On Optimum Conventional Quantization for Source Coding with Side Information at the Decoder

by

Lin Zheng

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, 2007

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.

Lin Zheng

Abstract

In many scenarios, side information naturally exists in point-to-point communications. Although side information can be present in the encoder and/or decoder and thus yield several cases, the most important case that worths particular attention is source coding with side information at the decoder (Wyner-Ziv coding) which requires different design strategies compared to the the conventional source coding problem. Due to the difficulty caused by the joint design of random variable and reconstruction function, a common approach to this lossy source coding problem is to apply conventional vector quantization followed by Slepian-Wolf coding. In this thesis, we investigate the best rate-distortion performance achievable asymptotically by practical Wyner-Ziv coding schemes of the above approach from an information theoretic viewpoint and a numerical computation viewpoint respectively.

From the information theoretic viewpoint, we establish the corresponding rate-distortion function $\hat{R}_{WZ}(D)$ for any memoryless pair (X, Y) and any distortion measure. Given an arbitrary single letter distortion measure d, it is shown that the best rate achievable asymptotically under the constraint that X is recovered with distortion level no greater than $D \geq 0$ is $\hat{R}_{WZ}(D) = \min_{\hat{X}}[I(X;\hat{X}) - I(Y;\hat{X})]$, where the minimum is taken over all auxiliary random variables $\hat{X}$ such that $Ed(X,\hat{X}) \leq D$ and $\hat{X} \to X \to Y$ is a Markov chain.

Further, we are interested in designing practical Wyner-Ziv coding. With the characterization at $\hat{R}_{WZ}(D)$, this reduces to investigating $\hat{X}$. Then from the viewpoint of numerical computation, the extended Blahut-Arimoto algorithm is proposed to study the rate-distortion performance, as well as determine the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$ which provids guidelines for designing practical Wyner-Ziv coding.

In most cases, the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$ is different from the random variable $\hat{X}'$ that achieves the classical rate-distortion $R(D)$ without side information at the decoder. Interestingly, the extended Blahut-Arimoto algorithm allows us to observe an interesting phenomenon, that is, there are indeed cases where $\hat{X} = \hat{X}'$. To gain deep

insights of the quantizer's design problem between practical Wyner-Ziv coding and classic rate-distortion coding schemes, we give a mathematic proof to show under what conditions the two random quantizers are equivalent or distinct. We completely settle this problem for the case where $\mathcal{X}$, $\mathcal{Y}$, and $\hat{\mathcal{X}}$ are all binary with Hamming distortion measure. We also determine sufficient conditions (equivalent condition) for non-binary alphabets with Hamming distortion measure case and Gaussian source with mean-squared error distortion measure case respectively.

Acknowledgements

The author wishes to express her sincere gratitude and appreciation to all of them who have made this thesis successful and meaningful. The author is grateful to her supervisor, Dr. En-hui Yang, for his invaluable and constant guidance through out her master study period. He is more a mentor than a mere advisor to her. He set a high standard for her that really helps her grow.

Also special thanks to Dr. Da-ke He of IBM T. J. Watson Research Center, Yorktown Heights, NY, Dr. Haiquan Wang and Professor L. L. Xie of Department of Electrical and Computer Engineering, University of Waterloo, for their valuable discussions.

Third, the author is deeply indebted to her friends in Multimedia Communications Laboratory at University of Waterloo, Dr. Wei Sun, Dr. Xudong Ma, Mr. Xiang Yu, Mr. Hui Zha, Miss Jiao Wang, Mr. Abir Mukherjee, Mr. Jin Meng, Mr. Yuhan Zhou and many other friends for their help, and discussions.

Last, but not the least, the author is much obliged to her parents, for their love, understanding and support in her life.

Contents

List of Figures

Chapter 1

Introduction

1.1 Distributed Source Coding

Source coding is a way to remove the uncontrolled redundancy occurring in the original information source so as to reduce the bandwidth of signal for it to be accommodated in the channel. For example, we hardly see the difference of consecutive frames in a slowly varying video sequence. Therefore, we can predict most pixels in the next frame by observing the first frame, such that the most pixels in next frame are redundant which can be removed hence "compresses" the source. Source coding can be either lossless or lossy. Lossless source coding is the compression of a signal where the decompression gives back to the original signal. Slepian-Wolf coding is a case of lossless coding. Lossy source coding achieves greater compression by throwing away some information of the signal that doesn't matter. Wyner-Ziv coding is a case of lossy coding.

Distributed source coding of correlated sources, refers to the compression of the outputs of two or more physically separated correlated sources which do not communicate with each other (hence distributed coding)([16], [17]). These sources send their compressed outputs to a central point (e.g., the base station) for joint decoding (See Fig 1.1).

Distributed source coding is a new coding paradigm based on two information theoretic results: Slepian-Wolf and Wyner-Ziv theorems which will be introduced in latter sections.

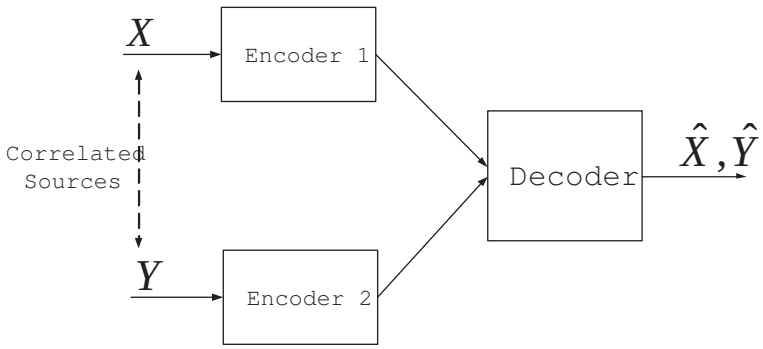

Figure 1.1: Distributed source coding with separated encoding and joint decoding

Based on the distributed source coding independent encoding and joint decoding configuration, many applications involves distributed source coding, such as data compression for network communications, sensor networks, upgrading of existing schemes and video compression comes.

1.1.1 Distributed Source Coding in Lossless Case

In this section, we are considering the case that the sources X and Y are recovered perfectly at the decoder in Fig 1.1 which is called Slepian-Wolf coding. The problem of lossless compression of finite alphabet sources takes its roots from the fundamental paper of Slepian and Wolf [4]. The Slepian-Wolf theorem shows that the output of two correlated sources can be compressed to the same extent without loss. Consider two correlated independent identically distributed (i.i.d.) finite-alphabet random sequences X and Y. With separate conventional entropy encoders and decoders, one can achieve $R_X \geq H(X)$ and $R_Y \geq H(Y)$ [2], where $H(X)$ and $H(Y)$ are the entropies of X and Y, respectively. Interestingly, we can do better with joint decoding. In this case, the Slepian-Wolf theorem establishes the rate region (Fig 1.2) which is bounded by the following inequalities.

$$\begin{aligned} R_X \geq H(X|Y), R_Y \geq H(Y|X) \\ R_X + R_Y \geq H(X,Y) \end{aligned} \tag{1.1}$$

Surprisingly, just as joint encoding of X and Y, the sum of rates $R_X + R_Y$ can achieve

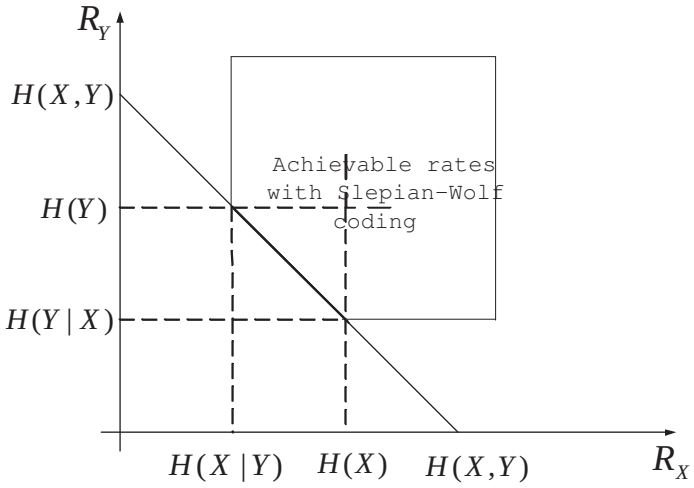

Figure 1.2: The Slepian-Wolf rate region for two sources

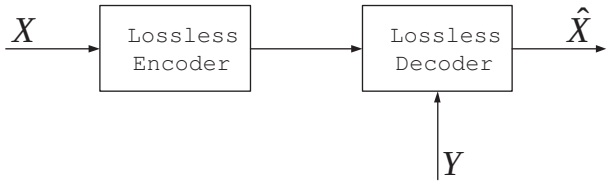

Figure 1.3: Lossless coding with side information

$H(X,Y)$, despite encoding X, Y separately. Compression with side information at the decoder (Fig 1.3) is a special case of the distributed coding problem in Fig 1.1. The source produces a sequence X with correlated side information Y, and at the decoder only X is recovered with an arbitrarily small probability of error. Since R_Y is achievable for conventional encoding, compression with receiver side information corresponds to one of the corners of the rate region in Fig 1.2, hence $R_X \geq H(X|Y)$.

1.1.2 Distributed Source Coding in Lossy Case

Consider again the problem of Fig 1.1. If there exists some distortion criterion between the sources and reconstruction outputs, i.e., lossy distributed source coding, the general problem is still open, although Gaussian case has been solved by Aaron, Saurabha, and

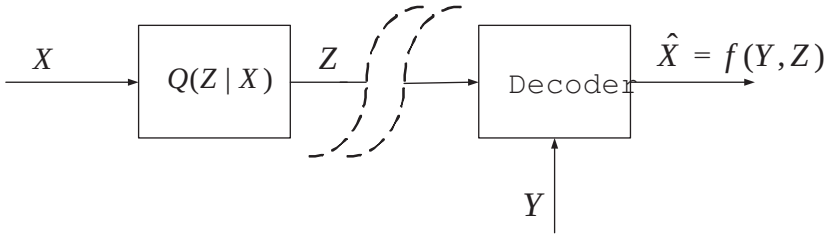

Figure 1.4: Rate distortion with side information

Pramod recently [5].

One special lossy case that has been discussed a lot is Wyner-Ziv coding (Fig 1.4). Shortly after Slepian-Wolf theorem, Wyner and Ziv have extended their work to establish information theoretic bounds for lossy compression with side information at the decoder. In [3], the Wyner-Ziv rate distortion function $R_{WZ}(D)$ gives the minimum rate necessary to reconstruct the source X with distortion constraint $Ed(X, \hat{X}) \leq D$,

$$R_{WZ}(D) = \inf_{p(z|x)} \inf_{f(y,z):Ed(x,f(y,z))\leq D} I(X;Z) - I(Y;Z) \tag{1.2}$$

where the minimization is taken over all $p(z|x)$ and all reconstruction functions $f(y, z)$ satisfying fidelity constraints. Z is an auxiliary random variable such that $Y \rightarrow X \rightarrow Z$ forms a Markov chain.

1.1.3 Practical Wyner-Ziv Coding and its Application

With the characterization of $R_{WZ}(D)$ in (1.2), we see that in order to achieve the best rate, one has to jointly design the auxiliary random variable and the reconstruction function with the given distortion measure. In general, this joint design problem is hard to solve. A simpler and more practically relevant problem is to design the auxiliary random variable for a fixed reconstruction function. In this thesis, we are focus on the commonly used practical Wyner-Ziv coding schemes comprising of conventional vector quantization followed by Slepian-Wolf coding (Fig 1.5).

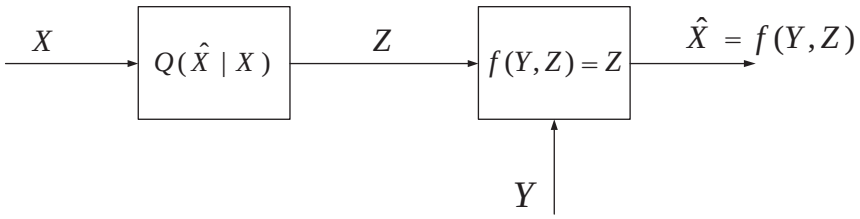

Figure 1.5: Practical W-Z system comprising of vector quantization and S-W coding

In this approach, it is implicitly assumed that $\mathcal{Z} = \hat{\mathcal{X}}$, and the reconstruction function is fixed as $f(Y, Z) = Z$, where Z can be regarded as the reconstructed output of a vector quantizer in response to input X.

In late chapters, we show that the minimum rate in bits per letter achievable asymptotically for (X, Y) under the constraint that X is recovered with distortion level no greater than $D \geq 0$ is given by

$$\hat{R}_{WZ}(D) \stackrel{\Delta}{=} \min_{\hat{X}} [I(X; \hat{X}) - I(Y; \hat{X})] \tag{1.3}$$

where the minimum is taken over all auxiliary random variables $\hat{X}$ from $\hat{\mathcal{X}}$ such that $\hat{X} \rightarrow X \rightarrow Y$ is a Markov chain, and $Ed(X, \hat{X}) \leq D$. A detailed explanation of this system and its actual implementation is explained in Chapter 3.

In the study of distributed source coding, it was considered to use the quantization at the encoder, thus practical Wyner-Ziv coding schemes provide practical solutions for designing the encoders and the decoders, for implementation of the source coding schemes. A source data observation is simply quantization-based encoded, and transmitted to a decoder. The decoder, with the help of an available uncoded source data correlated to the source, attempts to obtain the original source data.

Practical Wyner-Ziv coding, i.e., lossy compression with decoder side information, enables low-complexity video encoding where the bulk of the computation is shifted to the decoder. This idea is widely used in distributed video coding schemes, which is a new coding paradigm described by a configuration where the encoder has low-complexity at

the expense of a higher decoder complexity ([6], [7]).

1.2 Research Problems and Motivations

In this thesis, we are interested in the rate-distortion performance achievable asymptotically by practical Wyner-Ziv coding schemes comprising of conventional vector quantization followed by Slepian-Wolf coding.

The research problems to be investigated in this thesis are:

1. "From the viewpoint of information theory, for practical Wyner-Ziv coding schemes that referred to above, what is the best rate $\hat{R}_{WZ}(D)$ achievable asymptotically by this approach? Or what is the minimum rate in bits per letter under the given distortion constraint between the source and reconstruction output?"

2. "From the viewpoint of computation, how to calculate $\hat{R}_{WZ}(D)$ efficiently?"

3. "How to design practical Wyner-Ziv coding?"

4. "Under what conditions is the quantizer achieving $\hat{R}_{WZ}(D)$ the same as or different from the quantizer achieving the classical rate-distortion function? Equivalently, under what conditions should the design of conventional quantization in the case of side information be different from the case of no side information?"

The motivations for above research problems are three-fold. First, in existing information theoretic works on Wyner-Ziv coding, in order to achieve (1.2), one has to jointly design an auxiliary random variable and a reconstruction function with the given distortion measure. In general, this joint design problem is hard to solve, such that we investigate a commonly used approach to apply conventional vector quantization followed by Slepian-Wolf coding. To study the rate-distortion performance achievable asymptotically by this approach, we are naturally led to the first question.

Second, although the derived rate-distortion function $\hat{R}_{WZ}(D)$ can be characterized as optimization problem, the characterization does not mean that it can be calculated easily. Indeed, the closed forms of $\hat{R}_{WZ}(D)$ are known only to very few cases, such as the case when X, Y jointly Gaussian which will be discussed in Chapter 4. Generally, such optimization problem is difficult to solve. At the same time, how to provide guidelines for designing practical Wyner-Ziv coding is another key problem. With the characterization of $\hat{R}_{WZ}(D)$ in (1.3), this reduces to investigate $\hat{X}$. Therefore, it is important and necessary to propose an efficient algorithm (extended Blahut-Arimoto algorithm) for numerically computing the rate-distortion function, as well as determining the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$.

