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)$, by the conditional typicality lemma (Lemma 2), $P(\mathcal{E}_1^c \cap \mathcal{E}_2) \to 0$ as $n \to \infty$. Similarly, as $\epsilon > \epsilon'$, $\mathcal{E}_4^c = \{(\hat{U}^n(S), Y^n) \notin \mathcal{T}_{\epsilon'}^{(n)}(\hat{U},Y)\}$, and $X^n \mid \{\hat{U}^n(S) = \hat{u}^n, Y^n = y^n\} \sim \prod_{i=1}^n P_{X|\hat{U},Y}(x_i|\hat{u}_i,y_i) = \prod_{i=1}^n P_{X|Y}(x_i|y_i)$, by the conditional typicality lemma (Lemma 2), $P(\mathcal{E}_4^c \cap \mathcal{E}_5)$ tends to zero as $n \to \infty$.

To bound $P(\mathcal{E}_3)$ and $P(\mathcal{E}_6)$, we first use lemma 11.1 of [1] (see Appendix A.3) to find an upper bound as follows

$$P(\mathcal{E}_3) \le \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1)\}$$
$$P(\mathcal{E}_6) \le \Pr\{(\hat{U}^n(\tilde{s}), X^n) \in \mathcal{T}_\epsilon^{(n)}(\hat{U},X) \text{ for some } \tilde{s} \in \mathcal{B}(1)\}$$

For each $\tilde{l} \in \mathcal{B}(1)$, the sequence $U^n(\tilde{l}) \sim \prod_{i=1}^n P_U(u_i)$ is independent of Y^n. Now, by the packing lemma (see Appendix A.2), $\Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1)\}$ tends to zero as $n \to \infty$ if $R_1 - R < I(Y;U) - \delta(\epsilon)$. Thus, by lemma 11.1 of [1] and equation 4.11, $P(\mathcal{E}_3)$ tends to zero if

$$R > R_1 - I(Y;U) + \delta(\epsilon) > I(X;U) - I(Y;U) + \delta(\epsilon) + \delta(\epsilon') = I(X;U|Y) + \delta'(\epsilon) \tag{4.13}$$

In a similar way, for each $\tilde{s} \in \mathcal{B}(1)$, the sequence $\hat{U}^n(\tilde{s}) \sim \prod_{i=1}^n P_{\hat{U}}(\hat{u}_i)$ is independent of X^n. By the packing lemma, $\Pr\{(\hat{U}^n(\tilde{s}), X^n) \in \mathcal{T}_\epsilon^{(n)}(\hat{U},X) \text{ for some } \tilde{s} \in \mathcal{B}(1)\}$tends to zero as $n \to \infty$ if $R_2 - R < I(X;\hat{U}) - \delta(\epsilon)$. Again, by lemma 11.1 of [1] and equation 4.12, $P(\mathcal{E}_6)$ tends to zero if

$$R > R_2 - I(X;\hat{U}) + \delta(\epsilon) > I(Y;\hat{U}) - I(X;\hat{U}) + \delta(\epsilon) + \delta(\epsilon') = I(Y;\hat{U}|X) + \delta'(\epsilon) \tag{4.14}$$

Since in the codebook generation part, we fixed the the conditional PMFs $p(u|x)$ and $p(\hat{u}|y)$ as well as functions $\hat{x}(u,y)$ and $\hat{y}(\hat{u},x)$ such that $\mathbf{E}[d(X,\hat{X})] \le \frac{D}{1+\epsilon}$ and $\mathbf{E}[d(Y,\hat{Y})] \le \frac{D}{1+\epsilon}$, we can reformulate equations 4.13 and 4.14 as follows

$$R > \min_{p(u|x),\hat{x}(u,y):\, \mathbf{E}[d(X,\hat{X})] \le \frac{D}{1+\epsilon}} I(X;U|Y) + \delta'(\epsilon) = R_{X|Y}^{WZ}\left(\frac{D}{1+\epsilon}\right) + \delta'(\epsilon) \tag{4.15}$$

$$R > \min_{p(\hat{u}|y),\hat{y}(\hat{u},x):\, \mathbf{E}[d(Y,\hat{Y})]\le \frac{D}{1+\epsilon}} I(Y;\hat{U}|X) + \delta'(\epsilon) = R_{Y|X}^{WZ}(\frac{D}{1+\epsilon}) + \delta'(\epsilon) \tag{4.16}$$

