To Use the Codebook Information or Not: A Study of the Compress-and-Forward Relay Strategy

by

Xiugang Wu

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, 2008

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

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

Abstract

The motivation of this thesis is to understand how nodes can cooperate in a particular relay channel, say a relay channel with orthogonal link between relay and destination. We are especially interested in the scenario where relay cannot decode the message sent because the coding rate is too large vis-a-vis what it can handle, and try to investigate the optimality of compress-and-forward strategy for this scenario.

Specifically, noting that relay's compression is based on the unconditional distribution of its observation, it is natural to ask if relay can do better if it utilizes the source's codebook information, such that the performance of the relay network is improved? To answer this key question, we need to find the posterior Conditional Distribution of channel Output given Codebook Information (namely CDOCI) for the channel between source and relay.

Firstly, we model it as Binary Symmetric Channel (BSC) and show that under the now standard random coding framework, if the input distribution is uniform, then with high probability, the CDOCI is asymptotically uniform when coding rate is greater than the channel capacity and the block length is sufficiently large. Then it is shown that under the discrete memoryless channel (DMC), for those strongly typical output sequences, with high probability, the CDOCI is also asymptotically uniform and close to the unconditional distribution, for rate above capacity and sufficiently large block length. These results implicate that relay can hardly do better with codebook information used. To confirm this implication, we show that for sufficiently large block length, the rate needed for relay to forward its observation when the codebook information is utilized approaches the rate needed when the relay simply ignore the codebook information, if the coding rate at source is larger than channel capacity.

Now the answer to the above key question is apparent: in the cases of BSC and DMC, even if the relay tries to utilize the information obtained by knowing the codebook used at the source, it can hardly do better than simply ignore the codebook information. Therefore, the compress-and-forward strategy is kind of optimal in this sense, under the random coding framework.

Acknowledgements

First and foremost, I would like to thank my adviser Professor Liang-Liang Xie for his invaluable guidance, continuous encouragement and generous financial support during my master study period. He led me into the exciting area of network information theory, and taught me not only the approach to the specific problem in this thesis, but also his insights, inspirations and enthusiasm to conduct fundamental research.

Second, I also would like to thank the readers of this thesis, Professor Xuemin Shen and Professor Zhou Wang, for taking the time to read my thesis and providing constructive suggestions.

Last but not the least, I am deeply indebted to all my friends and my family, for their love and support.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problems and Motivations

Consider a randomly generated codebook for a *Binary Symmetric Channel* **(BSC)** shown in Figure 1.1, as in the seminal approach of Shannon [1]. Our result is to show that, if the input distribution is uniform, the conditional distribution of channel output sequence at the receiver end Y given the full information of the random codebook used at the source end X is asymptotically uniform with high probability when coding rate is greater than the channel capacity and the block length is sufficiently large. (For simplicity, we use the abbreviation **CDOCI** standing for the *Conditional Distribution of channel Output given Codebook Information* throughout this thesis.)

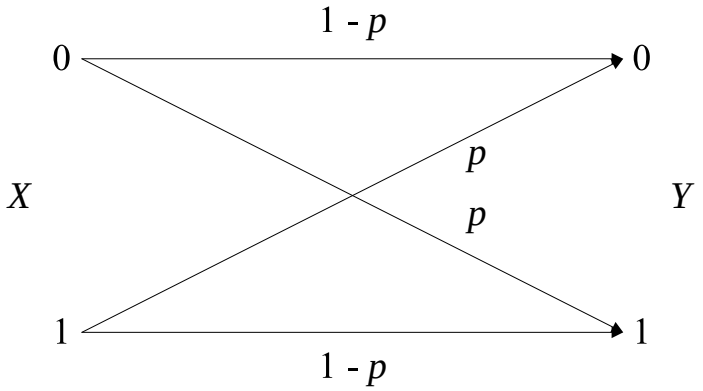

Figure 1.1: A binary symmetric channel.

To extend this result and get a general understanding of CDOCI, we then consider the *Discrete Memoryless Channel* (**DMC**), as depicted in Figure 1.2. We show that for those ***strongly typical*** output sequences, with high probability, the

CDOCI is also asymptotically uniform and close to the unconditional distribution, for rate above capacity and sufficiently large block length.

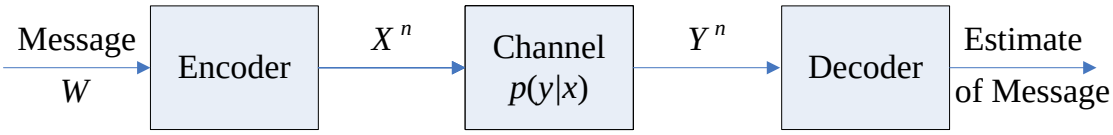

Figure 1.2: A discrete memoryless channel.

Our motivation to study this issue is to understand how nodes can cooperate in networks of the sort shown in Figure 1.3. This particular system shown contains a relay. The source node 1 wants to wirelessly send information to node 3, while node 2 is a relay that is available to help. Suppose that the channel between relay and destination is a wired channel, or more generally a channel that is "orthogonal" to channels 1–2 and 1–3, say one that transmits on a different frequency band. As the cut in Figure 1.3 shows, this network can be regarded as a broadcast channel from node 1 to node 2 and node 3, with output Y_2 and Y_3 respectively. Also assume that $p(y_2, y_3|x_1) = p(y_2|x_1) \cdot (y_3|x_1)$, so that the node 2 and node 3 obtain independent observations (given the message sent). As shown in Cover and El Gamal [2], if relay can decode the codeword sent by source, and link 1–2 is better than link 1–3, then relay should ***decode-and-forward*** and fortunately this strategy achieves the capacity.