Third, comparing with the classic rate-distortion function $R(D)$ ([2]) without side information at the decoder, the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$ should be generally different from the random variable $\hat{X}'$ that achieves $R(D)$, due to the presence of decoder only side information Y in practical Wyner-Ziv coding schemes. Interestingly, the extended Blahut-Arimoto algorithm allows us to observe that there are indeed cases where $\hat{X}' = \hat{X}$. To fully understand and characterize this important and rather surprising phenomenon, we are led to the fourth question which provides guidelines to design practical Wyner-Ziv coding and classic lossy coding.

1.3 Organization of Thesis and Contribution

This thesis is organized as follows.

In Chapter 2, we give a quick review of the background knowledge on conventional vector quantization and source coding techniques. Some previous theoretical results are presented such as classic rate-distortion function, Slepian-Wolf coding and Wyner-Ziv coding.

In chapter 3, we study the problem of practical Wyner-Ziv coding comprising of con-

ventional vector quantization followed by Slepian-Wolf coding. Given an arbitrary single letter distortion measure, the best rate-distortion function achievable asymptotically by this approach has been characterized. Next, we obtain a computationally efficient algorithm (extended Blahut-Arimoto algorithm) for this problem. The algorithm allows us to numerically calculate $\hat{R}_{WZ}(D)$, as well as compare the random variable $\hat{X}$ achieving $\hat{R}_{WZ}(D)$ and the random variable $\hat{X}'$ achieving the classical rate-distortion function, and thus observe an interesting phenomenon that in some cases these two random variables are exactly the same.

To fully understand and characterize this important and rather surprising phenomenon, in Chapter 4, we deal with the problem "under what conditions is the quantizer achieving $\hat{R}_{WZ}(D)$ the same as or different from the quantizer achieving the classical rate-distortion function?" Finally, the conclusion of this thesis is in Chapter 5, including some future works.

Now we summarize the contributions of this thesis. It characterizes the rate-distortion function of practical Wyner-Ziv coding comprising of conventional vector quantization and Slepian-Wolf coding. It proposes an extended Blahut-Arimoto algorithm to study the performance of practical Wyner-Ziv coding schemes, as well as provides guidelines for designing it. Although the presence of decoder only side information in practical Wyner-Ziv coding makes the quantization design generally different from the classic rate-distortion function without side information at the decoder, we give a mathematical proof to answer Question 4 raised in last section for binary alphabets with Hamming distortion measure. Furthermore, we determine sufficient conditions (equivalent condition) for non-binary alphabets with Hamming distortion measure case and Gaussian source with mean-squared error distortion measure case respectively.

1.4 Notations

Throughout the thesis, the following notations are adopted. We use capital letter to denote random variable, lowercase letter for its realization, and script letter for its alphabet. For

instance, X is a random variable over its alphabet $\mathcal{X}$ and $x \in \mathcal{X}$ is a realization. We use $p_X(x)$ to denote the probability distribution of a discrete random variables X taking values over its alphabet $\mathcal{X}$, and also to denote the probability density function of a continuous random variable X. If there is no ambiguity, sometimes $p_X(x)$ is omitted and we write $p(x)$ instead. Furthermore, E denote the expectation operator, $H(X)$ is the entropy of X, and $I(X;Y)$ denote the mutual information between X and Y.

Chapter 2

Theory for Source Coding

In this thesis, the investigated practical Wyner-Ziv coding uses the system model of source coding with side information comprising of conventional vector quantization followed by Slepian-Wolf coding. Before describing the detailed framework of practical Wyner-Ziv coding in Chapter 3, we present some of the required background knowledge on conventional vector quantization and source coding in this chapter, as well as the preliminaries on typicality which is an important tool in proving coding theorems. We reviewed the classic source coding and source coding with side-information including both the lossless case (Slepian-Wolf coding) and the lossy case with distortion constraint (Wyner-Ziv coding).

2.1 Conventional Vector Quantization Review

Quantization is one of the most common and direct techniques to achieve data compression. There are two basic quantization types: scalar and vector. Scalar quantization encodes data points individually, while vector quantization groups input data into vectors, each of which is encoded as a whole. Vector quantization typically searches a codebook (a collection of vectors) for the closest match to an input vector, yielding an output index. A dequantizer simply performs a table lookup in an identical codebook to reconstruct the original vector.

Generally, a vector quantization encoder or a vector quantization decoder has a single codebook containing a plurality of code vectors with indices. According to the indices, encoding and decoding processes are carried out. By increasing the quantity of the code vectors stored in the codebook, the quality of the reproduced signal may be improved.

The conventional vector quantization technique has only one codebook which stores a plurality of code vectors C_1 to C_n that are selectively output according to input indices. To accurately reproduce an encoded signal wave shape with such a conventional vector quantization encoder or decoder, it is necessary to reduce quantization distortion. To do so, the number of code vectors stored in the codebook must be increased. To increase the number of code vectors, it is necessary to increase the memory size of the codebook.

2.2 Preliminaries 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 ([2], [15]) needed in the latter proofs.

Definition 1 *A sequence $x^n \in \mathcal{X}^n$ is said to be ϵ-strongly typical with respect to a distribution $p(x)$ on $\mathcal{X}$ if*
1 for all $a \in \mathcal{X}$ with $p(a) > 0$, we have

$$\|\frac{1}{n}N(a|x^n) - p(a)\| < \frac{\epsilon}{|\mathcal{X}|}; \tag{2.1}$$

2 and for all $a \in \mathcal{X}$ with $p(a) = 0$, $N(a|x^n) = 0$,

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

Definition 2 *A pair of sequences $(x^n, y^n) \in \hat{\mathcal{X}}^n \times \hat{\mathcal{Y}}^n$ is said to be ϵ-strongly typical with respect to a distribution $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ if*

1 for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ *with* $p(a,b) > 0$*, we have*

$$\|\frac{1}{n}N(a,b|x^n,y^n) - p(a,b)\| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}; \tag{2.2}$$

2 and for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ *with* $p(a,b) = 0$*,* $N(a,b|x^n,y^n) = 0$*.*

where$N(a,b|x^n,y^n)$ *is the number of occurrences of the symbol* (a,b) *in the sequence* (x^n,y^n)*.*

The set of all ϵ-strongly typical sequences $x^n \in \mathcal{X}^n$ with respect to $p(x)$ is denoted by $A_\epsilon^{*(n)}(X)$, and the set of all jointly ϵ-strongly typical sequences $(x^n,y^n) \in \hat{\mathcal{X}}^n \times \hat{\mathcal{Y}}^n$ with respect to $p(x,y)$ is denoted by $A_\epsilon^{*(n)}(X,Y)$.

Lemma 1 *Let* X_i *be drawn i.i.d.* $\sim p(x)$*. Then* $Pr(A_\epsilon^{*(n)}(X)) \to 1$ *as* $n \to \infty$*.*

Lemma 2 *Let* (X_i,Y_i) *be drawn i.i.d.* $\sim p(x,y)$*. Then* $Pr(A_\epsilon^{*(n)}(X,Y)) \to 1$ *as* $n \to \infty$*.*

Lemma 3 *Let* $Y_1,Y_2,...,Y_n$ *be drawn i.i.d.* $\sim \prod p(y)$*. For* $x^n \in A_\epsilon^{*(n)}(X)$*, the probability that* $(x^n,Y^n) \in A_\epsilon^{*(n)}$ *is bounded by*

$$2^{-n(I(X;Y)+\epsilon_1)} \leq Pr((x^n,Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_1)} \tag{2.3}$$

where ϵ_1 *goes to* 0 *as* $\epsilon \to 0$ *and* $n \to \infty$*.*

Lemma 4 *Let* (X,Y,Z) *form a Markov chain* $X \to Y \to Z$*, i.e.,* $p(x,y,z) = p(x,y)p(z|y)$*. If for a given* $(y^n,z^n) \in A_\epsilon^{*(n)}(Y,Z)$*,* X^n *is drawn* $\sim \prod_{i=1}^n p(x_i|y_i)$*, then* $Pr\{(X^n,y^n,z^n) \in A_\epsilon^{*(n)}(X,Y,Z)\} > 1-\epsilon$ *for* n *sufficiently large.*

2.3 Source Coding Background Review

The function of the source code (consisting of the source encoder and source decoder) is to remove the uncontrolled redundancy naturally occuring in the original information sources so as to provide an efficient representation for the source output. The primary benefit

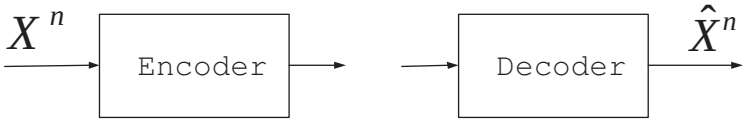

Figure 2.1: Source Coding

gained from the application of source coding is a reduced symbol throughput requirement and thus a better bandwidth efficiency.

Depending on the nature of the source output, source codes can be either lossless or lossy. In this section, we review source coding including lossless coding and coding with a distortion constraint.

2.3.1 Source Coding without Side Information

Before discussing source coding with side-information, we review basic source coding as shown in Fig 2.1.

Lossless Encoding

Let X be a discrete random variable taking values in a finite set $\mathcal{X}$, and $\{X_i\}_{i=1}^n$ be an i.i.d. sequence drawn according to distribution $p(x)$. The encoding and decoding mappings with block length n are:

$$f : \mathcal{X}^n \to \{0,1\}^* \tag{2.4}$$

$$g : \{0,1\}^* \to \mathcal{X}^n \tag{2.5}$$

The decoder is interested in recovering $\{X_i\}_{i=1}^n$ (which we write as X^n) with high probability, i.e.

$$P_e^{(n)} = P(g(f(X^n)) \neq X^n) \to 0 \text{ as } n \to \infty \tag{2.6}$$

and we define its average transmission rate in bits per symbol as:

$$\begin{aligned} R &= \frac{1}{n} E|f(X^n)| \\ &= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} p(x^n)|f(x^n)| \end{aligned} \tag{2.7}$$

where $|b|$ denotes the length of a binary string b.

The set of representation codewords $(g_n(1), g_n(2), ..., g_n(2^{|f(x^n)|}))$ constitutes the codebook corresponding to x^n. The source is mapped to one of the codewords in the codebook and the index of that codeword is made available to the decoder. R bits per symbol on average is required. We are interested in designing the mappings f and g so as to minimize the average transmission rate R in order for the receiver to recover the original source X^n perfectly. Information theory states that the rate region here is:

$$R \geq H(X) \tag{2.8}$$

Encoding with a distortion criterion

The description of an arbitrary real number requires an infinite number of bits, so a finite representation of a continuous random variable can never be perfect. Hence, after defining a distortion measure which is a measure of distance between the random variable and its reproduction, we remove the constraint on X to be discrete and allow it to be both discrete and continuous.

Consider again the problem of Fig 2.1. We are now interested in recovering X^n at the decoder within a distortion constraint D for some distortion measure $d(x, \hat{x})$.

Definition 3 *A distortion measure is a mapping*

$$d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathcal{R}^+ \tag{2.9}$$

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 $\hat{x}$.

Definition 4 *A distortion measure is said to be bounded if*

$$d_{max} \stackrel{\Delta}{=} \sup_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty \tag{2.10}$$

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

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{2.11}$$

The following distortion measures are widely used in practice.

Hamming (probability of error) distortion: The Hamming distortion is given by

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases}, \tag{2.12}$$

which results in a probability of error distortion, since $Ed(X, \hat{X}) = Pr(X \neq \hat{X})$. This distortion measure is usually used for discrete alphabets.

Squared error distortion: The squared error distortion,

$$d(x, \hat{x}) = (x - \hat{x})^2, \tag{2.13}$$

is a popular distortion measure used for continuous alphabets. In latter chapters, we bring in these two common used distortion measures to discuss the optimum conventional quantization for discrete and Gaussian cases separately.

Definition 5 *A* $(2^{nR}, n)$ *rate distortion code consists of an encoding function,*

$$f_n : \mathcal{X}^n \to \{0, 1\}^*, \tag{2.14}$$

and a decoding(reproduction) function,

$$g_n : \{0, 1\}^* \to \hat{\mathcal{X}}^n. \tag{2.15}$$

The average transmission rate R in bits per symbol is defined as

$$\begin{aligned} R &= \frac{1}{n} E|f_n(X^n)| \\ &= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} p(x^n)|f_n(x^n)| \end{aligned} \tag{2.16}$$

where $|b|$ denotes the length of a binary string b.

The distortion associated with the $(2^{nR}, n)$ code is defined as

$$D = Ed(X^n, g_n(f_n(X^n))), \tag{2.17}$$

where the expectation is with respect to the probability distribution on X, i.e.,

$$D = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))). \tag{2.18}$$

Thus the input source space $\mathcal{X}^n$ is first partitioned into $2^{|f_n(X^n)|}$ disjoint regions through the mapping f_n. Each region in the partition is associated with a reconstruction codeword. The set of n-tuples $(g_n(1), g_n(2), ..., g_n(2^{|f_n(x^n)|})$, denoted by $\hat{X}^n(1), \hat{X}^n(2), ...\hat{X}^n(2^{|f_n(x^n)|})$, constitutes the codebook corresponding to x^n. The source is quantized to one of the codewords in the corresponding codebook and the index of that codeword is made available to the decoder. This requires R bits per symbol on average. The problem is to design the mappings f_n and g_n so as to minimize the average transmission rate R in order for the receiver to recover the original source X^n such that the distortion level is no greater than a given distortion constraint D.

We have the following definitions[2].

Definition 6 *A rate distortion pair (R, D) is said to be achievable if there exists a sequence of $(2^{nR}, n)$ rate distortion codes (f_n, g_n) with $\lim_{n\to\infty} Ed(X^n, g_n(f_n(X^n))) \leq D$.*

Definition 7 *The rate distortion region for a source is the closure of the set of achievable pairs (R, D).*

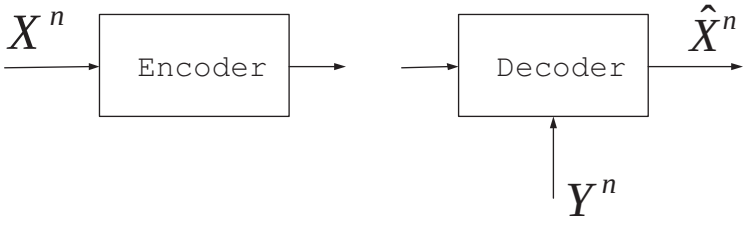

Figure 2.2: Source coding with side information only to the decoder

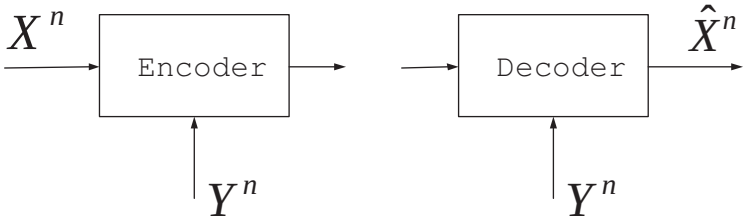

Figure 2.3: Source coding with side information to both the encoder and decoder

Definition 8 *The rate distortion function $R(D)$ is the infimum of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.*

We thus have the following theorem [2],[15]:

Theorem 1 *The rate distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is*

$$R(D) = \inf_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \tag{2.19}$$

2.3.2 Source Coding with Side Information

In this section, we review the concepts of source coding with side-information, also known as Distributed Source Coding. We consider two scenarios: one is that the side-information Y^n is available only to the decoder (see Fig 2.2), while the other one is that the side-information presents at both encoder and decoder (see Fig 2.3).

Lossless Encoding

In both cases of Fig 2.2 and Fig 2.3 , let X and Y be two random variables taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of independent copies of (X, Y). The decoder is interested in recovering X^n perfectly with high probability, i.e.,

$$P_e^{(n)} = P(\hat{X}^n \neq X^n) \to 0 \text{ as } n \to \infty \tag{2.20}$$

When the side-information is available to both encoder and decoder(Fig 2.3), then the problem of compressing X is well-understood: one can compress X at a theoretical rate of its conditional entropy given Y, $H(X|Y)$. Surprisingly, Slepian and Wolf [4] showed that, if Y were known only at the decoder for X and not at the encoder (see Fig 2.2), one can still compress X using only $H(X|Y)$ bits, the same as the case where the encoder does know Y. That is , by just knowing the joint distribution of X and Y, without explicitly knowing Y, the encoder of X can perform as well as an encoder which explicitly knows Y.