,respectively. Now, provided that $R > \max\{R_{X|Y}^{WZ}(\frac{D}{1+\epsilon}), R_{Y|X}^{WZ}(\frac{D}{1+\epsilon})\} + \delta'(\epsilon)$, then there will be no "Error" i.e., $P(\mathcal{E}) \to 0$.

When there is no "Error", $(U^n(L), X^n, Y^n) \in \mathcal{T}_\epsilon^n(U,X,Y)$ and $(\hat{U}^n(S), X^n, Y^n) \in \mathcal{T}_\epsilon^n(\hat{U},X,Y)$. Thus, by the law of total expectation and the typical average lemma (Lemma 1), the asymptotic distortion averaged over the random codebook and encoding is upper bounded as

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbf{E}[d(X^n,\hat{X}^n)] \le \limsup_{n\to\infty} (P(\mathcal{E})\mathbf{E}[d(X^n,\hat{X}^n)|\mathcal{E}] + P(\mathcal{E}^{\mathrm{c}})\mathbf{E}[d(X^n,\hat{X}^n)|\mathcal{E}^{\mathrm{c}}]) \\
&\le \limsup_{n\to\infty} (d_{max}P(\mathcal{E}) + P(\mathcal{E}^{\mathrm{c}})(1+\epsilon)\mathbf{E}[d(X,\hat{X})] \le (1+\epsilon)\frac{D}{(1+\epsilon)} = D
\end{aligned}$$

where $d_{max} = \max_{(x,\hat{x})\in \mathbf{A}\times\mathbf{B}} d(x,\hat{x})$. With the same procedure, $\limsup_{n\to\infty} \mathbf{E}[d(Y^n,\hat{Y}^n)] \le D$ as $n \to \infty$ if $R > \max\{R_{X|Y}^{WZ}(\frac{D}{(1+\epsilon)}), R_{Y|X}^{WZ}(\frac{D}{(1+\epsilon)})\} + \delta'(\epsilon)$.

In the last step, from the continuity of $R_{X|Y}^{WZ}(D)$ and $R_{Y|X}^{WZ}(D)$ in D [20], taking $\epsilon \to 0$ shows that any rate distortion pair (R,D) with $R > \max\{R_{X|Y}^{WZ}(D), R_{Y|X}^{WZ}(D)\}$ is achievable.

The converse part: We need to show that for any block code $C_n = (f_1, f_2, f, g_1, g_2)$ of order n with

$$\limsup_{n\to\infty} \mathbf{E}[d(X^n,\hat{X}^n)] \le D \tag{4.17}$$

$$\limsup_{n\to\infty} \mathbf{E}[d(Y^n,\hat{Y}^n)] \le D, \tag{4.18}$$

we must have $R \ge \max\{R_{X|Y}^{WZ}(D), R_{Y|X}^{WZ}(D)\}$.

Let M denotes the chosen indices at Encoder which sends to both Decoders 1 and 2. The key to the proof is identify U_i and $\hat{U}_i$. In general, $\hat{X}_i$ is a function of (M, Y^n). We would like $\hat{X}_i$ to be a function of (U_i, Y_i), so we identify the auxiliary random variable $U_i = (M, Y^{i-1}, Y_{i+1}^n)$. Similar to the proof of the converse for Wyner-Ziv coding in [1],