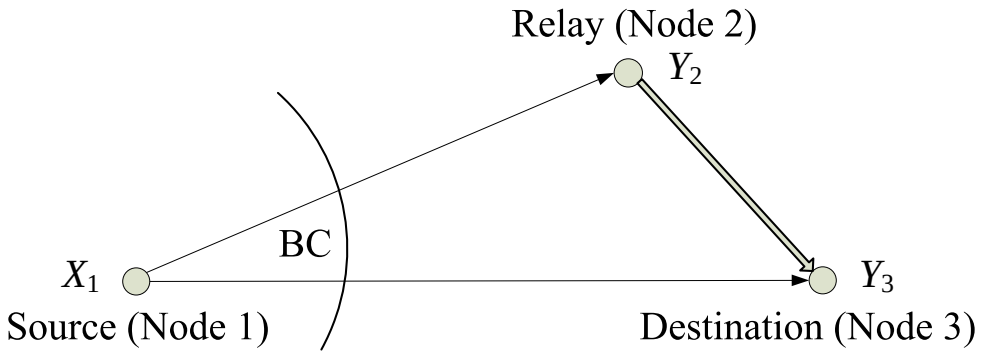

Figure 1.3: A relay channel with orthogonal link between relay and destination.

In this thesis, however, we are interested in the opposite scenario where relay cannot decode the message sent because the coding rate is too large vis-a-vis what it can handle. Cover and El Gamal also proposed their second strategy in [2] for this scenario, which is now often called ***compress-and-forward***. The idea of compress-and-forward strategy is that the relay node 2 transmits a quantized and

compressed version $\hat{Y}_2$ of its channel output Y_2 to the destination node 3 and node 3 decodes by making use of both $\hat{Y}_2$ and its own output Y_3. Specifically, what the relay does in this strategy is simply quantizing what it has received according to the distribution of channel output and using Wyner-Ziv source coding to exploit side information at the destination [3].

Notice that in this strategy, relay's compression is based on the **unconditional distribution** of the output of channel 1–2. However, the fact is that different randomly generated codebooks might result in different distributions of the output for channel 1–2 at the relay. An extreme example is when all the codewords in the randomly generated codebook happen to be the same, many outputs might very unlikely occur and the number of "possible" outputs might reduce dramatically, meaning that the relay only need to take care of those possible outputs and the compression performance can be easily improved compared to compress-and-forward strategy. So, up to now, a fundamental question naturally comes out: **Can the relay do better compression if it utilizes the codebook information, such that the performance of the relay network is improved? In another word, is there any gain for the relay to compress its observation "wisely" by exploiting the knowledge which specific codebook is used at the source end?** This is the key question we want to explore in this thesis.

To answer this question, obviously, we need to find the conditional distribution of output for channel 1–2. (A related reference studying the conditional entropy of the jointly typical set is [4].) Specifically, given that the relay knows the codebook used at the source, is the output's conditional distribution the same as the unconditional distribution? As mentioned in the beginning, for BSC whose input distribution is uniform, the CDOCI of each output sequence is asymptotically uniform and approaching the unconditional distribution with high probability for rate above capacity and sufficiently large block length, and hence the compression performance can be hardly improved. For the general DMC, the CDOCI of the strongly typical output sequences is also asymptotically uniform and approaching the unconditional distribution for rate above capacity and sufficiently large block length. Noting that the the typical sequence set contains most of the probability, so the behavior of typical sequences dominates that of the ensemble and therefore the compression performance could be hardly improved either.

To strengthen our judgement, we study the rate needed for the relay to forward its observation under DMC scenario. It is shown that for sufficiently large block length, the rate needed when the codebook information is utilized approaches the rate needed when the relay simply ignore the codebook information, if the coding rate at source end is larger than Shannon capacity.

Now, the answer to the key question is apparent: even if the relay tries to utilize the information obtained by knowing the specific codebook used at the source end, it can hardly do better than simply ignore the codebook information. Therefore, under the random coding framework, the compress-and-forward strategy is kind of optimal in this sense.

1.2 Contributions

- Motivated by the relay problem depicted in Figure 1.3, we study the conditional distribution of channel output given codebook information when coding rate is greater than Shannon capacity, for both the binary symmetric channel and discrete memoryless channel. We show that with high probability, the CDOCI is asymptotically uniform and approaching the unconditional distribution when the block length is sufficiently large. This implicates that we can hardly do better than compress-and-forward strategy even if we let the relay compress its observation utilizing codebook information.

- For the general discrete memoryless channel, we characterize the rate needed for the relay to forward its observation. By showing that the rates needed are asymptotically the same no matter whether the relay utilize the codebook information or not, the above implication on the optimality of compress-and-forward is strengthened.

- As a by-product of studying the relay problem and CDOCI, we show that it is impossible to transmit at rate above capacity using the technique of random coding in the sense that with high probability the error probability will go to 1 for rate greater than capacity. This can be regarded as a strong converse to channel coding theorem under the now standard random codebook construction shema.

- It is worth to point out that although we aim to study the relay problem, our results and proof method may provide a new way to understand the random coding framework and consider other problems. On one hand, the property of the uniformity of CDOCI for rate greater than capacity may find future other applications in the field of information theory. On the other hand, the proposal of typical codebook set in the proof allows us to focus on the typical codebooks and analyze the behavior of them, as we have done for the typical sequences. This method of thinking is different from the classical way of analyzing averaged over all the codebooks, and adds additional insight to the random coding.

1.3 Thesis Outline

This thesis is organized as follows:

In Chapter 2, we give a review of the preliminaries of network information theory. As the prerequisite knowledge for the later discussion, we introduce some classical results in single-user information theory, such as the concepts and properties of weak typicality and strong typicality and channel capacity. Then we present several multi-user communication problems, including distributed source coding,

multiple access channels, broadcast channels and relay channels. With an emphasis on relay channels, we particularly review two strategies: decode-and-forward and compress-and-forward.

In Chapter 3, we model the channel between source and relay with binary symmetric channel and focus on the CDOCI. By describing the geometric distribution of codewords in the signal space, we propose the concept of typical codebook for the binary symmetric channel and show the probability of typical codebook goes to 1. Then, under the typical codebook, we study the CDOCI by accumulating the probability contributions from the codewords. It is shown that the CDOCI is asymptotically uniform with high probability for rate greater than capacity and sufficiently large block length. This indicates that the relay cannot do better even if it utilize the codebook information to compress the observation.