Typicality was used in [2] to encode and decode X^n to achieve the minimum rate $H(X|Y)$, which can be described as follows.

Generation of codebooks. Randomly bin all the sequences x^n into 2^{nR} bins by independently generating an index b uniformly distributed on $\{1, 2, ..., 2^{nR}\}$ for each x^n. Let $B(i)$ denote the set of sequences x^n allotted to bin i.

Encoding. The sender sends the index i of the bin in which x^n falls.

Decoding. The receiver looks for a unique $x^n \in B(i)$ such that $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$. If there is none or more than one, it declares an error.

Analysis of the probability of error.

1 The pair (X^n, Y^n) generated by the source is not typical. The probability of this is small if n is large.

2 There exists another typical $x^n \in B(i)$ which is jointly typical with y^n. The probability that any other x^n is jointly typical with y^n is less than $2^{-n(I(X;Y)-3\epsilon)}$, and therefore the probability of this kind of error is bounded above by

$$E[|B(i) \bigcap A_\epsilon^{*(n)}(X)|2^{-n(I(X;Y)-3\epsilon)}] \leq 2^{n(H(X)+\epsilon)}2^{-nR}2^{-n(I(X;Y)-3\epsilon)}, \tag{2.21}$$

which goes to 0 if $R > H(X|Y) + 2\epsilon$.

Later in 1999, Pradhan and Ramchandran's work [1] provided a constructive practical framework based on algebraic trellis codes dubbed as Distributed source coding using syndromes that can be applicable in a variety of settings. It is instructive to examine the following example from [1] inspired by Wyner's idea in 1974 [8].

Assume X and Y are equiprobable binary triplets with $X, Y \in \{0,1\}^3$ and they differ in at most one position. Then $H(X) = H(Y) = 3$ bits. Because the Hamming distance between X and Y is $d_H(X,Y) \leq 1$, for a given Y (e.g., [101]), then X is either the same as Y ([101]) or differ in one position such as $[001], [111], [100]$. Hence $H(X|Y) = 2$ bits. We will see how to describe X with the same compression rate $H(X|Y)$ so that it can be perfectly reconstructed at the decoder in the above two scenarios.

Scenario 1: In this scenario, the side-information Y is available at both encoder and decoder. Clearly X can be predicted from Y. There are only 4 possibilities for the modulo-two binary sum of X and Y and hence X can be encoded with 2 bits given Y.

Scenario 2: In this scenario, Y is only revealed to the decoder but not the encoder. However, the encoder does know the correlation structure and also knows that the decoder has access to Y. By the Slepian-Wolf theorem, it is still possible to describe X with just 2 bits and decode it without loss at the joint decoder. This can be done by first partitioning the set of all possible outcomes of X into four bins, $Z_{00} = \{000, 111\}, Z_{01} = \{001, 110\}, Z_{10} = \{010, 101\}$ and $Z_{11} = \{011, 100\}$. The encoder for X identifies the set containing the codeword for X and sends the index for the set (which can be done in 2 bits) instead of the individual codeword. In forming the bins Z_s, we make sure that each of them has two elements with Hamming distance $d_H = 3$. On the decoder side, on the

reception of the coset index, with the help of side information Y, we pick in bin Z_s the X with $d_H(X,Y) \leq 1$. Unique decoding is guaranteed because the two elements in each bin Z_s have Hamming distance $d_H = 3$.

In channel coding terminology, each coset is associated with a unique syndrome [14]. Since the encoder send the syndrome for the coset containing the codeword for X to the decoder, we refer to this operation as **Syndrome Coding**.

Encoding with a distortion criterion

Consider the problem of Fig 2.2. Wyner and Ziv [3] studied this system for lossy compression with side information at the decoder. Here, the constraint on X and Y to be discrete can be removed, and we allow them to be continuous random variables as well. The source X is encoded without access to the side information Y. The decoder, however, has access to Y, and must recover X^n within a distortion constraint D for some distortion measure $d(x,\hat{x})$. Let $\{X_i, Y_i\}_{i=1}^n$ be i.i.d. $\sim p(x,y)$ and let the distortion measure be $d(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i,\hat{x}_i)$. Then the Wyner-Ziv theorem ([2],[3]) states that the rate distortion function for this problem is

$$R_{WZ}(D) = \inf_{p(z|x)} \inf_{f(y,z)} I(X;Z) - I(Y;Z) \tag{2.22}$$

where the minimization is under the distortion constraint

$$\sum_x \sum_z \sum_y p(x,y)p(z|x)d(x,f(y,z)) \leq D \tag{2.23}$$

where the minimization is taken over all $p(z|x)$ and all reconstruction functions $f(y,z)$ satisfying fidelity constraints, and Z is the active source codeword and the term $I(Y;Z)$ is the rate rebate due to the presence of the side-information at the decoder. We denote by $R'_{WZ}(D)$ the rate required if the side information were available at the encoder as well. Wyner and Ziv proved that, a rate loss $R_{WZ}(D) - R'_{WZ}(D) \geq 0$ is generally incurred when the encoder does not have access to the side information. Surprisingly, for the case when X and Y are jointly Gaussian and the mean squared error is the distortion measure, [3]

showed that $R_{WZ}(D) - R'_{WZ}(D) = 0$. Later in [10] it was shown that this also holds true for the case of $X = Y + N$, where N is independent and identically distributed Gaussian, and the distortion measure is mean squared error.

For encoding and decoding using typicality to achieve the rate-distortion function (2.22), one can refer to [2] for detailed steps.

In practice, due to the difficulty caused by the joint design of random variable and reconstruction function, a common approach to this lossy source coding problem is to apply conventional vector quantization followed by Slepian-Wolf coding. In the next chapter, we investigate the best rate-distortion performance achievable asymptotically by practical Wyner-Ziv coding schemes of the above approach from an information theoretic viewpoint and a numerical computation viewpoint respectively.

Chapter 3

Practical Wyner-Ziv Coding and its Rate-Distortion Function $\hat{R}_{WZ}$

As alluded to in Chapter 2, a typical approach of practical Wyner-Ziv coding is to apply conventional vector quantization followed by Slepian-Wolf coding. In this chapter we first determine the best rate-distortion performance $\hat{R}_{WZ}(D)$ achievable asymptotically by this approach for memoryless source-side information pair (X, Y) with alphabet $\mathcal{X} \times \mathcal{Y}$ and any distortion measure d. Then, we extend the well-known Blahut-Arimoto algorithm to calculate the rate-distortion function $\hat{R}_{WZ}(D)$ and determine the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$. Finally, we observe an interesting phenomenon from the simulation results which will be justified mathematically in the next chapter.

3.1 The Rate-Distortion Function $\hat{R}_{WZ}(D)$

In view of (2.22), we see that in order to achieve $R_{WZ}(D)$, one has to jointly design Z and the reconstruction function f for (X, Y) and the given distortion measure d. In general this joint design problem is hard to solve. A simpler and more practically relevant problem is to design Z for a fixed reconstruction function f. Indeed, in the practice of Wyner-Ziv coding, a common approach is to use conventional vector quantization followed by Slepian-Wolf coding [4](See Fig 1.5).

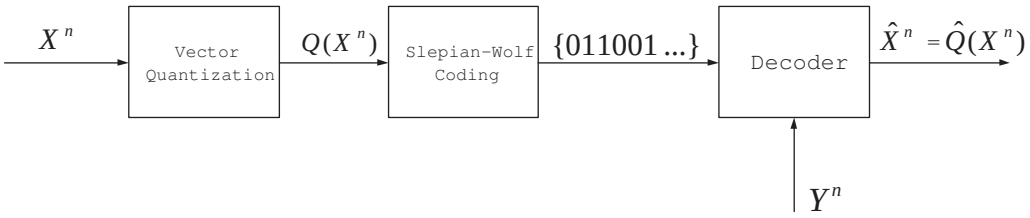

Figure 3.1: The architecture of practical W-Z system

In this approach, it is implicitly assumed that $\mathcal{Z} = \hat{\mathcal{X}}$, and the reconstruction function is fixed as $f(Y, Z) = Z$, where Z can be regarded as the reconstructed output of a vector quantizer in response to input X.

As shown in Fig 1.5, we are interested in the rate-distortion performance achievable asymptotically by the above approach, i.e., conventional vector quantization followed by Slepian-Wolf coding. The main result in this section is the determination of the minimum rate in bits per letter achievable asymptotically for (X, Y) under the constraint that X is recovered with distortion level no greater than $D \geq 0$.

Define, for $D \geq 0$, the quantity

$$\hat{R}_{WZ}(D) \stackrel{\Delta}{=} \min_{\hat{X}} [I(X; \hat{X}) - I(Y; \hat{X})] \tag{3.1}$$

where the minimum is taken over all auxiliary random variables $\hat{X}$ from $\hat{\mathcal{X}}$ such that $\hat{X} \to X \to Y$ is a Markov chain, and $Ed(X, \hat{X}) \leq D$.

To facilitate our discussion, let $C_n = (\phi_n \circ \psi_n, g_n)$ denote an order-n code where ϕ_n denotes a mapping from $\mathcal{X}^n$ to $\hat{\mathcal{X}}^n$, ψ_n denotes a mapping from $\phi_n(\mathcal{X}^n)$, the range of ϕ_n, to a prefix subset of $\{0, 1\}^*$ of finite binary strings, and g_n denotes a mapping from $\mathcal{Y}^n \times \{0, 1\}^*$ to $\phi_n(\mathcal{X}^n)$.

As shown in Fig 3.1 , to encode and decode a sequence $x^n \in \mathcal{X}^n$ with decoder side information y^n, C_n works as follows: on the encoder side, x^n is mapped to $\hat{x}^n = \phi_n(x^n)$ from $\hat{\mathcal{X}}$, and then $\hat{x}^n$ is encoded into a binary string $\psi_n(\hat{x}^n)$; on the decoder side, C_n decodes

$\hat{x}^n$ as $g_n(y^n, \psi_n(\hat{x}^n))$. In view of the above process, we see that if $\phi_n(X^n)$ is denoted by $\hat{X}^n$, then one can regard $\hat{X}^n$ as the reconstructed output of a vector quantizer in response to input X^n, and (ψ_n, g_n) as an order-n Slepian-Wolf code to encode $\hat{X}^n$ with decoder only side information Y^n. When (ψ_n, g_n) satisfies

$$\limsup_{n\to\infty} \Pr\{g_n(Y^n, \psi_n(\hat{X}^n)) \neq \hat{X}^n\} = 0,$$

the operational rate-distortion performance of C_n is characterized by

$$R_{C_n} \stackrel{\Delta}{=} \frac{1}{n} E|\psi_n(\hat{X}^n)|, \tag{3.2}$$

where $|b|$ denotes the length of a binary string b, and

$$D_{C_n} \stackrel{\Delta}{=} \frac{1}{n} E \sum_{i=1}^{n} d(X_i, \hat{X}_i). \tag{3.3}$$

Using the terminology similar to that in [3], we say that a rate distortion pair (R, D) is achievable if, for arbitrary $\epsilon > 0$, there exists a code $C_n = (\phi_n \circ \psi_n, g_n)$ such that

$$R_{C_n} \leq R + \epsilon, \text{ and } D_{C_n} \leq D + \epsilon. \tag{3.4}$$

Let $\hat{\mathcal{R}}$ denote the set of achievable (R, D) pairs, and define

$$R^*(D) \stackrel{\Delta}{=} \inf_{(R,D)\in\hat{\mathcal{R}}} R. \tag{3.5}$$

The following theorem shows that our information theoretic function $\hat{R}_{WZ}(D)$ defined in (3.1) is indeed equal to the rate-distortion function $R^*(D)$ defined in (3.5) above.

Theorem 2 *For any $D \geq 0$, $\hat{R}^*(D) = \hat{R}_{WZ}(D)$.*

Proof of Theorem 2: We first prove the converse of this theorem.

Consider any rate distortion code with side information. Let the encoding function be $f_n = (\phi_n \circ \psi_n) : \mathcal{X}^n \to \{0,1\}^*$ and let $g_{ni} : \mathcal{Y}^n \times \{0,1\}^* \to \hat{\mathcal{X}}^n$ denote the ith symbol produced by the decoding function. Let $T = f_n(X^n)$ denote the corresponding encoded version of X^n. We must show that if $Ed(X^n, g_n(Y^n, f_n(X^n))) \leq D$, then

$R_{C_n} = \frac{1}{n} E|\psi_n(\hat{X}^n)| \geq \hat{R}_{WZ}(D)$.

Similar to the converse part proof of rate distortion with side information in [2], the following chain of inequalities comes:

$$
\begin{aligned}
|\psi_n(\hat{X}^n)| &\geq H(T) \\
&\geq H(T|Y^n) \\
&\geq I(X^n;T|Y^n) \\
&= \sum_{i=1}^{n} I(X_i;T|Y^n,X^{i-1}) \\
&= \sum_{i=1}^{n} H(X_i|Y^n,X^{i-1}) - H(X_i|T,Y^n,X^{i-1}) \\
&= \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T,Y_i,Y^{i-1},Y_{i+1}^n,X^{i-1}) \\
&\geq \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T,Y_i) \\
&\overset{1)}{=} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T,Y_i,\hat{X}_i) \\
&\overset{2)}{\geq} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|Y_i,\hat{X}_i) \\
&= \sum_{i=1}^{n} I(X_i,\hat{X}_i|Y_i) \\
&\overset{3)}{=} \sum_{i=1}^{n} I(X_i,\hat{X}_i) - I(Y_i,\hat{X}_i) \\
&\geq \sum_{i=1}^{n} \hat{R}_{WZ}(Ed(X_i,g_{ni}(\hat{X}_i,Y_i))) \\
&= n\frac{1}{n}\sum_{i=1}^{n} \hat{R}_{WZ}(Ed(X_i,g'_{ni}(\hat{X}_i,Y_i))) \\
&\overset{4)}{\geq} n\hat{R}_{WZ}(E\frac{1}{n}\sum_{i=1}^{n} d(X_i,g'_{ni}(\hat{X}_i,Y_i))) \\
&\geq n\hat{R}_{WZ}(D)
\end{aligned}
$$

where 1) follows from the fact that $X_i \rightarrow (T,Y_i) \rightarrow \hat{X}_i$ forms a Markov chain; 2) fol-

lows from that conditioning reduces entropy; 3) follows from that $Y_i \to X_i \to \hat{X}_i$ forms a Markov chain; and 4) follows from Jensen's inequality and the convexity of $\hat{R}_{WZ}(D)$ (See Proposition 1 in the next section).

Therefore, we have,

$$\begin{aligned} R_{C_n} &= \frac{1}{n} E|\psi_n(\hat{X}^n)| \\ &= \frac{1}{n} \sum_{\hat{x}^n} p(\hat{x}^n)|\psi_n(\hat{x}^n)| \\ &\geq \hat{R}_{WZ}(D) \sum_{\hat{x}^n} p(\hat{x}^n) \\ &= \hat{R}_{WZ}(D) \end{aligned}$$

Next, we prove the achievability of this theorem.

Fix $Q_{\hat{X}|X}(\hat{x}|x)$. Calculate $q_{\hat{X}}(\hat{x}) = \sum_x p_X(x) Q_{\hat{X}|X}(\hat{x}|x)$.

Generation of codebook. Let $R_1 = I(X;\hat{X}) + \epsilon$. Generate 2^{nR_1} i.i.d codewords $\hat{X}^n(s) \sim \prod_{i=1}^n q_{\hat{X}}(\hat{x}_i)$, and index them by $s \in \{1, 2, ..., 2^{nR_1}\}$.