Consider

$$\begin{aligned}
nR &\geq H(M) \\
&\geq H(M|Y^n) \\
&= I(X^n; M|Y^n) \\
&= \sum_{i=1}^{n} I(X_i; M|Y^n, X^{i-1}) \\
&= \sum_{i=1}^{n} H(X_i|Y^n, X^{i-1}) - H(X_i|M, Y^n, X^{i-1}) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|M, Y^{i-1}, Y_i, Y_{i+1}^n) \\
&= \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|U_i, Y_i) \\
&= \sum_{i=1}^{n} I(X_i; U_i|Y_i) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} R_{X|Y}^{WZ}(\mathbf{E}[d(X_i, \hat{X}_i)]) \\
&\overset{(c)}{\geq} nR_{X|Y}^{WZ}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[d(X_i, \hat{X}_i)]) \\
&= nR_{X|Y}^{WZ}(\mathbf{E}[d(X^n, \hat{X}^n)])
\end{aligned}$$

where (a) follows since (X_i, Y_i) is independent of $(Y^{i-1}, Y_{i+1}^n, X^{i-1})$, (b) by the definition of $R_{X|Y}^{WZ}(D) = \min I(X; U|Y)$, and (c) follows by the convexity of $R_{X|Y}^{WZ}(D)$. Since $R_{X|Y}^{WZ}(D)$ is continuous and non-increasing in D, it follows from the bound on distortion in 4.17 that

$$\begin{aligned}
R &\geq \limsup_{n\to\infty} R_{X|Y}^{WZ}(\mathbf{E}[d(X^n, \hat{X}^n)]) \\
&\geq R_{X|Y}^{WZ}(\limsup_{n\to\infty} \mathbf{E}[d(X^n, \hat{X}^n)]) \\
&\geq R_{X|Y}^{WZ}(D)
\end{aligned} \tag{4.19}$$

Similarly, $\hat{Y}_i$ is a function of (M, X^n), so we set $\hat{U}_i = (M, X^{i-1}, X_{i+1}^n)$. Consider

$$\begin{aligned}
nR &\geq H(M) \\
&\geq H(M|X^n) \\
&= I(Y^n; M|X^n) \\
&= \sum_{i=1}^{n} I(Y_i; M|X^n, Y^{i-1}) \\
&\overset{(a)}{=} \sum_{i=1}^{n} I(Y_i; M, X^{i-1}, X_{i+1}^n, Y^{i-1}|X_i) \\
&\geq \sum_{i=1}^{n} I(Y_i; \hat{U}_i|X_i) \\
&\geq \sum_{i=1}^{n} R_{Y|X}^{WZ}(\mathbf{E}[d(Y_i, \hat{Y}_i)]) \\
&\overset{(b)}{\geq} nR_{Y|X}^{WZ}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[d(Y_i, \hat{Y}_i)])
\end{aligned}$$

where (a) follows since (X_i, Y_i) is independent of $(X^{i-1}, X_{i+1}^n, Y^{i-1})$ and (b) follows by the convexity of $R_{Y|X}^{WZ}(D)$. Since $R_{Y|X}^{WZ}(D)$ is non-increasing, by assumption 4.18, we can write

$$\begin{aligned}
R &\geq \limsup_{n\to\infty} R_{Y|X}^{WZ}(\mathbf{E}[d(Y^n, \hat{Y}^n)]) \\
&\geq R_{Y|X}^{WZ}(\limsup_{n\to\infty} \mathbf{E}[d(Y^n, \hat{Y}^n)]) \\
&\geq R_{Y|X}^{WZ}(D)
\end{aligned} \tag{4.20}$$