In Chapter 4, the CDOCI for general discrete memoryless channel is fully explored. We use the technique of strong version of typicality to define the typical codebook set for discrete memoryless channel, which is shown to take most probability. Similarly with the binary symmetric channel, we obtain that for those strongly typical output, the CDOCI is asymptotically uniform with high probability for rate greater than capacity and sufficiently large block length. This again demonstrates the futility of attempt for relay to utilize the codebook information and achieve better compression performance. To confirm our judgement, we particularly find the rate needed for the relay to forward its observation with codebook information used. Since this rate is asymptotically equal to the rate needed with codebook information unused, our judgement is strengthened. Besides, as a by-product of studying CDOCI, we give a strong converse of channel coding theorem under random coding framework.

Finally, we conclude this thesis and propose the possible future work to be done in Chapter 5.

Chapter 2

Preliminaries of Network Information Theory

2.1 Basic Tools and Results in Classical Information Theory

In this section, we mainly introduce some basic and important tools and results in information theory, which will be used throughout this thesis.

2.1.1 Weak Typicality

We introduce the definitions of weak typicality and jointly weak typicality, as well as the well-known asymptotic equipartition property. All these results with their proofs can be found in [5].

Weakly Typical Sequences

Definition 2.1.1. *The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property*

$$2^{-n[H(X)+\epsilon]} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n[H(X)-\epsilon]}. \tag{2.1}$$

Theorem 2.1.1 (Asymptotic Equipartition Property(AEP))**.** *Let X^n be a sequence of length n drawn i.i.d. according to $p(x^n) = \Pi_{i=1}^n p(x_i)$, then:*

1. *$Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.*
2. *$|A_\epsilon^{(n)}| \leq 2^{n[H(X)+\epsilon]}$, where $|A|$ denotes the number of elements in the set A.*
3. *$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n[H(X)-\epsilon]}$ for n sufficiently large.*

Proof. The proof is based on the Law of Large Numbers and one can find the details in [5]. □

Jointly Weakly Typical Sequences

Definition 2.1.2. *The set $A_\epsilon^{(n)}$ of jointly typical sequence (x^n, y^n) with respect to the distribution $p(x, y)$ is the set of n-sequences satisfying*

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon, \tag{2.2}$$

$$\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| \leq \epsilon, \tag{2.3}$$

$$\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| \leq \epsilon, \tag{2.4}$$

$$\tag{2.5}$$

where

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i). \tag{2.6}$$

Theorem 2.1.2 (Joint AEP)**.** *Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \Pi_{i=1}^n p(x_i, y_i)$, then:*

1. *$Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$, as $n \to \infty$.*
2. *$|A_\epsilon^{(n)}| \leq 2^{n[H(X,Y)+\epsilon]}$ and $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n[H(X,Y)-\epsilon]}$ for sufficiently large n.*
3. *If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then*

$$Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n[I(X;Y)-3\epsilon]}, \tag{2.7}$$

also, for sufficiently large n,

$$Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n[I(X;Y)+3\epsilon]}. \tag{2.8}$$

2.1.2 Strong Typicality

We give the definitions of strong typicality, strongly joint typicality, and conditionally strong typicality. For detailed discussion on strong typicality and the properties of strong typicality, see the book by Csiszár and Körner [21].

Strong Typicality

Definition 2.1.3. *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*

$$\left| \frac{1}{n} N(a|x^n) - p(a) \right| < \frac{\epsilon}{|\mathcal{X}|} \tag{2.9}$$

2. *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.

The set of sequences $x^n \in \mathcal{X}^n$ such that x^n is strongly typical is called the strongly typical set and is denoted by $A_\epsilon^{*(n)}(X)$ or $A_\epsilon^{*(n)}$ when the random variable is understood from the context.

Theorem 2.1.3 (Strong AEP)**.**

$$Pr(A_\epsilon^{*(n)}) \to 1, \ as \ n \to \infty. \tag{2.10}$$

Proposition 2.1.1. *For any* $x^n \in \mathcal{X}^n$, *if* $x^n \in A_\epsilon^{*(n)}(X)$, *then* $x^n \in A_\delta^{(n)}(X)$, *where* $\delta \to 0$, *as* $\epsilon \to 0$.

Strongly Joint Typicality

Definition 2.1.4. *A pair of sequences* $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ *is said to be* ϵ*-strongly jointly 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*

$$\left| \frac{1}{n} N(a, b|x^n, y^n) - p(a, b) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \tag{2.11}$$

2. *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 pair* (a, b) *in the pair of sequences* (x^n, y^n).

Theorem 2.1.4 (Consistency)**.** *If* $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$, *then* $x^n \in A_\epsilon^{*(n)}(X)$ *and* $y^n \in A_\epsilon^{*(n)}(Y)$.

Conditionally Strong Typicality

Definition 2.1.5. *A sequence* $y^n \in \mathcal{Y}^n$ *is said to be* ϵ*-strongly conditionally typical with the sequence* x^n *with respect to a conditional distribution* $V(\cdot|\cdot)$ *if:*

1. *For all* $(a, b) \in \mathcal{X} \times \mathcal{Y}$ *with* $V(b|a) > 0$, *we have*

$$\frac{1}{n} \left| N(a, b|x^n, y^n) - V(b|a) N(a|x^n) \right| < \frac{\epsilon}{|\mathcal{Y}| + 1} \tag{2.12}$$

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

The set of such sequences is called the conditionally typical set and it is denoted by $A_\epsilon^{*(n)}(Y|x^n)$.