Let $R_2 = I(X;\hat{X}) - I(Y;\hat{X}) + 5\epsilon$. Randomly assign the indices $s \in \{1, 2, ..., 2^{nR_1}\}$ to one of 2^{nR_2} bins using a uniform distribution over the bins. Let $B(i)$ denote the indices assigned to bin i. There are approximately $2^{n(R_1 - R_2)}$ indices in each bin.

Encoding. Given a source sequence x^n, the encoder looks for a codeword $\hat{x}^n(s)$ such that $(x^n, \hat{x}^n(s)) \in A_\epsilon^{*(n)}$. If there is no such $\hat{x}^n$, the encoder sets $s = 1$. If there is more than one such s, the encoder uses the lowest s. The encoder sends the index of the bin in which s belongs.

Decoding. The decoder looks for a $\hat{x}^n$ such that $s \in B(i)$ and $(\hat{x}^n(s), y^n) \in A_\epsilon^{*(n)}$. If it finds a unique s, it then gets the decoder output $\hat{x}_i = \hat{x}(s)$. If it doesn't find any such s or more than one such s, it sets $\hat{x}^n$ as an arbitrary sequence in $\hat{\mathcal{X}}^n$.

Analysis of the probability of error.
1. The pair $(x^n, y^n) \notin A_\epsilon^{*(n)}$. The probability of this event is small for large enough n by the weak law of large numbers.

2. The sequence x^n is typical, but there does not exit an s such that $(x^n, \hat{x}^n(s)) \in A_\epsilon^{*(n)}$. The probability of this event is small if $R_1 > I(X; \hat{X})$.

3. The pair of sequences $(x^n, \hat{x}^n(s)) \in A_\epsilon^{*(n)}$ but $(\hat{x}^n(s), y^n) \notin A_\epsilon^{*(n)}$. By the Markov Lemma (See Lemma 4), the probability of this event is small if n is large enough.

4. There exists another s' with the same bin index such that $(\hat{x}^n(s'), y^n) \in A_\epsilon^{*(n)}$. The probability of this event is:

$$Pr(\exists s' \in B(i) : (\hat{x}^n(s'), y^n) \in A_\epsilon^{*(n)}) \leq 2^{R_1 - R_2} 2^{-n(I(\hat{X};Y) - 3\epsilon)} \tag{3.6}$$

which goes to 0 since $R_1 - R_2 < I(\hat{X}; Y) - 3\epsilon$.

3.2 The Extended Blahut-Arimoto Algorithm

In this section, we extend the well-known Blahut-Arimoto algorithm [2], [13] to calculate the rate distortion function $\hat{R}_{WZ}(D)$ for any memoryless pair (X, Y) and any distortion measure d. Our extension is similar to that used to calculate $R_{WZ}(D)$ in [9], which has a more complicated objective function. Specifically, our extended Blahut-Arimoto algorithm is to compute $\min_{Q_{\hat{X}|X}: Ed(\hat{X};X) \leq D} I(X; \hat{X}|Y)$, while the Blahut-Arimoto algorithm is to calculate $\min_{Q_{\hat{X}|X}: Ed(\hat{X};X) \leq D} I(X; \hat{X})$. This extended algorithm serves two purposes in this thesis: first it allows us to study the rate performance of $\hat{R}_{WZ}(D)$; and second it provides guidelines for designing practical Wyner-Ziv coding. Before describing the algorithm and showing its convergence, some properties are given.

3.2.1 Properties of $\hat{R}_{WZ}(D)$

Proposition 1 *$\hat{R}_{WZ}(D) = \min_{Q_{\hat{X}|X}: Ed(\hat{X};X)\leq D} I(X;\hat{X}|Y)$ is a non-increasing continuous convex function of $D \geq 0$.*

Proof of Proposition 1: Since $Y \to X \to \hat{X}$ forms a Markov chain,

$$\hat{R}_{WZ}(D) = \min_{Q_{\hat{X}|X}: Ed(\hat{X};X)\leq D} [I(X;\hat{X}) - I(Y;\hat{X})] = \min_{Q_{\hat{X}|X}: Ed(\hat{X};X)\leq D} I(X;\hat{X}|Y) \tag{3.7}$$

The monotonicity of $\hat{R}_{WZ}(D)$ follows immediately from the fact that the domain of minimization in the definition of $\hat{R}_{WZ}(D)$ increases with D. Thus, $\hat{R}_{WZ}(D)$ is non-increasing of D.

Consider two rate distortion pairs $(\hat{R}_{WZ}(D_1), D_1)$ and $(\hat{R}_{WZ}(D_2), D_2)$ which lie on the rate-distortion curve given by $\hat{R}_{WZ}(D)$. Let the joint distributions that achieve these pairs be $p_1(x,y,\hat{x}) = p(y)p(x|y)Q_1(\hat{x}|x)$ and $p_2(x,y,\hat{x}) = p(y)p(x|y)Q_2(\hat{x}|x)$, respectively.

Let the conditional probabilities Q_1 and Q_2 achieve the points $(Q_1, I_{Q_1}(X;\hat{X}|Y))$ and $(Q_2, I_{Q_2}(X;\hat{X}|Y))$. Consider the distribution $Q_\lambda = \lambda Q_1 + (1-\lambda)Q_2$ that achieve the points $(Q_\lambda, I_{Q_\lambda}(X;\hat{X}|Y))$. Since the distortion

$$D(Q) = \sum_x \sum_y \sum_{\hat{x}} p(y)p(x|y)Q(\hat{x}|x)d(x,\hat{x}) \tag{3.8}$$

is a linear function of the distribution Q, we have $D(Q_\lambda) = \lambda D_1 + (1-\lambda)D_2$. Next, we show that $I(X;\hat{X}|Y)$ is a convex function of the conditional distribution $Q(\hat{x}|x)$.

$$\begin{aligned}
I(X;\hat{X}|Y) &= \sum_x \sum_y \sum_{\hat{x}} p(y)p(x|y)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{\sum_x p(x|y)Q(\hat{x}|x)} \\
&= \sum_x \sum_y p(y)p(x|y) \sum_{\hat{x}} Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{\sum_x p(x|y)Q(\hat{x}|x)} \\
&= \sum_x \sum_y p(y)p(x|y) D(Q(\hat{x}|x)|| \sum_x p(x|y)Q(\hat{x}|x)) \\
&= \sum_x \sum_y p(y)p(x|y) D(Q(\hat{x}|x)||p(\hat{x}|y))
\end{aligned} \tag{3.9}$$

Since the relative entropy $D(Q(\hat{x}|x)||p(\hat{x}|y))$ is a convex function of $(Q(\hat{x}|x), p(\hat{x}|y))$ (Theorem 2.7.2, [2]), and $I(X;\hat{X}|Y)$ is a linear function of $D(p||q)$, it follows that $I(X;\hat{X}|Y)$ is a convex function of the condition distribution Q.

Because of the convexity of $I(X;\hat{X}|Y)$, we have,

$$I_{Q_\lambda}(X;\hat{X}|Y) \leq \lambda I_{Q_1}(X;\hat{X}|Y) + (1-\lambda) I_{Q_2}(X;\hat{X}|Y) \tag{3.10}$$

Hence, by the definition of $\hat{R}_{WZ}(D)$,

$$\begin{aligned} \hat{R}_{WZ}(D_\lambda) &\leq I_{Q_\lambda}(X;\hat{X}|Y) \\ &\leq \lambda I_{Q_1}(X;\hat{X}|Y) + (1-\lambda) I_{Q_2}(X;\hat{X}|Y) \\ &= \lambda \hat{R}_{WZ}(D_1) + (1-\lambda)\hat{R}_{WZ}(D_2) \end{aligned} \tag{3.11}$$

where $D_\lambda = \lambda D_1 + (1-\lambda) D_2$. This proves that $\hat{R}_{WZ}(D)$ is a convex function of D.

Proposition 2

$$\hat{R}_{WZ}(D) = sD + \min_{Q_{\hat{X}|X}} [I(X;\hat{X}) - I(Y;\hat{X}) - sEd(X,\hat{X})] \tag{3.12}$$

for $s \leq 0$, where $D = \sum_x \sum_{\hat{x}} p_X(x) Q^*_{\hat{X}|X}(\hat{x}|x) d(x,\hat{x})$ and $Q^*_{\hat{X}|X}(\hat{x}|x)$ achieves the minimum in (3.12).

Proposition 3 *For $s \leq 0$ and two probability distributions $Q_{\hat{X}|X}, p_{\hat{X}|Y} > 0$, define*

$$\begin{aligned} F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) = \sum_x \sum_y \sum_{\hat{x}} p_X(x) p_{Y|X}(y|x) Q_{\hat{X}|X}(\hat{x}|x) \log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{p_{\hat{X}|Y}(\hat{x}|y)} \\ -s \sum_x \sum_{\hat{x}} p_X(x) Q_{\hat{X}|X}(\hat{x}|x) d(x,\hat{x}) \end{aligned} \tag{3.13}$$

Then (a),

$$\hat{R}_{WZ}(D) = sD + \min_{Q_{\hat{X}|X}, p_{\hat{X}|Y}} F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) \tag{3.14}$$

where $D = \sum_x \sum_{\hat{x}} p_X(x) Q^*_{\hat{X}|X}(\hat{x}|x) d(x,\hat{x})$ and $Q^*_{\hat{X}|X}(\hat{x}|x)$ achieves the minimum in (3.14).

(b), For fixed $Q_{\hat{X}|X}(\hat{x}|x)$, the optimal probability distribution $p^*_{\hat{X}|Y}(\hat{x}|y)$ to minimize $F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$ is given by

$$p^*_{\hat{X}|Y}(\hat{x}|y) = \sum_x p_{X|Y}(x|y) Q_{\hat{X}|X}(\hat{x}|x) \tag{3.15}$$

(c), For fixed $p_{\hat{X}|Y}(\hat{x}|y)$, the optimal probability distribution $Q^*_{\hat{X}|X}(\hat{x}|x)$ to minimize $F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$ is given by

$$Q^*_{\hat{X}|X}(\hat{x}|x) = \frac{g(x,\hat{x}) e^{sd(x,\hat{x})}}{\sum_{\hat{x'}} g(x,\hat{x'}) e^{sd(x,\hat{x'})}} \tag{3.16}$$

where, $g(x,\hat{x}) = e^{\sum_y p_{Y|X}(y|x) \log p_{\hat{X}|Y}(\hat{x}|y)}$

Proof of Proposition 3:(a), For a given probability distribution $Q_{\hat{X}|X}(\hat{x}|x)$, let

$$p^*_{\hat{X}|Y}(\hat{x}|y) = \sum_x p_{X|Y}(x|y) Q_{\hat{X}|X}(\hat{x}|x) \tag{3.17}$$

Then, for any probability $p_{\hat{X}|Y}(\hat{x}|y)$,

$$\begin{aligned}
& F_s(Q_{\hat{X}|X}, p^*_{\hat{X}|Y}) - F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) \\
= & \sum_x \sum_y \sum_{\hat{x}} p_X(x) p_{Y|X}(y|x) Q_{\hat{X}|X}(\hat{x}|x) [\log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{p^*_{\hat{X}|Y}(\hat{x}|y)} - \log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{p_{\hat{X}|Y}(\hat{x}|y)}] \\
= & -\sum_y p_Y(y) \sum_x \sum_{\hat{x}} p_{X|Y}(x|y) Q_{\hat{X}|X}(\hat{x}|x) \log \frac{p^*_{\hat{X}|Y}(\hat{x}|y)}{p_{\hat{X}|Y}(\hat{x}|y)} \\
= & -\sum_y p_Y(y) D(p^*_{\hat{X}|Y}(\hat{x}|y) || p_{\hat{X}|Y}(\hat{x}|y)) \\
\leq & \ 0
\end{aligned} \tag{3.18}$$

So by proposition 2,

$$\hat{R}_{WZ}(D) = sD + \min_{Q_{\hat{X}|X}, p_{\hat{X}|Y}} F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) \tag{3.19}$$

(b), is obvious from (a);

(c), Using the Lagrange multipliers $\nu(x)$, we define $G_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$ as follows:

$$G_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) = F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) + \sum_x \nu(x) \sum_{\hat{x}} Q_{\hat{X}|X}(\hat{x}|x) \tag{3.20}$$

Now note that $G_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$ is a convex function of $Q_{\hat{X}|X}(\hat{x}|x)$. For fixed $p_{\hat{X}|Y}(\hat{x}|y)$, one has

$$\begin{aligned}
\frac{\partial G_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})}{\partial Q_{\hat{X}|X}(\hat{x}|x)} &= \sum_y \frac{\partial(\sum_x \sum_{\hat{x}} p_X(x) p_{Y|X}(y|x) Q_{\hat{X}|X}(\hat{x}|x) \log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{p_{\hat{X}|Y}(\hat{x}|y)})}{\partial Q_{\hat{X}|X}(\hat{x}|x)} \\
&- s \sum_x \sum_{\hat{x}} \frac{\partial p_X(x) Q_{\hat{X}|X}(\hat{x}|x) d(x,\hat{x})}{\partial Q_{\hat{X}|X}(\hat{x}|x)} + \frac{\partial \sum_x \nu(x) \sum_{\hat{x}} Q_{\hat{X}|X}(\hat{x}|x)}{\partial Q_{\hat{X}|X}(\hat{x}|x)} \\
&= \frac{\partial}{\partial Q_{\hat{X}|X}(\hat{x}|x)} [\sum_y p_Y(y) \sum_x \sum_{\hat{x}} p_{X|Y}(x|y) Q_{\hat{X}|X}(\hat{x}|x) \log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{p_{\hat{X}|Y}(\hat{x}|y) e^{sd(x,\hat{x})}} \\
&+ \sum_x \nu(x) \sum_{\hat{x}} Q_{\hat{X}|X}(\hat{x}|x)] \\
&= \frac{\partial}{\partial Q_{\hat{X}|X}(\hat{x}|x)} [\sum_x \sum_{\hat{x}} p_X(x) Q_{\hat{X}|X}(\hat{x}|x) (\log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{e^{sd(x,\hat{x})}} \\
&- \sum_y p_{Y|X}(y|x) \log p_{\hat{X}|Y}(\hat{x}|y)) + \sum_x \nu(x) \sum_{\hat{x}} Q_{\hat{X}|X}(\hat{x}|x)] \\
&= \frac{\partial}{\partial Q_{\hat{X}|X}(\hat{x}|x)} [\sum_x \sum_{\hat{x}} p_X(x) Q_{\hat{X}|X}(\hat{x}|x) \log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{e^{\sum_y p_{Y|X}(y|x) \log p_{\hat{X}|Y}(\hat{x}|y) + sd(x,\hat{x})}} \\
&+ \sum_x \nu(x) \sum_{\hat{x}} Q_{\hat{X}|X}(\hat{x}|x)] \\
&= p_X(x) \log \frac{Q_{\hat{X}|X}(\hat{x}|x)}{e^{\sum_y p_{Y|X}(y|x) \log p_{\hat{X}|Y}(\hat{x}|y) + sd(x,\hat{x})}} + \nu(x)
\end{aligned} \tag{3.21}$$

The derivative is equal to zero if the minimum is achieved. Assume $p_X(x) > 0$ for all x. By the Karush-Kuhn-Tucker (KKT) conditions and $\sum_{\hat{x}} Q_{\hat{X}|X}(\hat{x}|x) = 1$, if $Q^*_{\hat{X}|X}(\hat{x}|x) > 0$, then it is not hard to get,

$$Q^*_{\hat{X}|X}(\hat{x}|x) = \frac{g(x,\hat{x}) e^{sd(x,\hat{x})}}{\sum_{\hat{x'}} g(x,\hat{x'}) e^{sd(x,\hat{x'})}} \tag{3.22}$$

where, $g(x,\hat{x}) = e^{\sum_y p_{Y|X}(y|x)\log p_{\hat{X}|Y}(\hat{x}|y)}$, which is optimal to minimize $F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$ for fixed $p_{\hat{X}|Y}(\hat{x}|y)$.