By combining 4.19 and 4.20, we derive

$$R \geq \max\{R_{X|Y}^{WZ}(D), R_{Y|X}^{WZ}(D)\}.$$

This completes the proof of Theorem 2. □

Remark 2. *We can combine two Wyner-Ziv coding with a simple linear-network code. Encoder maps X^n to a binary sequence by a Wyner-Ziv Code [20]. This code behaves Y^n as side information at decoder 2, but it ignores Y^n at Encoder. At the same time, Y^n is mapped to a binary sequence by a Wyner-Ziv code which handle X^n as side information*

at Decoder 1*, but it ignores* X^n *at Encoder. Therefore, we can easily achieve the following rate by the separately precoded broadcast code:*

$$R \geq \max\{R_{X|Y}^{WZ}(D), R_{Y|X}^{WZ}(D)\}.$$

It is instructive to compare $R_{sb}(X, Y, D)$ with $R(X, Y, D)$. When X_1 and Y_1 are jointly Gaussian, and $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$ is IID with $d(x, \hat{x}) = (x - \hat{x})^2$, we have

$$R_{sb}(X, Y, D) = R(X, Y, D).$$

In general, however, it is expected that the inequality in 4.7 is strict, which is the case, for example, when (X, Y) is the source pair in the example of DSBS.

Since the separately precoded broadcast coding is a especial case of a coding diagram shown in Figure 3.1, Theorems 3, 4, and 5 of Chapter 3 are holding.

4.4 Summary

In this chapter, $R_{\mathcal{C}}(X, Y, D)$ is characterized in terms of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information when (X, Y) is a DMS pair and $\mathcal{C}$ consists only of so-called separately precoded broadcast codes within the coding paradigm; its distance property among different sources is also presented.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

In this thesis, we have proposed a new information distance between any two data objects X and Y which has the universality from the information distance as defined in [16], and the computability and applicability to both discrete and continuous-valued data as in SMID $d_\phi(X, Y)$. In other words, we bring distortion into the information distance, and universality to the SMID. For this reason, we give a new coding paradigm such that X and Y are compressed into a sequence of coded bits specifying a codeword which would, in turn, convert Y into $\hat{X}$, and X into $\hat{Y}$ with the condition that both the distortion between X and $\hat{X}$ and the distortion between Y and $\hat{Y}$ are less than or equal to a prescribed threshold D. Since we need universality to some extent, we analyze a class $\mathcal{C}$ of coding schemes within the coding paradigm. Given a class $\mathcal{C}$, the information distance between X and Y at the distortion level D is defined as the smallest number of coded bits caused by coding schemes from $\mathcal{C}$. We characterize and analyze the information distance for some classes $\mathcal{C}$. For example, when $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$ is an IID pair, we establish upper and lower bounds to our new information distance. For the finite alphabets, we defined a pseudo distance and analyzed the distance properties. When the alphabets are not finite but the prescribed threshold D is small, our new information distance is also valid and it satisfies the triangle inequality. In the last part of this thesis, when $\mathcal{C}$ is the class of separately precoded broadcast codes, the information distance is equal to the maximum of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information.

5.2 Future Works

In this thesis, where $\mathcal{C}$ consists of all codes within the coding paradigm, upper and lower bounds to our new information distance are established and are shown to be tight in some special cases such as jointly Gaussian and DSBS. Single letter characterization of $R(X, Y, D)$ remains open in general. One of our future work is finding the closed-form solution for this multi-user information theory problem when the sources are stationary, ergodic.

For the finite alphabets, the new information distance is completely done. When the alphabet sources are not finite, our pseudo distance is not valid in general. We studied the case which is IID source and the distortion level D is small (For example high-resolution images or videos). Our plan for future work concentrates on the case in which the alphabet is abstract and we do not have any constraint on distortion level D. One of the most interesting future work plan is defining the information distance for stationary, totally ergodic sources which we are working on it now.

Finally, we can analyse the applicability of the new distance measure to image classification and compare it to the current database classifications as an interesting future work.

References

[1] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, Jan. 2012.

[2] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. Vol. 40, No. 2, Article 5, pp.1-60, April 2008.

[3] E.-H. Yang, J. Meng, and X. Yu, "Display, visualization, and management of photos based on content analytics," *US Patent Application*, vol. No. 14/790,650, July 2, 2015.