Theorem 2.1.5. *Let Y^n be a sequence of length n generated with respect to the sequence x^n and a conditional distribution $V(\cdot|\cdot)$, then*

$$Pr(Y^n \in A_\epsilon^{(n)}(Y|x^n)) \to 1, \text{ as } n \to \infty. \tag{2.13}$$

Theorem 2.1.6. *If $x^n \in A_{\epsilon_1}^{*(n)}(X)$ and $y^n \in A_{\epsilon_2}^{*(n)}(Y|x^n)$, then $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$, where $\epsilon \to 0$ as $\epsilon_1 \to 0$ and $\epsilon_2 \to 0$.*

Proof. See the proof in the book by Csiszár and Körner [21]. □

Theorem 2.1.7. *There exists a sequence $\epsilon(n) \to 0$ respectively for $A_{\epsilon(n)}^{(n)}(X)$, $A_{\epsilon(n)}^{(n)}(X,Y)$, $A_{\epsilon(n)}^{*(n)}(X)$, $A_{\epsilon(n)}^{*(n)}(X,Y)$ and $A_{\epsilon(n)}^{*(n)}(Y|x^n)$ so that the probabilities of all these typical sets go to 1 as $n \to \infty$.*

Claim 2.1.1. *We use the the same convention in [21] throughout this thesis, i.e., in the asymptotic analysis, without reassertion, we use ϵ-typical sequences with $\epsilon = \epsilon(n)$ such that*

$$\epsilon(n) \to 0 \quad \text{and} \quad \sqrt{n} \cdot \epsilon(n) \to \infty \quad \text{as} \quad n \to \infty. \tag{2.14}$$

Remark 2.1.1. *In (2.14), the constraint $\epsilon(n) \to 0$ is useful for the precise asymptotic analysis, i.e., to make ϵ diminish to 0 as n grows, while the necessity of the constraint $\sqrt{n} \cdot \epsilon(n) \to \infty$ can be seen by lower bounding the probability of typical set using Chebyshev's inequality. In other words, the convention to choose $\epsilon(n)$ not only makes it possible to get rid of the undesired effect of the quantity ϵ in the asymptotic analysis, but also ensures the probability of typical set still goes to 1. For details, see (2.9), CONVENTION 2.11, LEMMA 2.12 and LEMMA 2.13 in [21].*

2.1.3 Channel Capacity

One of the most fundamental questions in information theory is: what is the ultimate transmission rate of communication. Channel coding theorem is the very answer to this question. It says that there always exsits a maximum achievable rate for a channel, called the channel capacity, below which the reliable communication can be implemented while above which the reliable communication is impossible. Moreover, it is fortunate that we have an elegant expression for the channel capacity. Take the discrete memoryless channel shown in Figure 1.2 for an example and we have the following theorem:

Theorem 2.1.8 (The Channel Coding Theorem)**.** *The capacity of a discrete memoryless channel $p(y|x)$ is defined as*

$$C = \max_{p(x)} I(X;Y), \tag{2.15}$$

where X and Y are respectively the input and the output of the channel, and the maximum is taken over all input distributions $p(x)$. Information can be transmitted as reliably as desired if and only if the rate is below the capacity C.

Random Coding and Achievability

At the first glance, the result is rather counter-intuitive. How can one correct all the errors introduced by the noisy channel and implement the reliable communication? To prove the achievability, we need some ideas which are firstly stated by Shannon in his original 1948 paper [1]. These original ideas are as follows:

Firstly, we allow a vanishing probability of error instead of requiring zero probability of error. Secondly, to put the law of large numbers into effect, we use the channel many times instead of only once. And finally, we randomly generate the codebook and calculate the probability of error averaged over the code ensemble. By showing the error probability averaged over all the codebooks goes to 0, we argue that there exists at least one good code with vanishing probability of error.

Converse

The capacity becomes a good dividing point not only because of the achievability part but also because we cannot achieve an arbitrarily low error probability at rates above capacity, which is just the converse part. To prove the converse, we need the *Fano's Inequality* and the details can be found in [5]. This converse is sometimes called the weak converse to channel coding theorem.

Moreover, a strong converse to the channel coding theorem can be proved, which states that for rates above capacity, the probability of error goes exponentially to 1 [22]. However, as stated in [22], since the strong converse is proved under block code scenario, compared to weak converse, it does not necessarily preclude the possibility of reliable transmission at rates above capacity. But this result still add some insight into the nature of channel capacity. Similarly, in Section 4.4, we provide a proof to the strong converse under random coding framework.

To summarize this subsection, we know that the channel coding theorem consists of two parts: achievability and converse. The idea of random coding is employed to prove the achievability and the converse can be proved by using Fano's Inequality.

2.2 A Brief Review of Several Triumphs in Multi-user Information Theory

In this section, we consider several multi-user communication models, including distributed source coding, multiple access channel and broadcast channel. Particularly, the relay channel will be discussed in details in the next section. As a quick review, all the results will be given directly without proof.

2.2.1 Distributed Source Coding

Slepian-Wolf Coding

We know that a rate $R > H(X)$ is sufficient to encode the source X. Now, consider the source coding problem presented in the Figure 2.1, where X and Y are correlated but encoded separately. It is natural to ask what is the sufficient rate pairs for the decoder to reconstruct both X and Y.

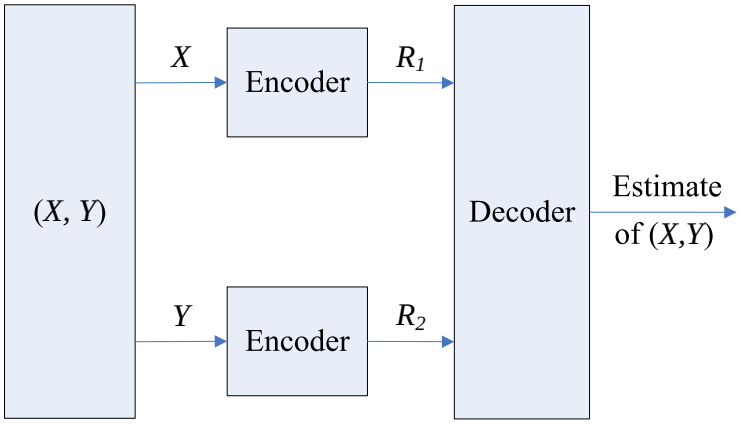

Figure 2.1: Slepian-Wolf Coding.