Proposition 4 *Probability distributions $Q_{\hat{X}|X}(\hat{x}|x)$ and $p_{\hat{X}|Y}(\hat{x}|y)$ achieve the rate-distortion function $\hat{R}_{WZ}(D)$ in (3.14) if and only if they satisfy*

$$p_{\hat{X}|Y}(\hat{x}|y) = \sum_x p_{X|Y}(x|y)Q_{\hat{X}|X}(\hat{x}|x), Q_{\hat{X}|X}(\hat{x}|x) = \frac{g(x,\hat{x})e^{sd(x,\hat{x})}}{\sum_{\hat{x'}} g(x,\hat{x'})e^{sd(x,\hat{x'})}} \tag{3.23}$$

where, $g(x,\hat{x}) = e^{\sum_y p_{Y|X}(y|x)\log p_{\hat{X}|Y}(\hat{x}|y)}$.

3.2.2 The Extended Blahut-Arimoto Algorithm

The extended Blahut-Arimoto algorithm, as the standard Blahut-Arimoto algorithm, is an iterative algorithm, and works as follows. Starting with an initial guess $p^{(0)}_{\hat{X}|Y}$, we first find $Q^{(1)}_{\hat{X}|X}$ that minimizes $F_s(Q_{\hat{X}|X}, p^{(0)}_{\hat{X}|Y})$. Then we fix $Q^{(1)}_{\hat{X}|X}$, and find $p^{(1)}_{\hat{X}|Y}$ that minimizes $F_s(Q^{(1)}_{\hat{X}|X}, p_{\hat{X}|Y})$. Iteratively doing so, we then have a sequence of distribution pairs $\{(Q^{(i)}_{\hat{X}|X}, p^{(i-1)}_{\hat{X}|Y}); i \geq 1\}$ such that

$$Q^{(i)}_{\hat{X}|X}(\hat{x}|x) \;=\; \frac{g(x,\hat{x})e^{sd(x,\hat{x})}}{\sum_{\hat{x'}\in\hat{\mathcal{X}}} g(x,\hat{x'})e^{sd(x,\hat{x'})}} \tag{3.24}$$

where $g(x,\hat{x}) = e^{\sum_{y\in\mathcal{Y}} p_{Y|X}(y|x)\log p^{(i-1)}_{\hat{X}|Y}(\hat{x}|y)}$, and for any $(\hat{x},y) \in \hat{\mathcal{X}}\times\mathcal{Y}$,

$$p^{(i)}_{\hat{X}|Y}(\hat{x}|y) \;=\; \sum_{x\in\mathcal{X}} p_{X|Y}(x|y)Q^{(i)}_{\hat{X}|X}(\hat{x}|x). \tag{3.25}$$

Using an idea similar to [11], the following theorem shows that the sequence $\{(Q^{(i)}_{\hat{X}|X}, p^{(i-1)}_{\hat{X}|Y}); i \geq 1\}$ obtained by our extended Blahut-Arimoto algorithm converges to a pair of distributions that achieves $F_s \stackrel{\Delta}{=} \inf_{Q_{\hat{X}|X}, p_{\hat{X}|Y}} F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$. Here, the infimum is taken over all possible pairs of transition probability distributions from $\mathcal{X}$ to $\hat{\mathcal{X}}$ and from $\mathcal{Y}$ to $\hat{\mathcal{X}}$ (Note that since $\hat{\mathcal{X}}$ and $\mathcal{Y}$ are assumed to be finite, the infimum is achievable). For brevity, let $Q_{\hat{X}|X}(p_{\hat{X}|Y})$ denote the transition probability function from $\mathcal{X}$ to $\hat{\mathcal{X}}$ obtained from $p_{\hat{X}|Y}$ through (3.24).

Theorem 3 *If the side information alphabet and the reproducing alphabet are finite and* $p_Y(y) > 0$ *for any* $y \in \mathcal{Y}$*, there exists a* $p^*_{\hat{X}|Y}$ *such that* $p^{(n)}_{\hat{X}|Y} \to p^*_{\hat{X}|Y}, Q^{(n)}_{\hat{X}|X} \to Q^*_{\hat{X}|X} = Q_{\hat{X}|Y}(p^*_{\hat{X}|Y})$*, and* $F_s(Q^*_{\hat{X}|X}, p^*_{\hat{X}|Y}) = F_s$ *as* $n \to \infty$*.*

Proof of Theorem 3: From the algorithm, starting from an arbitrary $p^{(0)}_{\hat{X}|Y} > 0$, set recursively generate $Q^{(n)}_{\hat{X}|X} = Q(p^{(n-1)}_{\hat{X}|Y})$, $p^{(n)}_{\hat{X}|Y} = p(Q^{(n)}_{\hat{X}|X}), n = 1, 2, 3....$ Then by (3.18), clearly,

$$F_s(Q^{(1)}_{\hat{X}|X}, p^{(0)}_{\hat{X}|Y}) \geq F_s(Q^{(1)}_{\hat{X}|X}, p^{(1)}_{\hat{X}|Y}) \geq F_s(Q^{(2)}_{\hat{X}|X}, p^{(1)}_{\hat{X}|Y}) \geq ... \tag{3.26}$$

Since $Y \to X \to \hat{X}$ forms a Markov chain, then $Q_{\hat{X}|X} = Q_{\hat{X}|X,Y}$. For arbitrary $Q_{\hat{X}|X}, p_{\hat{X}|Y}$, consider the "backward probability",

$$L_{X|\hat{X},Y}(x|\hat{x}, y) = \frac{Q_{\hat{X}|X,Y}(\hat{x}|x, y)p_{X|Y}(x|y)}{p_{\hat{X}|Y}(\hat{x}|y)} \tag{3.27}$$

$$= \frac{p_{X|Y}(x|y)Q_{\hat{X}|X}(\hat{x}|x)}{p_{\hat{X}|Y}(\hat{x}|y)} \tag{3.28}$$

and let $L_{y\hat{x}}$ denote the corresponding distribution for fixed y and $\hat{x}$.

From the easily checked identity

$$F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) + \sum_y p_Y(y) \sum_{\hat{x}} p_{\hat{X}|Y}(\hat{x}|y) D(L_{y\hat{x}}(x|y, \hat{x})||L^{(n)}_{y\hat{x}}(x|y, \hat{x}))$$

$$-F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) = \sum_y p_Y(y) \sum_{\hat{x}} p_{\hat{X}|Y}(\hat{x}|y) \log \frac{p^{(n)}_{\hat{X}|Y}(\hat{x}|y)}{p^{(n-1)}_{\hat{X}|Y}(\hat{x}|y)} \tag{3.29}$$

where $L^{(n)}_{X|\hat{X},Y}(x|y, \hat{x})$ is defined as:

$$L^{(n)}_{X|\hat{X},Y}(x|y, \hat{x}) = \frac{p_{X|Y}(x|y)g^{(n-1)}(x, \hat{x})e^{sd(x,\hat{x})}}{p^{(n)}_{\hat{X}|Y}(\hat{x}|y) \sum_{\hat{x'}} g^{(n-1)}(x, \hat{x'})e^{sd(x,\hat{x'})}}, \tag{3.30}$$

where $g^{(n-1)}(x, \hat{x'}) = e^{\sum_y p_{Y|X}(y|x) \log p^{(n-1)}_{\hat{X}|Y}(\hat{x'}|y)}$.

Supposing

$$\begin{aligned} F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y}) &\leq \lim_{n\to\infty} F_s(Q^{(n)}_{\hat{X}|X}, p^{(n)}_{\hat{X}|Y}) && (3.31)\\ &= \lim_{n\to\infty} F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) && (3.32)\end{aligned}$$

(3.29) implies, for any $N \geq M \geq 1$,

$$\begin{aligned} 0 &\leq \sum_{n=M+1}^{N} [F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) - F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})]\\ &= \sum_{n=M+1}^{N} [\sum_y p(y) \sum_{\hat{x}} p_{\hat{X}|Y}(\hat{x}|y) \log \frac{p^{(n)}_{\hat{X}|Y}(\hat{x}|y)}{p^{(n-1)}_{\hat{X}|Y}(\hat{x}|y)}\\ &- \sum_y p_Y(y) \sum_{\hat{x}} p_{\hat{X}|Y}(\hat{x}|y) D(L_{y\hat{x}}(x|y,\hat{x}) || L^{(n)}_{y\hat{x}}(x|y,\hat{x}))]\\ &\leq \sum_{n=M+1}^{N} \sum_y p(y) \sum_{\hat{x}} p_{\hat{X}|Y}(\hat{x}|y) \log \frac{p^{(n)}_{\hat{X}|Y}(\hat{x}|y)}{p^{(n-1)}_{\hat{X}|Y}(\hat{x}|y)}\\ &= \sum_y p(y) \sum_{\hat{x}} p_{\hat{X}|Y}(\hat{x}|y) \log \frac{p^{(N)}_{\hat{X}|Y}(\hat{x}|y)}{p^{(M)}_{\hat{X}|Y}(\hat{x}|y)}\\ &= \sum_y p(y)[I(p_{\hat{X}|Y}||p^{(M)}_{\hat{X}|Y}) - I(p_{\hat{X}|Y}||p^{(N)}_{\hat{X}|Y})] && (3.33)\end{aligned}$$

If the side information alphabet is finite, this shows, the series $\sum_n [F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) - F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})]$ converges; thus

$$\lim_{n\to\infty} [F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) - F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})] = 0 \quad (3.34)$$

We are saying that,

$$\begin{aligned} \lim_{n\to\infty} [F_s(Q^{(n)}_{\hat{X}|X}, p^{(n)}_{\hat{X}|Y})] &= \lim_{n\to\infty} [F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y})]\\ &= \inf F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})\\ &= F_s && (3.35)\end{aligned}$$

The infinum can be approached by $p_{\hat{X}|Y}$ with $I(p_{\hat{X}|Y}||p^{(1)}_{\hat{X}|Y}) < \infty$. If the reproduction alphabet is finite, the condition $I(p_{\hat{X}|Y}||p^{(1)}_{\hat{X}|Y}) < \infty$ is satisfied; if it is countable, the sufficient condition is

$$d^* = d_{max} = \min_{\hat{x}} \sum_x p(x)d(x,\hat{x}) < \infty \tag{3.36}$$

Pick a convergent subsequence $p^{(n_i)}_{\hat{X}|Y} \to p^*_{\hat{X}|Y}$, say, of $p^{(n)}_{\hat{X}|Y}$. Then, $Q^{(n_i+1)}_{\hat{X}|X} = Q(p^{(n_i)}_{\hat{X}|Y}) \to Q(p^*_{\hat{X}|Y}) = Q^*_{\hat{X}|X}$ and

$$F_s(Q^{(n_i+1)}_{\hat{X}|X}, p^{(n_i)}_{\hat{X}|Y}) \to F_s(Q^*_{\hat{X}|X}, p^*_{\hat{X}|Y}) \tag{3.37}$$

In view of (3.35), we have $F_s(Q^*_{\hat{X}|X}, p^*_{\hat{X}|Y}) = F_s$, thus $p^*_{\hat{X}|Y} = p(Q^*_{\hat{X}|X})$.

Since $F_s(Q^*_{\hat{X}|X}, p^*_{\hat{X}|Y}) = F_s = \inf F_s(Q_{\hat{X}|X}, p_{\hat{X}|Y})$, then for every $F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y})$, we have $F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) \geq F_s(Q^*_{\hat{X}|X}, p^*_{\hat{X}|Y})$.

Apply (3.33) for $Q^*_{\hat{X}|X}$ and $p^*_{\hat{X}|Y}$,

$$\begin{aligned} 0 &\leq \sum_{n=M+1}^{N} [F_s(Q^{(n)}_{\hat{X}|X}, p^{(n-1)}_{\hat{X}|Y}) - F_s(Q^*_{\hat{X}|X}, p^*_{\hat{X}|Y})] \\ &\leq \sum_y p(y)[I(p^*_{\hat{X}|Y}||p^{(M)}_{\hat{X}|Y}) - I(p^*_{\hat{X}|Y}||p^{(N)}_{\hat{X}|Y})] \end{aligned} \tag{3.38}$$

is always true for any $N \geq M \geq 1$. Then we are saying that $I(p^*_{\hat{X}|Y}||p^{(n)}_{\hat{X}|Y})$ is a nonincreasing sequence. By $p^{(n_i)}_{\hat{X}|Y} \to p^*_{\hat{X}|Y}$, we have $p^{(n_i)}_{\hat{X}|Y} \to p^*_{\hat{X}|Y}$ for each y with $p_Y(y) > 0$, which implies $I(p^*_{\hat{X}|Y}||p^{(n_i)}_{\hat{X}|Y}) \to 0$. Hence, $I(p^*_{\hat{X}|Y}||p^{(n)}_{\hat{X}|Y}) \to 0$ which means $I(p^*_{\hat{X}|y}||p^{(n)}_{\hat{X}|y}) \to 0$ for each y, thus $p^{(n)}_{\hat{X}|Y} \to p^*_{\hat{X}|Y}$.

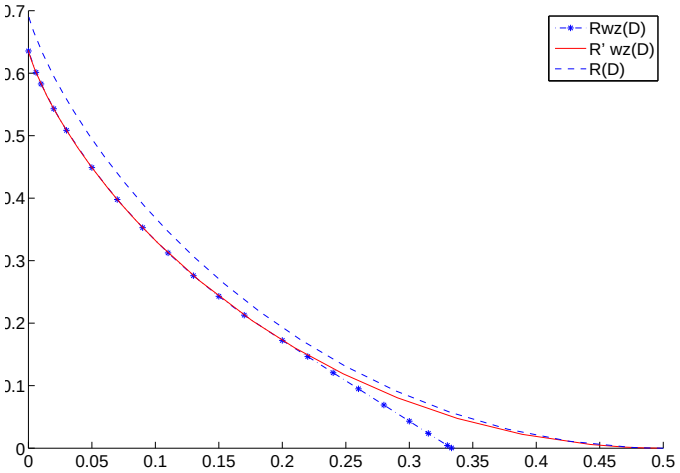

Figure 3.2: Comparison of $R_{WZ}(D)$, $\hat{R}_{WZ}(D)$, and $R(D)$

3.2.3 An Interesting Phenomenon Observed

The extended Blahut-Arimoto algorithm allows us to observe some interesting phenomena. One is naturally derived by computing the rate-distortion function $\hat{R}_{WZ}(D)$. Comparing with the rate performance $R_{WZ}(D)$ in (2.22) and $R(D)$ in (2.19), we see that for the case of unbiased input to a binary symmetric channel from the source X to side information Y with crossover probability $p_0 = \frac{1}{3}$, $R_{WZ}(D) < \hat{R}_{WZ}(D) < R(D)$ (Fig 3.2).

As expected, since the reconstruction function in practical Wyner-Ziv system has been fixed, the rate-performance $\hat{R}_{WZ}(D)$ is not as good as $R_{WZ}(D)$. Furthermore, with the help of side information Y, $\hat{R}_{WZ}(D)$ outperforms $R(D)$. From Fig 3.2, we see that the gap between $\hat{R}_{WZ}(D)$ and $R_{WZ}(D)$ is small. However, the deep insights of the gap should be mathematically discussed, which is referred to the future work in Chapter 5.

Another more interesting phenomenon is that, in most cases, the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$ is different from the random variable $\hat{X}'$ that achieves the classical rate-distortion $R(D)$ in (2.19). Interestingly, there are indeed cases where $\hat{X} = \hat{X}'$. To put

this into perspective, let us look at an example.

Example 1: Suppose that $\mathcal{X} = \mathcal{Y} = \hat{\mathcal{X}} = \{0, 1\}$, and that the Hamming distortion measure is used. We consider two cases.

Case 1: $p_{Y|X}(0|1) = p_{Y|X}(1|0) = 0.25$

Case 2: $p_{Y|X}(0|1) = 0.45, p_{Y|X}(1|0) = 0.05$.