[4] F. Emmert-Streib and M. D. Eds., *Information Theory and Statistical Learning*. Springer Science+Business Media, LLC, 2009.

[5] E.-H. Yang, X. Yu, and J. Meng, "Set mapping induced image perceptual similarity distance," *Proc. of the 2015 Information Theory and Applications Workshop*, vol. San Diego, California, U.S.A., Feb. 1-Feb. 6, 2015.

[6] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Netw.*, vol. Vol. 10, No. 5, pp.1055-1064, Sept. 1999.

[7] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," *Proc. of the 2003 Int. Conf. on Image process.*, vol. Vol. 3, pp. 513-516, Sept. 2003.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, 2, pp. 91-110, 2004.

[9] T. Lindeberg, "Scale invariant feature transform," *Scholarpedia*, vol. 7 (5): 10491, 2012.

[10] A. Vedaldi, *An implementation of SIFT detector and descriptor*, 2008. http://www.robots.ox.ac.uk/~vedaldi/code.html.

[11] L. Kang, C. Hsu, H. Chen, C. L. C. Lu, and S. Pei, "Feature-based sparse representation for image similarity assessment," *IEEE Trans. Multimedia*, vol. Vol. 13, No. 5, pp. 1019-1030, October 2011.

[12] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. Vol. 29, No. 5, pp. 840-853, May 2007.

[13] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *Int. Journal of Computer Vision*, vol. 35(3), pp. 245-268, 1999.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory (second edition)*. Hoboken, NJ: Wiley, 2006.

[15] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Trans.*, vol. Vol. 1, pp. 1-7, 1965, 1965.

[16] C. H. Bennet and *et al*, "Information distance," *IEEE Trans. Inf. Theory*, vol. Vol. 1, pp. 1-7, 1965, Jul. 1998.

[17] R. W. Yeung, *Information Theory and Network Coding*. Springer Science+Business Media, LLC, 2008.

[18] A. Orlitsky and J. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. Vol. 47, no. 3, pp. 903–917, Mar. 2001.

[19] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981.

[20] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. Vol. 22, No. 1, pp. 1-10, Jan. 1976.

[21] A. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Trans. Inf. Theory*, vol. Vol. 40, No. 6, pp. 2031-2034, Nov. 1994.

[22] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inf. Theory*, vol. Vol. 31, No. 6, pp. 727-734, Nov. 1985.

[23] Z. Zhang, E. Yang, and V. Wei, "The redundancy of source coding with a fidelity criterion - part one: Known statistics," *IEEE Trans. Inf. Theory*, vol. vol. 43, no. 1, pp. 71–91, Jan. 1997.

[24] R. Timo, A. Grant, and G. Kramer, "Lossy broadcasting with complementary side information," *IEEE Trans. Inf. Theory*, vol. Vol. 59, no. 1, pp. 104-131, Jan. 2013.

[25] R. Gray, "Conditional rate-distortion theory," *Stanford Univ.*, vol. Stanford, CA, 1972.

[26] A. Kimura and T. Uyematsu, "Multiterminal source coding with complementary delivery," *Int. Symp. on Inf. Theory Appl. (ISITA). Seoul, Korea,*, vol. pp. 189–194, 2006.

[27] S. M. Ross, *A First Course in Probability*. Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[28] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. Vol. IT-19, pp. 480-489, July 1973.

[29] T. Linder and R. Zamir, "On the asymptotic tightness of the shannon lower bound," *IEEE Trans. Inf. Theory*, vol. Vol. 40, no. 6, pp. 2026–2031, Nov. 1994.

[30] R. Zamir and T. Berger, "Multiterminal source coding with high resolution," *IEEE Trans. Inf. Theory*, vol. Vol. 45, no. 1, pp. 106–117, Jan. 1999.

[31] R. Zamir, "Gaussian codes and shannon bounds for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. Vol. 45, no. 7, pp. 2629–2636, Nov. 1999.