To accurate formulate and solve the distributed source coding problem, we use the following definition:

Definition 2.2.1. *A rate pair (R_1, R_2) is said to be achievable if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ distributed source codes with probability of error $P_e^{(n)} \to 0$, and the closure of the set of achievable rate pairs is called the achievable rate region.*

Slepian and Wolf gave the answer to this problem in their famous and fundamental paper [6], which is now often called *Slepian-Wolf Theorem*.

Theorem 2.2.1 (Slepian-Wolf Theorem)**.** *For the distributed source coding problem for the source* (X, Y) *drawn i.i.d.* $\sim p(x, y)$*, the achievable rate region is given by*

$$R_1 \geq H(X|Y), \tag{2.16}$$
$$R_2 \geq H(Y|X), \tag{2.17}$$
$$R_1 + R_2 \geq H(X, Y). \tag{2.18}$$

Random Binning and Achievability

The Slepian-Wolf Theorem was then extended to jointly ergodic sources by Cover [7]. In his paper, Cover used a binning argument, which has evolved to one of the most significant techniques beyond Shannon's random coding.

Briefly speaking, the technique is to randomly assign $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences to 2^{nR_1} and 2^{nR_2} indexed bins respectively, such that there are $2^{n(H(X)-R_1)}$ or $2^{n(H(Y)-R_2)}$ sequences in each bin. Given a realization x^n (or y^n), the encoder just simply transmits the index of the bin containing x^n (or y^n) and the decoder use the method of jointly decoding, i.e., find the jointly typical pair (x^n, y^n) contained in the bins corresponding to the received bin index. Readily we can see that, if $R_1 + R_2 > H(X, Y)$, then the probability that there exists another jointly typical sequence pair can be driven to 0 as $n \to \infty$. This is the idea of random binning and the outline of the achievability of Slepian-Wolf Theorem.

Wyner-Ziv Rate Distortion Function

We know that $R(D)$ is sufficient to describe X within distortion D. What if the side information Y is given? Formally, we have the following result:

Let (X, Y) be drawn i.i.d. $\sim p(x, y)$ and $d(\cdot, \cdot)$ be a distortion measure. Y is directly available to the decoder as the side information.

$$R_{WZ}(D) = \min I(X; W|Y) \tag{2.19}$$

where the minimization is taken over all W such that $W \to X \to Y$ forms a Markov Chain and all the functions $f : \mathcal{Y} \times \mathcal{W} \to \hat{\mathcal{X}}$ with $Ed(f(W, Y), X) \leq D$.

Basically, the Wyner-Ziv coding consists of two parts: quantizing as in the classical rate-distortion theory and random binning as in the Slepian-Wolf coding. The reader is referred to Wyner and Ziv [3] for the details of the proof. The Wyner-Ziv coding is employed in the compress-and-forward strategy for the relay channel problem, which will be discussed in 2.3.2.

2.2.2 Multiple Access Channel

The first multi-user channel we study is the multiple access channel, depicted in Figure 2.2, where sender 1 chooses an index W_1 uniformly from the set $\{1, 2, \cdots, 2^{nR_1}\}$

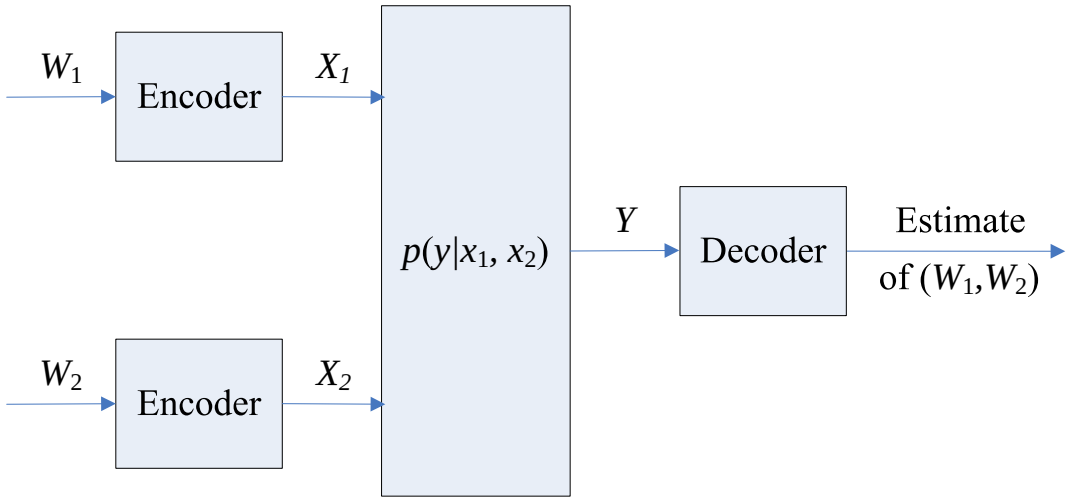

Figure 2.2: The multiple access channel.

and sends the corresponding codeword over the channel and sender 2 does likewise simultaneously.

A rate pair (R_1, R_2) is said to be achievable for the multiple access channel if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$, and the closure of the set of achievable rate pairs is called the capacity region. The multiple access channel capacity region was found by Ahlswede [8] and Liao [9] and is stated as follows:

Theorem 2.2.2. *The capacity region of a multiple access channel is given by the convex hull of all (R_1, R_2) satisfying*

$$R_1 \leq I(X_1; Y|X_2), \tag{2.20}$$

$$R_2 \leq I(X_2; Y|X_1), \tag{2.21}$$

$$R_1 + R_2 \leq I(X_1, X_2; Y) \tag{2.22}$$

for some product distribution $p_1(x_1)p_2(x_2)$ on $\mathcal{X}_1 \times \mathcal{X}_2$.

2.2.3 Broadcast Channel

The broadcast channel was firstly introduced by Cover in [10]. This channel describes the scenario where there is one sender and multiple (two or more) receivers, as illustrated in Figure 2.3. About this channel, one main concern is to find out the capacity region. Although the capacity region for general broadcast channels is still unknown, it has been made clear for some special classes, for example, the degraded broadcast channels [11], [12], [13].