In Case 1, it is observed that regardless the marginal distribution p_X, the transition probability distribution $Q_{\hat{X}'|X}$ achieving $R(D)$ for X always achieves $\hat{R}_{WZ}(D)$ for (X, Y). The above observation, however, does not hold in Case 2. For example, when $p_X(0) = 1/3$, and $D = 0.01$, the optimum $Q_{\hat{X}'|X}$ for $R(D)$ and the optimum $Q_{\hat{X}|X}$ for $\hat{R}_{WZ}(D)$ are as follows:

$$Q_{\hat{X}'|X}(0|1) \sim 0.005, \quad Q_{\hat{X}'|X}(1|0) \sim 0.02.$$
$$Q_{\hat{X}|X}(0|1) \sim 0.0038, \quad Q_{\hat{X}|X}(1|0) \sim 0.0225.$$

To fully understand and characterize this important and rather surprising phenomenon, we are led to the following questions:

Q1: Under what conditions is $\hat{X}$ achieving $\hat{R}_{WZ}(D)$ the same as $\hat{X}'$ achieving $R(D)$? Equivalently, under what conditions should the design of conventional quantization in the case of side information be the same as the case of no side information?

Q2: Under what conditions is $\hat{X}$ achieving $\hat{R}_{WZ}(D)$ different from $\hat{X}'$ achieving $R(D)$? Equivalently, under what conditions should the design of conventional quantization in the case of side information be different from the case of no side information?

Example 1 above seems to suggest that $p_{Y|X}$ being symmetric be the answer to above questions for binary alphabets and Hamming distortion measure. This is indeed proved in the next chapter. Note that due to finite precision in computation, running the extended Blahut-Arimoto algorithm is not a mathematical proof.

Chapter 4

Optimum Conventional Quantization

In this chapter, we settle Question Q1 raised at the end of Chapter 3 for the following cases:

1, $\mathcal{X}$, $\mathcal{Y}$, and $\hat{\mathcal{X}}$ are all binary alphabets, and the distortion measure d is the Hamming distortion measure;

2, $\mathcal{X}$, $\mathcal{Y}$, and $\hat{\mathcal{X}}$ are all discrete non-binary alphabets, where $\hat{X} \in$ the optimum interior set V [12], and the distortion measure d is the Hamming distortion measure;

3, $X \sim \mathcal{N}(0, \sigma_x^2)$, while the side information $Y = X + N$ with X and N independent and $EN = 0$, $EN^2 = {\sigma_N}^2$, and the distortion measure d is the mean-squared error distortion measure.

We also settle Question Q2 completely for the binary case with Hamming distortion measure.

4.1 Binary Case

Throughout this section, we assume that $\mathcal{X} = \mathcal{Y} = \hat{\mathcal{X}} = \{0, 1\}$, and d denotes the Hamming distortion measure. With these assumptions, we settle Question Q1 and Q2 completely in Theorems 4 and 5 [21].

Definition 9 *A binary memoryless channel $p_{Y|X}$ with input X and output Y is said to be symmetric if and only if the input and output relationship can be expressed as $Y = X \oplus N$, where N is independent of X. Throughout this thesis, $\oplus$ denotes modulo-2 addition. The probability of the event $N = 1$ is called the crossover probability of the symmetric channel.*

Theorem 4 *Let X denote a binary random variable with $p_X(0) = p \in (0, 0.5]$. If the channel $p_{Y|X}$ from X to Y is symmetric with crossover probability $0 \leq q \leq 1$, then for any $D \geq 0$, the random variable $\hat{X}'$ that achieves $R(D)$ for X also achieves $\hat{R}_{WZ}(D)$ for (X, Y).*

Proof of Theorem 4: Theorem 4 obviously holds when $D \geq p$ or $D = 0$. Assume now that $D \in (0, p)$. Observe that

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &\geq H(X) - H(X \oplus \hat{X}) \end{aligned} \tag{4.1}$$

It follows from (4.1) that if $I(X;\hat{X})$ were to be minimized subject to a Hamming distortion constraint, one should look for $\hat{X}$ such that $p_{X|\hat{X}}$ is symmetric, if such $\hat{X}$ exists. Interestingly, with $0 < D < p$, one can always find such a random variable $\hat{X}'$ with marginal distribution $q_{\hat{X}'}$

$$q_{\hat{X}'}(0) = 1 - q_{\hat{X}'}(1) = \frac{p - D}{1 - 2D} \tag{4.2}$$

such that the conditional probability distribution $p_{X|\hat{X}'}$ of X given $\hat{X}'$ (also called the test channel) is given by

$$p_{X|\hat{X}'}(x|\hat{x}) = \begin{cases} D & x \neq \hat{x} \\ 1 - D & \text{otherwise} \end{cases}, \tag{4.3}$$

where $x, \hat{x} \in \{0, 1\}$.

Let us look at the case where the side information Y is available only at the decoder. Then

$$\begin{aligned} & I(X;\hat{X}) - I(Y;\hat{X}) \\ = \; & H(X) - H(Y) + [H(Y|\hat{X}) - H(X|\hat{X})]. \end{aligned} \tag{4.4}$$

In view of (4.4), we see that the presence of Y complicates the problem of minimizing $I(X;\hat{X})-I(Y;\hat{X})$ subject to $Ed(X,\hat{X}) \leq D$. More specifically, it is no longer clear that in order to minimize $I(X;\hat{X})-I(Y;\hat{X})$ under the distortion constraint, one should look for a symmetric test channel $p_{X|\hat{X}}$. In the following, we shall argue that the optimum test channel $p_{X|\hat{X}}$ (in the sense of minimizing$I(X;\hat{X}) - I(Y;\hat{X})$ under the constraint $Ed(X,\hat{X}) \leq D$) is not only symmetric but also equal to $p_{X|\hat{X}'}$ in (4.3).

For brevity, let $r_0 = p_{X|\hat{X}}(1|0)$, and $r_1 = p_{X|\hat{X}}(0|1)$. Let $q_{\hat{X}}$ denote the marginal distribution of $\hat{X}$. Then

$$\begin{aligned} & H(Y|\hat{X}) - H(X|\hat{X}) \\ = \; & H(Y \oplus \hat{X}|\hat{X}) - H(X \oplus \hat{X}|\hat{X}) \\ = \; & q_{\hat{X}}(0)[H(r_0 * q) - H(r_0)] + \\ & q_{\hat{X}}(1)[H(r_1 * q) - H(r_1)] \\ \overset{1)}{\geq} \; & H(r * q) - H(r) \\ \overset{2)}{\geq} \; & H(D * q) - H(D), \end{aligned} \tag{4.5}$$

where $r = q_{\hat{X}}(0)r_0 + q_{\hat{X}}(1)r_1$, and for two real numbers a, b, $a * b \overset{\Delta}{=} a(1-b) + (1-a)b$. In the above, the inequality 1) follows from Lemma 5 after Theorem 5; and the inequality 2) is due to the constraint $r = Ed(X,\hat{X}) \leq D$ and Lemma 5. Combining (4.4) and (4.5), we have

$$\begin{aligned} & I(X;\hat{X}) - I(Y;\hat{X}) \\ \geq \; & H(X) - H(Y) + [H(D * q) - H(D)]. \end{aligned} \tag{4.6}$$

It is thus clear that the right-hand-side of (4.6) is indeed achievable with the random variable $\hat{X}'$ satisfying (4.2) and (4.3). This completes the proof of Theorem 4.

Theorem 5 *Let X denote a binary random variable with $p_X(0) = p \in (0, 0.5]$. If the channel $p_{Y|X}$ from X to Y is asymmetric with $p_{Y|X}(1|0) \neq p_{Y|X}(0|1)$ and $p_{Y|X}(1|0) + p_{Y|X}(0|1) \neq 1$, then for $0 < D < p$, the random variable $\hat{X}'$ that achieves $R(D)$ for X cannot achieve $\hat{R}_{WZ}(D)$ for (X,Y).*

Proof of Theorem 5: For brevity, denote $p_{Y|X}(1|0)$ and $p_{Y|X}(0|1)$ by q_0 and q_1, respectively. Without loss of generality, we assume that $q_0 < q_1$. Note that from our assumptions, $q_0 + q_1 \neq 1$. For $0 < D < p$, let us assume that $\hat{X}'$ is a random variable achieving $R(D)$ for X, i.e., $\hat{X}'$ satisfies (4.2) and (4.3). We prove Theorem 5 by contradiction.

Suppose that $\hat{X}'$ achieves $\hat{R}_{WZ}(D)$. Let $Q_{\hat{X}'|X}$ and $p_{\hat{X}'|Y}$ denote the conditional probability distribution of $\hat{X}'$ given X and Y, respectively. In view of (3.13), we see that $F_s(Q_{\hat{X}'|X}, p_{\hat{X}'|Y}) = F_s$. For fixed $q_{\hat{X}'}$, take the first derivative of $F_s(p_{\hat{X}'|X}, p_{\hat{X}'|Y})$ over $p_{X|\hat{X}'}$, and set it to be 0. This implies that

$$\begin{aligned} & \sum_{y\in\{0,1\}} p_{Y|X}(y|1) \log \frac{p_{X|\hat{X}'}(1|0)}{p_{Y|\hat{X}'}(y|0)} \\ = & \sum_{y\in\{0,1\}} p_{Y|X}(y|0) \log \frac{p_{X|\hat{X}'}(0|1)}{p_{Y|\hat{X}'}(y|1)}, \end{aligned} \tag{4.7}$$

where $p_{X|\hat{X}'}$ and $p_{Y|\hat{X}'}$ denote the conditional probability distribution of X given $\hat{X}'$ and Y given $\hat{X}'$, respectively.

Let $\delta = q_1 - q_0$. Then the left-hand-side of (4.7) can be written as

$$\begin{aligned} & (q_0 + \delta) \log \frac{D}{1 - D * q_0 + D\delta} + \\ & (1 - q_0 - \delta) \log \frac{D}{D * q_0 - D\delta}. \end{aligned} \tag{4.8}$$

Similarly, the right-hand-side of (4.7) can be written as

$$\begin{aligned} & (1 - q_0) \log \frac{D}{D * q_0 + (1 - D)\delta} + \\ & q_0 \log \frac{D}{1 - D * q_0 - (1 - D)\delta}. \end{aligned} \tag{4.9}$$

Let $F(D, \delta)$ denote the result of subtracting (4.9) from (4.8). It is easy to verify that

$$F(0.5, \delta) = (1 - 2q_0 - \delta) \log \frac{1 + \delta}{1 - \delta} \tag{4.10}$$

When $q_0 + q_1 < 1$, since $\delta > 0$, we have that for any $D \in (0, p)$,

$$\frac{\partial F(D, \delta)}{\partial D} < 0, \text{ and } F(D, \delta) > F(0.5, \delta) > 0. \tag{4.11}$$

Similarly, when $q_0 + q_1 > 1$, we have that for any $D \in (0, p)$

$$\frac{\partial F(D,\delta)}{\partial D} > 0, \text{ and } F(D,\delta) < F(0.5,\delta) < 0 \tag{4.12}$$

Remark 1 *Proof of (4.11):*
Denote the sum in (4.8) by $F_1(D,\delta)$, and the sum in (4.9) by $F_2(D,\delta)$, so $F(D,\delta) = F_1(D,\delta) - F_2(D,\delta)$. We have

$$\frac{\partial F(D,\delta)}{\partial D} = \frac{\partial F_1(D,\delta)}{\partial D} - \frac{\partial F_2(D,\delta)}{\partial D} \tag{4.13}$$

$$\begin{aligned} F_1(D,\delta) = (q_0+\delta)(\log D - \log(1 - D * q_0 + D\delta)) + \\ (1 - q_0 - \delta)(\log D - \log(D * q_0 - D\delta)) \end{aligned} \tag{4.14}$$

$$\begin{aligned} F_2(D,\delta) = (1-q_0)(\log D - \log(D * q_0 + (1-D)\delta)) + \\ q_0(\log D - \log(1 - D * q_0 - (1-D)\delta)) \end{aligned} \tag{4.15}$$

And we can easily derive,

$$\begin{aligned} \frac{\partial F_1(D,\delta)}{\partial D} &= \frac{1}{D} - \frac{(q_0+\delta)(-1+2q_0+\delta)}{1 - D * q_0 + D\delta} \\ &\quad - \frac{1 - q_0 - \delta)(1 - 2q_0 - \delta)}{D * q_0 - D\delta} \\ &= \frac{1}{D} - \frac{(1 - \delta - 2q_0)(1 - q_0 - D * q_0 - (1-D)\delta)}{(D * q_0 - D\delta)(1 - D * q_0 + D\delta)} \end{aligned} \tag{4.16}$$

$$\begin{aligned} \frac{\partial F_2(D,\delta)}{\partial D} &= \frac{1}{D} - \frac{q_0(-1+2q_0+\delta)}{1 - D * q_0 - (1-D)\delta} \\ &\quad - \frac{1 - q_0)(1 - 2q_0 - \delta)}{D * q_0 + (1-D)\delta} \\ &= \frac{1}{D} - \frac{(1 - \delta - 2q_0)(1 - q_0 - D * q_0 - (1-D)\delta)}{(D * q_0 + (1-D)\delta)(1 - D * q_0 - (1-D)\delta)} \end{aligned} \tag{4.17}$$

Observing the right-hand-side of (4.16)and (4.17), the only difference is the denominator of the second item. Let $A = D*q_0 - D\delta$, $B = D*q_0 + (1-D)\delta$, and $B(1-B) - A(1-A)$

denote the difference of the two denominators. We can easily check that $B = A + \delta$, *and* $B(1-B) - A(1-A) = \delta(1-2A-\delta)$. *With* $q_0 + q_1 < 1$, *we have*

$$\begin{aligned} 2A+\delta &= 2(D*q_0 - D\delta) + \delta \\ &= 2[D + (1-2D)q_0 - D\delta] + \delta \\ &= 2[D + (1-2D)q_0] + \delta(1-2D) \\ &= 2[D + (1-2D)q_0] + (q_1 - q_0)(1-2D) \\ &= 2D + (q_0+q_1)(1-2D) \\ &< 2D + 1 - 2D \\ &= 1 \end{aligned} \tag{4.18}$$

For fixed $\delta > 0$, *and* $q_0 + q_1 < 1$, *(4.18) implies that* $B(1-B) - A(1-A) > 0$, *and further implies*

$$\frac{\partial F_1(D,\delta)}{\partial D} < \frac{\partial F_2(D,\delta)}{\partial D} \tag{4.19}$$

(4.13) and (4.19) together imply that for any $\delta > 0$, $q_0 + q_1 < 1$,

$$\frac{\partial F(D,\delta)}{\partial D} < 0, \quad 0 < D < p \tag{4.20}$$

Lemma 5 *Let* r, q *be two real numbers such that* $0 \le r, q \le 1$. *Then* $H(r*q) - H(r)$ *is convex with respect to* r. *Furthermore,* $H(r*q) - H(r)$ *is a monotonically decreasing function of* r *when* $r \le 0.5$.

Proof of Lemma 5: For brevity, we assume natural logarithm in this proof. Taking the first order derivative of $H(r*q) - H(r)$ with respect to r, we get

$$\frac{d[H(r*q) - H(r)]}{dr} = (1-2q)\log\frac{1-r*q}{r*q} - \log\frac{1-r}{r}. \tag{4.21}$$

The convexity of $H(r*q) - H(r)$ with respect to r can then be verified by checking that

$$\frac{d^2[H(r*q) - H(r)]}{dr^2} = -\frac{(1-2q)^2}{(1-r*q)(r*q)} + \frac{1}{r(1-r)} \ge 0.$$

Observe that $|0.5 - r*q| = |0.5-r||1-2q| \le |0.5-r|$. This, together with (4.21), implies that $H(r*q) - H(r)$ is a monotonically decreasing function of r when $r \le 0.5$. This concludes the proof of Lemma 5.