[32] S. Wang, Y. Fang, and S. Cheng, *Distributed Source Coding: Theory and Practice*. Wiley, 2017.

APPENDICES

Appendix A

Useful Lemmas In Information Theory

A.1 Covering Lemma

The covering lemma generalizes the bound on the probability of the encoding error event ε in the achievability proof of the lossy source coding theorem. The lemma will be used in the achievability proofs of several multi-user source and channel coding theorems[1].

Lemma 1. ***(Covering Lemma)*** *For any $\epsilon > 0$, we can find n such that if $X^n(m)$, $m \in [1 : 2^{nR}]$, are independently drawn from $\prod_{i=1}^{n} P_X(x_i)$ and $Y^n \in \mathcal{T}_{\epsilon'}^{(n)}(Y)$, $\epsilon' < \epsilon$, is independent of each $X^n(m)$, then*

$$\Pr\{(X^n(m), Y^n) \in \mathcal{T}_{\epsilon}^{(n)}(X, Y) \text{ for some } m\} \to 1$$

if $R > I(X;Y) + \delta(\epsilon)$, where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ [32].

Proof.

$$
\begin{aligned}
&\Pr\{(X^n(m), Y^n) \notin \mathcal{T}_\epsilon^{(n)}(X,Y) \text{ for all } m\} \\
&= \sum_{y^n \in \mathcal{T}_{\epsilon'}^{(n)}(Y)} p(y^n) \Pr\{(X^n(m), y^n) \notin \mathcal{T}_\epsilon^{(n)}(X,Y) \text{ for all } m \mid y^n\} \\
&= \sum_{y^n \in \mathcal{T}_{\epsilon'}^{(n)}(Y)} p(y^n) \prod_{m=1}^{2^{nR}} \Pr\{(X^n(m), y^n) \notin \mathcal{T}_\epsilon^{(n)}(X,Y)\} \\
&= \sum_{y^n \in \mathcal{T}_{\epsilon'}^{(n)}(Y)} p(y^n) (\Pr\{(X^n(m), y^n) \notin \mathcal{T}_\epsilon^{(n)}(X,Y)\})^{2^{nR}} \\
&\le (1 - 2^{-n(I(X;Y)+\delta(\epsilon))})^{2^{nR}} \le e^{-2^{n(R-I(X;Y)-\delta(\epsilon))}}
\end{aligned}
$$

which tends to zero as $n \to 0$ if $R > I(X;Y) + \delta(\epsilon)$. □

We can generalize the covering lemma by replacing the independence to the conditional independence condition. Then we have the following lemma.

Lemma 2. ***(Conditional Covering Lemma)*** *For any $\epsilon > 0$, we can find n such that if $X^n(m)$, $m \in [1 : 2^{nR}]$, are independently drawn from $\prod_{i=1}^n P_{X|U}(x_i|u_i)$ and $Y^n \in \mathcal{T}_{\epsilon'}^{(n)}(Y)$, $\epsilon' < \epsilon$, is conditionally independent of each $X^n(m)$ given U^n, then*

$$\Pr\{(X^n(m), Y^n, U^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,U) \text{ for some } m\} \to 1$$

if $R > I(X;Y|U) + \delta(\epsilon)$, where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ [32].

A.2 Packing Lemma

The packing lemma generalizes the bound on the probability of the decoding error event in the achievability proof of the channel coding theorem. The lemma will be used in the achievability proofs of many multi-user source and channel coding theorems.

Lemma 3. ***(Packing Lemma)*** *For any $\epsilon > 0$, we can find n such that if $X^n(m)$, $m \in [1 : 2^{nR}]$, are independently drawn from $\prod_{i=1}^n P_X(x_i)$ and $Y^n \in \mathcal{T}_\epsilon^{(n)}(Y)$ is independent of each $X^n(m)$, then*

$$\Pr\{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y) \text{ for some } m\} \to 0$$

if $R < I(X;Y) - \delta(\epsilon)$, where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ [32].