Definition 2.2.2. *A broadcast channel is said to be physically degraded if*

$$p(y_1, y_2|x) = p(y_1|x)p(y_2|x). \tag{2.23}$$

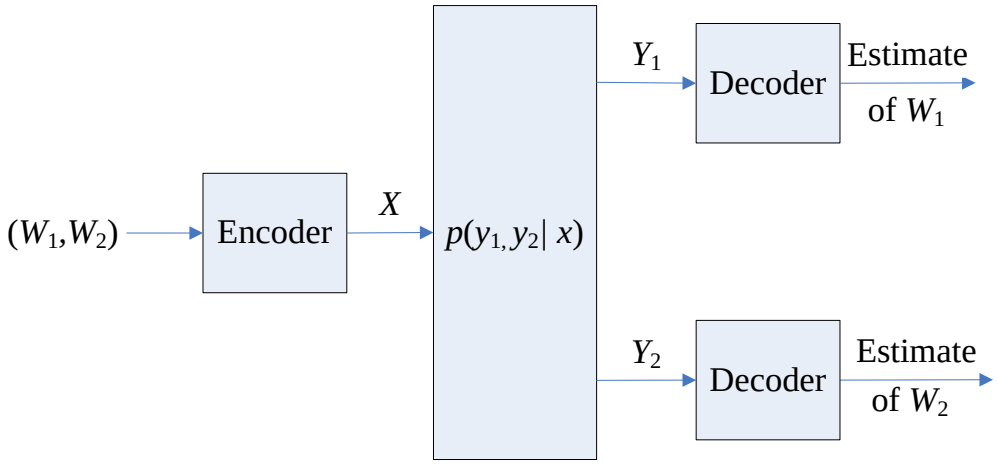

Figure 2.3: The broadcast channel.

A broadcast channel is said to be stochastically degraded if its conditional marginal distributions are the same as that of a physically degraded broadcast channel.

Since the capacity region of a broadcast channel depends only on the conditional marginal distributions $p(y_1|x)$ and $p(y_2|x)$ [5], both the physically degraded and the stochastically degraded broadcast channels have the same capacity region if they share the same conditional marginal distributions. The capacity region of degraded broadcast channel was first conjectured by Cover in [10], and then proved to be achievable by Bergmans [11], using the idea of ***superposition coding***. Finally Bergmans [13] and Gallager [12] established the converse.

Theorem 2.2.3. *The capacity region for sending independent information over the degraded broadcast channel $X \rightarrow Y_1 \rightarrow Y_2$ is the convex hull of the closure of all (R_1, R_2) satisfying*

$$R_2 \leq I(U;Y_2), \tag{2.24}$$
$$R_1 \leq I(X;Y_1|U) \tag{2.25}$$

for some joint distribution $p(u)p(x|u)p(y_1,y_2|x)$, where the auxiliary random variable U has cardinality bounded by $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\}$.

2.3 Relay Channel

The relay channel was introduced in the pioneering work of van der Meulen [14] [15]. A general model for the three-node discrete memoryless relay channel is depicted in Figure 2.4, where the relay and the source cooperate to resolve the receiver's uncertainty. One can easily find that this channel can be regarded as a combination of a broadcast channel(from source to relay and destination) and a multiple access

channel(from source and relay to destination). In fact, even this simplest relay channel is complex enough such that after several decades' effort, the exact capacity is still unknown except several special cases like physically degraded relay channel, which will be shown later. However, some substantial advances on this channel were made by Cover and El Gamal in their 1979 work [2], where two fundamental coding strategies, namely, Decode-and-Forward (DF) and Compress-and-Forward (CF), were developed.

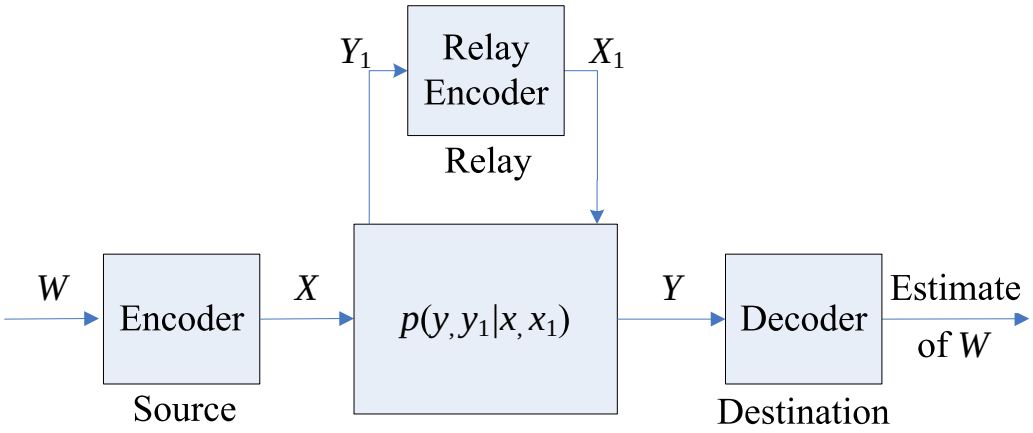

Figure 2.4: The relay channel.

2.3.1 Decode-and-Forward

Before proceeding to the two coding strategies mentioned above, we first introduce the concept of degraded relay channel, whose capacity has been established and can be achieved by decode-and-forward coding strategy.

Definition 2.3.1. *The relay channel* $(\mathcal{X} \times \mathcal{X}_1, p(y, y_1|x, x_1), \mathcal{Y} \times \mathcal{Y}_1)$ *is said to be degraded if relay receiver* y_1 *is better than the ultimate receiver* y *in the sense that*

$$p(y, y_1|x, x_1) = p(y_1|x, x_1)p(y|y_1, x_1). \tag{2.26}$$

Theorem 2.3.1. *The capacity* C *of the degraded relay channel is given by*

$$C = \sup_{p(x,x_1)} \min\{I(X, X_1|Y); I(X; Y_1|X_1)\} \tag{2.27}$$

where the supremum is over all joint distributions $p(x, x_1)$ *on* $\mathcal{X} \times \mathcal{X}_1$.