4.2 Non-Binary Case

Throughout this section, we assume that $\mathcal{X} = \mathcal{Y} = \hat{\mathcal{X}} = \{0, 1, 2, ..., n-1\}, n \geq 2$, and d denotes the Hamming distortion measure. With these assumptions, we determine the sufficient condition to answer Question Q1 in Theorems 6.

Definition 10 *For a given matrix P with the following structure,*

$$P = \begin{pmatrix} p_0 & p_1 & ... & p_1 \\ p_1 & p_0 & ... & p_1 \\ . & & & \\ . & & & \\ p_1 & p_1 & ... & p_0 \end{pmatrix} \tag{4.22}$$

let P_p denote the set of all matrices obtained from P by finite numbers of permutation over rows and columns, and the sum of each row and each column is 1.

Theorem 6 *Let X denote a random variable taking values from $\{0, 1, 2, ..., n-1\}, n \geq 2$. We index the source letters in order of decreasing probability, $P_0 \geq P_1 \geq ... \geq P_{n-1}$, then for any $D \in [0, (n-1)P_{n-1}]$, if the channel $p_{Y|X}$ from X to Y is in the set P_p, the random variable $\hat{X}'$ that achieves $R(D)$ for X also achieves $\hat{R}_{WZ}(D)$ for (X, Y).*

Proof of Theorem 6: For brevity, let $q_{ij} = q_{X|\hat{X}}(i|j), Ed(X, \hat{X}) = d$. Observe the classic rate-distortion function,

$$\begin{aligned} R(D) = I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &\geq H(X) - H(X - \hat{X}) \end{aligned} \tag{4.23}$$

For Hamming distortion measure, interestingly, for any $D \in [0, (n-1)P_{n-1}]$, which implies that $\hat{X} \in$ the optimum interior set V [12], one can always find such a random variable $\hat{X}'$ with the conditional probability distribution $p_{X|\hat{X}'}$ of X given $\hat{X}'$ (test channel) given by,

$$Q = \begin{pmatrix} q_{00} & q_{10} & ... & q_{(n-1)0} \\ q_{01} & q_{11} & ... & q_{(n-1)1} \\ . & & & \\ . & & & \\ q_{0(n-1)} & q_{1(n-1)} & ... & q_{(n-1)(n-1)} \end{pmatrix} = \begin{pmatrix} 1-D & \frac{D}{n-1} & ... & \frac{D}{n-1} \\ \frac{D}{n-1} & 1-D & ... & \frac{D}{n-1} \\ . & & & \\ . & & & \\ \frac{D}{n-1} & \frac{D}{n-1} & ... & 1-D \end{pmatrix} \tag{4.24}$$

Let us look at the case where the side information available only at the decoder. Then

$$\begin{aligned} & I(X;\hat{X}) - I(Y;\hat{X}) \\ = \;& H(X) - H(Y) + [H(Y|\hat{X}) - H(X|\hat{X})]. \end{aligned} \tag{4.25}$$

In the following, we would argue that if the channel $p_{Y|X}$ from X to Y is in the set P_p, denoted by P for brevity, the optimum test channel $p_{X|\hat{X}}$ (in the sense of minimizing$I(X;\hat{X}) - I(Y;\hat{X})$ under the constraint $Ed(X,\hat{X}) \leq D$) is equal to $p_{X|\hat{X}'}$ in (4.24).

Let $q_{\hat{X}}$ denote the marginal distribution of $\hat{X}$. Then,

$$\begin{aligned} & H(Y|\hat{X}) - H(X|\hat{X}) \\ = \;& \sum_{i=0}^{n-1} q_{\hat{X}}(i)[H(\begin{pmatrix} q_{0i} & q_{1i} & ... & q_{(n-1)i} \end{pmatrix} P) - H\begin{pmatrix} q_{0i} & q_{1i} & ... & q_{(n-1)i} \end{pmatrix}] \\ \overset{1)}{=} \;& \sum_{i=0}^{n-1} q_{\hat{X}}(i)[H(\begin{pmatrix} q_{ii} & q_{(i+1)i} & ... & q_{(i-1)i} \end{pmatrix} P) - H\begin{pmatrix} q_{ii} & q_{(i+1)i} & ... & q_{(i-1)i} \end{pmatrix}] \\ \overset{2)}{\geq} \;& H(\begin{pmatrix} q_0 & q_1 & ... & q_{n-1} \end{pmatrix} P) - H\begin{pmatrix} q_0 & q_1 & ... & q_{n-1} \end{pmatrix} \\ \overset{3)}{\geq} \;& H\left(\begin{pmatrix} 1-d & \frac{d}{n-1} & ... & \frac{d}{n-1} \end{pmatrix} P\right) - H\begin{pmatrix} 1-d & \frac{d}{n-1} & ... & \frac{d}{n-1} \end{pmatrix} \\ \overset{4)}{\geq} \;& H(\begin{pmatrix} 1-D & \frac{D}{n-1} & ... & \frac{D}{n-1} \end{pmatrix} P) - H\begin{pmatrix} 1-D & \frac{D}{n-1} & ... & \frac{D}{n-1} \end{pmatrix} \end{aligned} \tag{4.26}$$

where $q_0 = \sum_{i=0}^{n-1} q_{\hat{X}}(i) q_{X|\hat{X}}(i|i) = 1-d$. The equality 1) holds regardless of the permutation of q_{ij} due to the special structure of P; the inequalities 2) and 3) follow from Lemma 6 and Lemma 7 below separately; and the inequality 4) is due to the constraint $d = Ed(X,\hat{X}) \leq D$ and Lemma 7.

Lemma 6 *$H(Y|\hat{X}) - H(X|\hat{X})$ is a convex function of $q_{X|\hat{X}}$ for fixed $q_{\hat{X}}$.*

Proof of Lemma 6: Denote $J(q_{X|\hat{X}}) = H(Y|\hat{X}) - H(X|\hat{X})$. Let the conditional probability assignments $q'_{X|\hat{X}}$ and $q''_{X|\hat{X}}$ achieve the points $(q'_{X|\hat{X}}, J(q'_{X|\hat{X}}))$ and $(q''_{X|\hat{X}}, J(q''_{X|\hat{X}}))$, respectively. Note that

$$q^*_{X|\hat{X}} = \lambda q'_{X|\hat{X}} + (1-\lambda) q''_{X|\hat{X}} \tag{4.27}$$

is also a conditional probability assignment, then for fixed $q_{\hat{X}}$,

$$\begin{aligned}
J(q^*_{X|\hat{X}}) &= \sum_x \sum_y p_{Y|X} \sum_{\hat{x}} q^*_{X|\hat{X}} q_{\hat{X}} \log \frac{q^*_{X|\hat{X}}}{p^*_{Y|\hat{X}}} \\
&= \sum_x \sum_y p_{Y|X} \sum_{\hat{x}} (\lambda q'_{X|\hat{X}} + (1-\lambda) q''_{X|\hat{X}}) q_{\hat{X}} \log \frac{(\lambda q'_{X|\hat{X}} + (1-\lambda) q''_{X|\hat{X}}) q_{\hat{X}}}{\sum_{x'} (\lambda q'_{X'|\hat{X}} + (1-\lambda) q''_{X'|\hat{X}}) p_{Y|X'} q_{\hat{X}}} \\
&\overset{1)}{\leq} \lambda \sum_x \sum_y p_{Y|X} \sum_{\hat{x}} q'_{X|\hat{X}} q_{\hat{X}} \log \frac{q'_{X|\hat{X}} q_{\hat{X}}}{\sum_{x'} q'_{X'|\hat{X}} p_{Y|X'} q_{\hat{X}}} \\
&+ (1-\lambda) \sum_x \sum_y p_{Y|X} \sum_{\hat{x}} q''_{X|\hat{X}} q_{\hat{X}} \log \frac{q''_{X|\hat{X}} q_{\hat{X}}}{\sum_{x'} q''_{X'|\hat{X}} p_{Y|X'} q_{\hat{X}}} \\
&= \lambda J(q'_{X|\hat{X}}) + (1-\lambda) J(q''_{X|\hat{X}}) \qquad (4.28)
\end{aligned}$$

In the above, the inequality 1) is due to log sum inequality [2]. Therefore, the lemma follows, and $H(Y|\hat{X}) - H(X|\hat{X})$ is a convex function of $q_{X|\hat{X}}$ for fixed $q_{\hat{X}}$.

Lemma 7 *Let $D \in [0, (n-1)P_{n-1}]$, $q_i \in [0,1]$, and $\sum_{i=0}^{n-1} q_i = d \leq D$. P is in the set P_p by Definition 10. Then,*

$$H\left(\begin{pmatrix} 1-d & q_1 & \dots & q_{n-1} \end{pmatrix} P\right) - H\begin{pmatrix} 1-d & q_1 & \dots & q_{n-1} \end{pmatrix}$$

is minimized by $q_i = \frac{d}{n-1}$. Furthermore,

$$H\left(\begin{pmatrix} 1-d & \frac{d}{n-1} & \dots & \frac{d}{n-1} \end{pmatrix} P\right) - H\begin{pmatrix} 1-d & \frac{d}{n-1} & \dots & \frac{d}{n-1} \end{pmatrix}$$

is a monotonically decreasing function of d when $d \in (0, D]$.

Proof of Lemma 7: For brevity, we let

$$f(q_i) = H\left(\begin{pmatrix} 1-d & q_1 & \dots & q_{n-1} \end{pmatrix} P\right) - H\begin{pmatrix} 1-d & q_1 & \dots & q_{n-1} \end{pmatrix}$$

By taking the first order derivative of $f(q_i)$ with respect to q_i, one can check that $\frac{df(q_i)}{dq_i} = 0$ when $q_i = \frac{d}{n-1}$. Since $f(q_i)$ is a convex function of q_i by lemma 6, we see that $f(q_i)$ is minimized by $q_i = \frac{d}{n-1}$.

Let

$$g(d) = H\left(\begin{pmatrix} 1-d & \frac{d}{n-1} & \ldots & \frac{d}{n-1} \end{pmatrix} P\right) - H\begin{pmatrix} 1-d & \frac{d}{n-1} & \ldots & \frac{d}{n-1} \end{pmatrix}$$

Taking the first order derivative of $g(d)$ with respect to d, we get,

$$\begin{aligned} \frac{dg(d)}{dd} &= (p_1 - p_0)(-\log W_0 - 1) + (p_0 - p_1)(-\log W_1 - 1) - [\log(1-d) - \log\frac{d}{n-1}] \\ &= (p_1 - p_0)(\log W_1 - \log W_0) + \log\frac{d}{n-1} - \log(1-d), \end{aligned} \tag{4.29}$$

where $W_0 = (1-d)p_0 + dp_1$, and $W_1 = \frac{d}{n-1}p_0 + (1-\frac{d}{n-1})p_1$.

Let $\delta = p_1 - p_0$. Then the right-hand-side of (4.29) can be written as

$$F(d,\delta) = \frac{dg(d)}{dd} = \delta\log\frac{(1-\frac{d}{n-1})\delta + p_0}{d\delta + p_0} + \log\frac{d}{n-1} - \log(1-d) \tag{4.30}$$

When $\delta > 0$, since $p_0 + (n-1)p_1 = 1$, and $p_0, p_1 > 0$, then $\delta \in (0, \frac{1}{n-1})$. It is easy to verify both

$$\frac{\partial F(d,\delta)}{\partial\delta} \geq 0, \delta \in (0, \frac{1}{n-1}) \tag{4.31}$$

and

$$F(d, \frac{1}{n-1}) < 0 \tag{4.32}$$

(4.31) and (4.32) together imply that for any $\delta \in (0, \frac{1}{n-1})$, and any $d \in (0, D]$,

$$F(d,\delta) < F(d, \frac{1}{n-1}) < 0 \tag{4.33}$$

When $\delta \leq 0$,since $p_0 + (n-1)p_1 = 1$, and$p_0, p_1 > 0$, then $\delta \in (-1, 0)$. It is easy to verify that

$$\frac{\partial F(d,\delta)}{\partial\delta} \leq 0, \delta \in (-1, 0) \tag{4.34}$$

Further verify that

$$F(d, -1) < 0 \tag{4.35}$$

(4.34) and (4.35) together imply that for any $\delta \in (-1, 0)$, and any $d \in (0, D]$,

$$F(d, \delta) < F(d, -1) < 0 \tag{4.36}$$

Thus, (4.33) and (4.36) constitute the desired result that $g(d)$ is a monotonically decreasing function of d when $d \in (0, D]$.

4.3 Gaussian Case

Assume now the source $X \sim \mathcal{N}(0, \sigma_x^2)$, and side information Y is a continuous random variable. In practical use, we generally assume that Y is a noisy version of X, such that

$$Y = X + N_2 \tag{4.37}$$

where X, N_2 are independent random variables with $EN_2 = 0$, and $EN_2^2 = \sigma_2^2$.

The random variable $\hat{X}$ which would achieve $\hat{R}_{WZ}(D)$ can be written as,

$$X = \hat{X} + N_1 \tag{4.38}$$

where $EN_1 = 0$, $EN_1^2 = \sigma_1^2$, and $Y \to X \to \hat{X}$ forms a Markov chain. We use mean-squared error distortion measure with the constraint $Ed(X - \hat{X}) = \sigma_1^2 \leq D$, and assume N_1, N_2 independent.

It is well known that, in the classic rate-distortion system, for a $\mathcal{N}(0, \sigma_x^2)$ source, if $D \leq \sigma_x^2$, the probability density $p_{X|\hat{X}'}$ that governs the additive noise $N_1' = X - \hat{X}'$ in the "backward channel" is

$$N_1' \sim \mathcal{N}(0, D), \text{and independent of } \hat{X}'. \tag{4.39}$$

Theorem 7 *Let the source $X \sim \mathcal{N}(0, \sigma_x^2)$, and side information Y be continuous. For any $D \geq 0$, the quantizer $\hat{X}'$ that achieves $R(D)$ for X also achieves $\hat{R}_{WZ}(D)$ for (X, Y), if X and Y are jointly Gaussian.*

Proof of Theorem 7: Theorem 7 obviously holds when $D > \sigma_x^2$ or $D = 0$. Assume that $D \in (0, \sigma_x^2]$. Since X, Y are jointly Gaussian, without loss of generality, we can always write Y as (4.37), where X, N_2 are two independent Gaussian variables.

Observe that

$$\begin{aligned} & I(X;\hat{X}) - I(Y;\hat{X}) \\ = \; & h(X) - h(Y) + [h(Y|\hat{X}) - h(X|\hat{X})] \end{aligned} \tag{4.40}$$

In order to minimize $I(X;\hat{X}) - I(Y;\hat{X})$ subject to $Ed(X,\hat{X}) \leq D$, one should look for a Gaussian test channel $N_1 \sim \mathcal{N}(0, D)$ which is equal to the optimum test channel of the classic rate distortion system in (4.39).

$$\begin{aligned} h(Y|\hat{X}) - h(X|\hat{X}) &= h(Y - \hat{X}|\hat{X}) - h(X - \hat{X}|\hat{X}) \\ &= h(N_1 + N_2|\hat{X}) - h(N_1|\hat{X}) \\ &= \int_{\hat{x}} p_{\hat{X}}(\hat{x})[h(p_{N_1|\hat{X}}(x|\hat{x}) * p_{N_2}(x)) - h(p_{N_1|\hat{X}}(x|\hat{x}))]d\hat{x} \\ &\overset{1)}{\geq} h(\int_{\hat{x}} p_{\hat{X}}(\hat{x}) p_{N_1|\hat{X}}(x|\hat{x})d\hat{x} * p_{N_2}(x)) - h(\int_{\hat{x}} p_{\hat{X}}(\hat{x}) p_{N_1|\hat{X}}(x|\hat{x})d\hat{x}) \\ &= h(p_{N_1}(x) * p_{N_2}(x)) - h(p_{N_1}(x)) \\ &= h(N_1 + N_2) - h(N_1) \\ &\overset{2)}{\geq} \frac{1}{2}\log\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2} \\ &\overset{3)}{\geq} \frac{1}{2}\log(1 + \frac{\sigma_2^2}{D}) \end{aligned} \tag{4.41}$$

where $p * q = \int_{x_1} p(x_1)q(x - x_1)dx_1$. In the above, the inequality 1) follows from Jensen's inequality and Lemma 8, the inequality 2) follows from Lemma 9, and the inequality 3) is due to the constraint $\sigma_1^2 \leq D$.