Proof.

$$\begin{aligned}
&\Pr\{(X^n(m), Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y) \text{ for some } m\} \\
&= \sum_{y^n \in \mathcal{T}_\epsilon^{(n)}(Y)} p(y^n) \Pr\{(X^n(m), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y) \text{ for some } m \mid y^n\} \\
&\leq \sum_{y^n \in \mathcal{T}_\epsilon^{(n)}(Y)} p(y^n) \sum_{m=1}^{2^{nR}} \Pr\{(X^n(m), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\} \\
&\leq \sum_{y^n \in \mathcal{T}_\epsilon^{(n)}(Y)} p(y^n) 2^{nR}.2^{-n(I(X;Y)-\delta(\epsilon))} \\
&\leq 2^{-n(I(X;Y)-R-\delta(\epsilon))})
\end{aligned}$$

which tends to zero as $n \to 0$ if $R < I(X;Y) - \delta(\epsilon)$. This completes the proof of the packing lemma. □

We can generalize the packing lemma by replacing the independence to the conditional independence condition. Then we have the following lemma.

Lemma 4. ***(Conditional Packing Lemma)*** *For any $\epsilon > 0$, we can find n such that if $X^n(m)$, $m \in [1 : 2^{nR}]$, are independently drawn from $\prod_{i=1}^n P_{X|U}(x_i|u_i)$ and $Y^n \in \mathcal{T}_\epsilon^{(n)}(Y)$ is conditionally independent of each $X^n(m)$ given U^n, then*

$$\Pr\{(X^n(m), Y^n, U^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,U) \text{ for some } m\} \to 0$$

if $R < I(X;Y|U) - \delta(\epsilon)$, where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ [32].

A.3 Proof of Lemma 11.1. of Network Information Theory book (El-Gamal book)

Lemma 5.

$$\begin{aligned}
&\Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(M'),\ \tilde{l} \neq L\} \\
&\leq \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1)\}[1].
\end{aligned}$$

Proof. We first show that

$$\begin{aligned}&\Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(m'),\ \tilde{l} \neq L \mid M' = m'\}\\ &\leq \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1) \mid M' = m'\}\end{aligned}$$

This holds trivially when $m' = 1$. For $m' \neq 1$, consider

$$\begin{aligned}&\Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(m'),\ \tilde{l} \neq L \mid M' = m'\}\\ &= \sum_{l \in \mathcal{B}(m')} p(l|m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(m'),\ \tilde{l} \neq L \mid L = l, M' = m'\}\\ &\overset{(a)}{=} \sum_{l \in \mathcal{B}(m')} p(l|m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(m'),\ \tilde{l} \neq L \mid L = l\}\\ &\overset{(b)}{=} \sum_{l \in \mathcal{B}(m')} p(l|m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in [1 : 2^{n(R_1-R)} - 1] \mid L = l\}\\ &\leq \sum_{l \in \mathcal{B}(m')} p(l|m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1) \mid L = l\}\\ &\overset{(c)}{=} \sum_{l \in \mathcal{B}(m')} p(l|m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1) \mid L = l, M' = m'\}\\ &= \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1) \mid M' = m'\}\end{aligned}$$

where (a) and (c) follow since M is a function of L and (b) follows since given $L = l$, any collection of $2^{n(R_1-R)} - 1$ codewords $U^n(\tilde{l})$ with $\tilde{l} \neq l$ has the same distribution. Therefore, We have

$$\begin{aligned}P(\mathcal{E}_3) &= \sum_{m'} p(m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(m'),\ \tilde{l} \neq L \mid M' = m'\}\\ &\leq \sum_{m'} p(m') \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1) \mid M' = m'\}\\ &= \Pr\{(U^n(\tilde{l}), Y^n) \in \mathcal{T}_\epsilon^{(n)}(U,Y) \text{ for some } \tilde{l} \in \mathcal{B}(1)\}[1].\end{aligned}$$

□