Outline of proof. Directly applying the max-flow-min-cut theorem for general multi-terminal networks to the relay channel, an upper bound of the capacity is obtained, i.e., for any relay channel, the capacity is bounded by

$$C \leq \sup_{p(x,x_1)} \min\{I(X, X_1|Y); I(X; Y, Y_1|X_1)\}. \tag{2.28}$$

Notice that due to degradedness, $I(X;Y,Y_1|X_1) = I(X;Y_1|X_1)$, which establishes the converse part of Theorem 2.3.1.

To achieve the capacity in Theorem 2.3.1, we need to employ the decode-and-forward coding strategy. Briefly, this strategy involves a combination of random coding, Slepian-Wolf binning, superposition coding and block Markov encoding at the relay and source and we provide an outline as follows:

- Random coding and binning: First randomly generate 2^{nR_0} i.i.d. sequences according to $p(x_1^n) = \Pi_{i=1}^n p(x_{1i})$, indexed as $x_1^n(s),\ s \in [1, 2^{nR_0}]$ and for each $x_1^n(s)$, generate 2^{nR} conditionally independent sequences $x^n(w|s),\ w \in [1, 2^{nR}]$ according to $p(x^n|x_1^n(s)) = \Pi_{i=1}^n p(x_i|x_{1i}(s))$. Then randomly distribute the indexes $1, \cdots, 2^{nR}$ to 2^{nR_0} bins $S_1, \cdots, S_{2^{nR_0}}$ such that each message index w is corresponding to a bin index s, i.e., contained in the bin S_s.

- Encoding: At block i, let w_i be the new index to be sent and assume $w_{i-1} \in S_{s_i}$. The source sends $x^n(w_i|s_i)$ while the relay estimates w_{i-1} by $\hat{\hat{w}}_{i-1}$ and sends $x_1^n(\hat{\hat{s}}_i)$ assuming $\hat{\hat{w}}_{i-1} \in S_{\hat{\hat{s}}_i}$.

- Decoding: At the end of block i, the decoding is implemented as follows:

  1. Upon estimating s_i by $\hat{\hat{s}}_i$ and receiving $y_1^n(i)$, the relay claims that the message $\hat{\hat{w}}_{i-1} = w$ is sent iff there exists a unique w such that $(x^n(w|\hat{\hat{s}}_i), y_1^n, x_1^n(\hat{\hat{s}}_i))$ are jointly typical. This decoding error probability can be arbitrarily small if $R < I(X;Y_1|X_1)$.
  2. Upon receiving $y^n(i)$, the receiver claims that the message $\hat{s}_i = s$ is sent iff there exists a unique s such that $(y^n, x_1^n(s))$ are jointly typical. This decoding error probability can be arbitrarily small if $R_0 < I(X_1;Y)$.
  3. The receiver calculates his ambiguity set $\mathcal{L}(y^n(i-1))$ consisting of all w_{i-1} such that $(x^n(w_{i-1}|\hat{s}_{i-1}), y^n(i-1), x_1^n(\hat{s}_{i-1}))$ are jointly typical. Assuming that s_i is decoded successfully, the receiver claims that $\hat{w}_{i-1} = w$ is sent iff there exists a unique $w \in S_{s_i} \cap \mathcal{L}(y^n(i-1))$. This decoding error probability can be arbitrarily small if $R < I(X;Y|X_1) + R_0$. Obviously, the receiver is always one block behind. In B blocks of transmission, a sequence of $B-1$ indices will be sent, resulting in the actual rate $R(B-1)/B$ is arbitrarily close to R as $B \to \infty$. Combining all the above, we have $R < \min\{I(X;Y_1|X_1), I(X_1;Y) + I(X;Y|X_1)\}$, which makes the achievability part of Theorem 2.3.1.

□

Remark 2.3.1. *Note that although the above decode-and-forward strategy is used to achieve the capacity of degraded relay channel, it can apply to arbitrary relay channel. The difference is that the degradedness property is needed to justify that the maximum achievable rate by decode-and-forward is indeed the capacity, while in the case of general relay channel no such claim can be made.*

2.3.2 Compress-and-Forward

We are now in a position to discuss the nature of the general relay channel, whose capacity, unfortunately, is undetermined yet. However, an achievable rate is proposed based on compress-and-forward strategy:

Theorem 2.3.2. *The rate R^* is achievable for any discrete memoryless relay channel, where*

$$R^* = \sup I(X; \hat{Y}_1, Y|X_1) \tag{2.29}$$

subject to the constraint

$$I(X_1; Y) \geq I(Y_1; \hat{Y}_1|Y, X_1) \tag{2.30}$$

where the supremum is over all joint distributions

$$p(x, x_1, y, y_1, \hat{y}_1) = p(x)p(x_1)p(y, y_1|x, x_1)p(\hat{y}_1|x_1, y_1) \tag{2.31}$$

and $\hat{Y}_1$ has a finite range.

Outline of proof. Still, a block Markov Encoding is used, i.e., at the end of block i, the x_1 information is used to resolve the uncertainty of the receiver about w_{i-1}.

- Random coding and binning: Randomly generate 2^{nR_0} sequences according to $p(x_1^n) = \Pi_{i=1}^n p(x_{1i})$, indexed as $x_1^n(s),\ s \in [1, 2^{nR_0}]$ and 2^{nR} sequences according to $p(x^n) = \Pi_{i=1}^n p(x_i)$, indexed as $x^n(w),\ w \in [1, 2^{nR}]$. For each $x_1^n(s)$, generate $2^{n\hat{R}}$ sequences according to $p(\hat{y}_1^n|x_1^n(s)) = \Pi p(\hat{y}_{1i}|x_{1i}(s))$, where $p(\hat{y}_1|x_1) = \Sigma_{x,y_1,y} p(\hat{y}_1|x_1, y_1)p(x)p(y_1, y|x, x_1)$, indexed as $\hat{y}_1^n(z|s),\ z \in [1, 2^{n\hat{R}}],\ s \in [1, 2^{nR_0}]$. Then randomly distribute the indexes $1, \cdots, 2^{n\hat{R}}$ to 2^{nR_0} bins $S_1, \cdots, S_{2^{nR_0}}$.