Combining (4.40) and (4.41), we have

$$\begin{aligned} & I(X;\hat{X}) - I(Y;\hat{X}) \\ \geq\ & h(X) - h(Y) + \frac{1}{2}\log(1+\frac{\sigma_2^2}{D}) \\ =\ & \frac{1}{2}\log\frac{\sigma_x^2(D+\sigma_2^2)}{(\sigma_x^2+\sigma_2^2)D}, 0 < D \leq \sigma_x^2 \end{aligned} \tag{4.42}$$

If $D > \sigma_x^2$, we choose $\hat{X} = 0$ with probability 1, achieving $\hat{R}_{WZ}(D) = 0$. Hence, we conclude that for X, Y jointly Gaussian,

$$\hat{R}_{WZ}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma_x^2(D+\sigma_2^2)}{(\sigma_x^2+\sigma_2^2)D} & 0 < D \leq \sigma_x^2 \\ 0 & D > \sigma_x^2 \end{cases}, \tag{4.43}$$

The right-hand-side of (4.42) is indeed achievable with the random variable $\hat{X}'$ that achieve $R(D)$.

In (4.43), we derived a lower bound of $\hat{R}_{WZ}(D)$, which is called the generalized Shannon lower bound for Gaussian source X with square error criterion. For non-Gaussian channel from X to Y, however, this lower bound is not achievable, which provides the following useful upper bound to $\hat{R}_{WZ}(D)$ when X is Gaussian.

Theorem 8 *: Let $X \sim \mathcal{N}(0, \sigma_x^2)$, and p_{N_2} be any probability density with mean zero and variance σ_2^2. That is, suppose*

$$\int_x x p_{N_2}(x) dx = 0 \tag{4.44}$$

and

$$\int_x x^2 p_{N_2}(x) dx = \sigma_2^2 \tag{4.45}$$

let $Y = X + N_2$ with X, N_2 independent, and p_{N_1} of (4.39) governs the transition from $\hat{X}$ to X where $X = \hat{X} + N_1$. Then,

$$\hat{R}_{WZ}(D) \leq \frac{1}{2}\log\frac{\sigma_x^2}{D} - I(Y;\hat{X}) \leq \frac{1}{2}\log\frac{\sigma_x^2(D+\sigma_2^2)}{(\sigma_x^2+\sigma_2^2)D} \tag{4.46}$$

with equality iff $p_(N_2)$ *is* $\mathcal{N}(0,\sigma_2^2)$.

Proof of Theorem 8: By the definition of $\hat{R}_{WZ}(D)$, $X \sim N(0,\sigma_x^2)$ and (4.39), we have,

$$\begin{aligned}
\hat{R}_{WZ}(D) &\leq I(X;\hat{X}) - I(Y;\hat{X}) \\
&= h(X) - h(X|\hat{X}) - I(Y;\hat{X}) \\
&= \frac{1}{2}\log\frac{\sigma_x^2}{D} - I(Y;\hat{X})
\end{aligned} \tag{4.47}$$

This establishes the left side of inequality (4.46). In order to obtain the right side, we try to minimize $I(Y;\hat{X})$.

By (4.39), we see that $\hat{X} \sim \mathcal{N}(0,\sigma_x^2 - D)$ is independent of $N_1 \sim \mathcal{N}(0,D)$, also independent of N_2. Let $N = N_1 + N_2$ with $EN = 0, EN^2 = D + \sigma_2^2$, we have

$$\begin{aligned}
I(\hat{X};Y) &= I(\hat{X};X+N_2) \\
&= I(\hat{X};\hat{X}+N_1+N_2) \\
&= I(\hat{X};\hat{X}+N) \\
&\overset{1)}{\geq} I(\hat{X};\hat{X}+N^*) \\
&= \frac{1}{2}\log\frac{\sigma_2^2+\sigma_x^2}{\sigma_2^2+D}
\end{aligned} \tag{4.48}$$

In the above, $N^* \sim \mathcal{N}(0, D+\sigma_2^2)$. 1) follows from Lemma II.2 in [20], with the equality holds iff $N = N^*$. Combine (4.47) and (4.48), the theorem follows, that is,

$$\begin{aligned}
\hat{R}_{WZ}(D) &\leq I(X;\hat{X}) - I(Y;\hat{X}) \\
&= \frac{1}{2}\log\frac{\sigma_x^2}{D} - \frac{1}{2}\log\frac{\sigma_2^2+\sigma_x^2}{\sigma_2^2+D} \\
&= \frac{1}{2}\log\frac{\sigma_x^2(D+\sigma_2^2)}{(\sigma_x^2+\sigma_2^2)D}, 0 < D \leq \sigma_x^2
\end{aligned} \tag{4.49}$$

where the equality holds iff $N \sim \mathcal{N}(0, D + \sigma_2^2)$. By Cramer's theorem, $N_1 + N_2 \sim \mathcal{N}(0, D + \sigma_2^2)$ iff N_1, N_2 are both Gaussian, hence $N_2 \sim \mathcal{N}(0, \sigma_2^2)$.

Lemma 8 *For the continuous alphabet, if $Y \to X \to \hat{X}$ forms a Markov chain, then $f_p = h(Y|\hat{X}) - h(X|\hat{X})$ is a convex function with respect to $p_{X|\hat{X}}$ for fixed $p_{\hat{X}}$.*

Proof of Lemma 8: Let the conditional probability assignments $p'_{X|\hat{X}}$ and $p''_{X|\hat{X}}$ achieve the points $(p'_{X|\hat{X}}, f'_p)$ and $(p''_{X|\hat{X}}, f''_p)$, respectively, and note that

$$p^*_{X|\hat{X}} = \lambda p'_{X|\hat{X}} + (1 - \lambda) p''_{X|\hat{X}} \tag{4.50}$$

is also a conditional probability assignment. We employ the well-known inequality

$$\log x \leq x - 1 \tag{4.51}$$

in order to complete the following proof.

$$
\begin{aligned}
f_{p^*} &= \int_x \int_y \int_{\hat{x}} p^*_{X|\hat{X}} p_{\hat{X}} q_{Y|X} \log \frac{p^*_{X|\hat{X}}}{p^*_{Y|\hat{X}}} dx dy d\hat{x} \\
&= \lambda \int_x \int_y \int_{\hat{x}} p'_{X|\hat{X}} p_{\hat{X}} q_{Y|X} \log \frac{p'_{X|\hat{X}} p^*_{X|\hat{X}} p'_{Y|\hat{X}}}{p'_{Y|\hat{X}} p^*_{Y|\hat{X}} p'_{X|\hat{X}}} dx dy d\hat{x} \\
&+ (1-\lambda) \int_x \int_y \int_{\hat{x}} p''_{X|\hat{X}} p_{\hat{X}} q_{Y|X} \log \frac{p''_{X|\hat{X}} p^*_{X|\hat{X}} p''_{Y|\hat{X}}}{p''_{Y|\hat{X}} p^*_{Y|\hat{X}} p''_{X|\hat{X}}} dx dy d\hat{x} \\
&\leq \lambda \int_x \int_y \int_{\hat{x}} p'_{X|\hat{X}} p_{\hat{X}} q_{Y|X} [\log \frac{p'_{X|\hat{X}}}{p'_{Y|\hat{X}}} + \frac{p^*_{X|\hat{X}} p'_{Y|\hat{X}}}{p^*_{Y|\hat{X}} p'_{X|\hat{X}}} - 1] dx dy d\hat{x} \\
&+ (1-\lambda) \int_x \int_y \int_{\hat{x}} p''_{X|\hat{X}} p_{\hat{X}} q_{Y|X} [\log \frac{p''_{X|\hat{X}}}{p''_{Y|\hat{X}}} + \frac{p^*_{X|\hat{X}} p''_{Y|\hat{X}}}{p^*_{Y|\hat{X}} p''_{X|\hat{X}}} - 1] dx dy d\hat{x} \\
&= \lambda [f_{p'} - 1 + \int_{\hat{x}} \int_y \frac{p'_{Y|\hat{X}} p_{\hat{X}}}{p^*_{Y|\hat{X}}} (\int_x q_{Y|X} p^*_{X|\hat{X}} dx) dy d\hat{x}] \\
&+ (1-\lambda) [f_{p''} - 1 + \int_{\hat{x}} \int_y \frac{p''_{Y|\hat{X}} p_{\hat{X}}}{p^*_{Y|\hat{X}}} (\int_x q_{Y|X} p^*_{X|\hat{X}} dx) dy d\hat{x}] \\
&= \lambda [f_{p'} - 1 + \int_{\hat{x}} \int_y \frac{p'_{Y,\hat{X}}}{p^*_{Y|\hat{X}}} p^*_{Y|\hat{X}} dy d\hat{x}] \\
&+ (1-\lambda) [f_{p''} - 1 + \int_{\hat{x}} \int_y \frac{p''_{Y,\hat{X}}}{p^*_{Y|\hat{X}}} p^*_{Y|\hat{X}} dy d\hat{x}] \\
&= \lambda f_{p'} + (1-\lambda) f_{p''} \qquad (4.52)
\end{aligned}
$$

Combining (4.50) and (4.52) yields the desired result.

Lemma 9 *Let $Y^* \sim \mathcal{N}(0, \sigma_y^2)$, and X is a continuous random variable independent of Y with mean zero and variance σ_x^2, then*

$$h(X + Y^*) - h(X) \geq h(X^* + Y^*) - h(X^*) \qquad (4.53)$$

where $X^* \sim \mathcal{N}(0, \sigma_x^2)$, and equality holds iff $X = X^*$.

Proof of Lemma 9: By the Entropy-Power Inequality, we have,

$$h(X+Y^*) - h(X) \;\geq\; \frac{1}{2}\ln(e^{2h(X)} + e^{2h(Y^*)}) - h(X) \tag{4.54}$$

where the equality holds iff $X = X^*$.

Observe that, the right-hand-side of (4.54) is a monotonically decreasing function of $h(X)$. Since with fixed variance, normal distribution maximizes the differential entropy, thus we have,

$$h(X+Y^*) - h(X) \;\geq\; h(X^*+Y^*) - h(X^*) \tag{4.55}$$

where equality holds iff $X = X^*$.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In point-to-point communication, side information gives some extra information about the source and channel to the transmitter and/or the receiver. For instance, the side information can be the nature and the format of the source. Although side information can be present in the encoder and/or the decoder and yields several cases, the most important case that is worth particular attention is source coding with side information at the decoder (Wyner-Ziv coding) which requires different design strategies from the conventional source coding problem. Due to the difficulty caused by the joint design of the random variable and reconstruction function, a common approach to this lossy source coding problem is to apply conventional vector quantization followed by Slepian-Wolf coding. In this thesis, we investigated the best rate-distortion performance achievable asymptotically by practical Wyner-Ziv coding schemes of above approach from an information theoretic viewpoint and a numerical computation viewpoint respectively.

From the information theoretic viewpoint, we established the corresponding rate-distortion function $\hat{R}_{WZ}(D)$ for any memoryless (X, Y) and any distortion measure. We proved the achievability and converse of the derived rate-distortion function, and the convexity property was also shown.

To gain deep insights on $\hat{R}_{WZ}(D)$, on the other hand, from the viewpoint of numerical computation, we focus on the algorithm development to study the rate-distortion performance and provide guidelines for designing practical Wyner-Ziv coding which reduces to investigating $\hat{X}$. Based on the iteration idea of the Blahut-Arimoto algorithm for computing the classic rate-distortion function, an extended Blahut-Arimoto algorithm was proposed, and the convergence of the algorithm was also proved.

Interestingly, the extended Blahut-Arimoto algorithm allows us to observe an important phenomenon where the random variable $\hat{X}$ that achieves $\hat{R}_{WZ}(D)$ is generally different from the random variable $\hat{X}'$ that achieves the classical rate-distortion $R(D)$ in (2.19). Surpringly, there are indeed cases where $\hat{X} = \hat{X}'$. To fully understand and characterize this important and rather surprising phenomenon, we are led to the question that under what conditions are the two random variables equivalent or distinct. We completely settle this problem for the case where $\mathcal{X}$, $\mathcal{Y}$, and $\hat{\mathcal{X}}$ are all binary, the two random variables are the same if and only if the channel from source X to side information Y is symmetric. Furthermore, we also determine the sufficient condition (equivalent condition) for non-binary alphabets case with Hamming distortion measure case, and the case of Gaussian source with mean-squared error distortion measure case respectviely.

5.2 Future Work

Practical Wyner-Ziv problem has been recently an active research field in information theory. From the viewpoint of information theory, there remains many open problems.

1. In Chapter 4, for non-binary alphabets with Hamming distortion measure case and Gaussian source with mean-squared error distortion measure case, only sufficient conditions are determined, i.e., only the conditions under which the quantizer that achieves $\hat{R}_{WZ}(D)$ is the same as quantizer that achieves the classical rate-distortion $R(D)$ are proved. However, we don't know whether the conditions are necessary or not. It is valuable to find sufficient and necessary conditions similar to that of the binary case.

2. What's the performance gap between the traditional Wyner-Ziv system and practical Wyner-Ziv system by our approach? We ran some simulations on binary symmetric source for the two cases (See Fig 3.2), and the gap is minor. Is the gap still acceptable for arbitrary cases? Mathematic proof should be given to gain deep insights.

3. After finish the above open problems, it should be interesting to investigate some other practical Wyner-Ziv schemes which might be more complicated.

Bibliography

[1] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and Construction," *Proceedings of the Data Compression Conference (DCC)*, Snowbird, UT, March 1999.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: NY: Wiley, 1991.

[3] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, pp. 1–10, 1976.

[4] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, 1973.

[5] A. B. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the Quadratic Gaussian two-encoder source-coding problem," *ISIT 2006, pp. 1404-1408*, July 2006, Seattle, WA

[6] Catarina Brites, Fernando Pereira, "Distributed video coding: bring new applications to life," *Proc Conf. on Telecommunications,* ConfTele , Tomar , Portugal

[7] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE, Special Issue on Video Coding and Delivery,* vol. 93, no. 1, pp. 71-83, January 2005

[8] A. D. Wyner, "Recent results in the Shannon theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 2–10, Jan. 1974.

[9] F. M. J Willems, "Computation of the Wyner-Ziv rate-distortion function," Katholieke Universiteit Leuven, Department Wiskunde, October 1982.

[10] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1181-2003, July 2003

[11] I. Csiszar, "On the computation of rate distortion functions," *IEEE Trans. Inform. Theory*, vol. 20, pp. 122–124, 1974.

[12] T. Berger, *Rate-distortion Theory: A Mathematical Basis for Data Compression,* Englewood Cliffs, N.J. :Prentice-Hall. 1971.

[13] R. E. Blahut, "Computation of channel capacity and rate-distortion function," *IEEE Trans. Inform. Theory*, vol. 18, pp. 460–473, 1972.

[14] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, New York: Springer-Verlag, 1988

[15] R. G. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968.

[16] S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed compression in a dense microsensor network," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 51-60, 2002

[17] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding: Channel coding for compression," *IEEE Signal Processing Magazine*, Sept. 2004

[18] Harald Cramer, *Random Variables and Probability Distributions,* Cambridge Tracts in Mathematics and Mathematical Physics, London, Cambridge University Press, 1970.

[19] M. Loeve, *Probability Theory,* Van Nostrand Reinhold, New York, 1955

[20] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covatiance constraint," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, 2001.

[21] Lin Zheng, Da-ke He, and En-hui Yang, "On Optimum Conventional Quantization For Source Coding With Side Information At The Decoder," *The 10th Canadian Workshop on Information Theory,* Edmonton, Canada, June, 2007