- Encoding: At block i, let w_i be the new index to be sent and assume

  $$(\hat{y}_1^n(z_{i-1}|s_{i-1}), y_1^n(i-1), x_1^n(s_{i-1}))$$

  are jointly typical and $z_{i-1} \in S_{s_i}$. The codeword pair $(x^n(w_i), x_1^n(s_i))$ are sent.

- Decoding: At the end of block i, the decoding is implemented as follows:

  1. Upon receiving $y^n(i)$, the receiver claims that the message $\hat{s}_i = s$ is sent iff there exists a unique s such that $(y^n, x_1^n(s))$ are jointly typical. This decoding error probability can be arbitrarily small if $R_0 < I(X_1; Y)$.
  2. The receiver calculates a set $\mathcal{L}(y^n(i-1))$ consisting of all z such that $(\hat{y}_1^n(z|\hat{s}_{i-1}), x_1^n(\hat{s}_{i-1}), y^n(i-1))$ are jointly typical. The receiver claims that z_{i-1} is sent in block $i-1$ if $\hat{z}_{i-1} \in S_{s_i} \cap \mathcal{L}(y^n(i-1))$. This decoding error probability can be arbitrarily small if $\hat{R} < I(\hat{Y}_1; Y|X_1) + R_0$.

3. The receiver declares that $\hat{w}_{i-1}$ was sent in block $i-1$ if

$$(x^n(\hat{w}_{i-1}), \hat{y}_1^n(\hat{z}_{i-1}|\hat{s}_{i-1}), x_1^n(\hat{s}_{i-1}), y^n(i-1))$$

are jointly typical. This decoding error probability can be arbitrarily small if $R < I(X; \hat{Y}_1, Y|X_1)$.

4. Upon receiving $y_1^n(i)$, the relay declares that z is "received" if

$$(\hat{y}_1^n(z|\hat{s}_i), x_1^n(\hat{s}_i), y_1^n(i))$$

are jointly typical. There will exist such a z if $\hat{R} > I(Y_1; \hat{Y}_1|X_1)$. Combining all the above, we obtain the constraint $I(X_1; Y) \geq I(Y_1; \hat{Y}_1|Y, X_1)$.

□

Chapter 3

A Geometric Approach to the CDOCI for Binary Symmetric Channel

3.1 Modeling with Binary Symmetric Channel and Results

Assume that both the channel 1–2 and 1–3 in Figure 1.3 are binary symmetric channels as shown in Figure 1.1, with crossover probability p and q respectively. We do such an assumption because the BSC, on one hand, is the simplest model of noisy channel and easy for analysis, but still captures most of the complexity of the general problem on the other hand. Since the main concern in this thesis is about the relay's behavior, we will focus on the channel 1–2 and directly analyze the single user binary symmetric channel depicted in Figure 1.1 instead. Without loss of generality, we assume the error probability $p < \frac{1}{2}$ in the following discussion.

To transmit at a rate R, consider a *random codebook* generated by selecting a distribution $p(x)$ on the input alphabet $\mathcal{X} = \{0,\ 1\}$ and generating 2^{nR} *i.i.d.* random codewords. To accurately formulate the problem, we use the following notation.

Notation 3.1.1. *The n-dimensional signal space for the n-used binary symmetric channel is defined as*

$$\begin{aligned}\mathcal{A}^n := \{&a^n(1) = 000\cdots00, a^n(2) = 000\cdots01, \cdots,\\ &a^n(2^n-1) = 111\cdots10, a^n(2^n) = \underbrace{111\cdots11}_{n\ bits}\},\end{aligned} \tag{3.1}$$

where each element in $\mathcal{A}^n$ consists of n bits.

Note that both the input and output sequences for the n-used binary symmetric channel exist in this discrete n-dimensional signal space.

Notation 3.1.2. *The codebook corresponding to rate R under n-dimensional signal space for the binary symmetric channel is defined as*

$$\mathcal{C}^{(n,R)} := \{X^n(w) \in \mathcal{A}^n,\ w = 1, \cdots, 2^{nR}\}, \tag{3.2}$$

where each of the 2^{nR} sequences in $\mathcal{C}^{(n,R)}$ represents a codeword of length n, randomly generated according to the distribution,

$$p(x^n) = \prod_{i=1}^{n} p(x_i). \tag{3.3}$$

From the classical information theory [5], we readily get the conclusion that the capacity of binary symmetric channel is $C = 1 - H(p)$, and it is achieved when $p_X(0) = p_X(1) = \frac{1}{2}$. Using the above notation, we have a vivid geometric interpretation of capacity and the method of *typical set decoding* for binary symmetric channel. Consider the following sequence of events, all of which happen in the n-dimensional signal space $\mathcal{A}^n$:

1. Choose 2^{nR} sequences from $\mathcal{A}^n$ to form the codebook. As mentioned above, $p_X(0) = p_X(1) = \frac{1}{2}$ is required to achieve the capacity. In words, from the geometric point of view, the capacity is achieved only if we *unbiasedly* choose 2^{nR} codewords *at random* from the signal space $\mathcal{A}^n$ to form the codebook, which means that each sequence in $\mathcal{A}^n$ is equally likely to become one codeword in the codebook. Formally, to achieve the capacity of Binary Symmetric Channel, the codebook $\mathcal{C}^{(n,R)}$ should be so formed that $Pr(X^n(w) = a^n(i)) = \frac{1}{2^n}$, for any $w \in \{1, \cdots, 2^{nR}\}$ and any $i \in \{1, \cdots, 2^n\}$. For the purpose of achieving capacity, self-evidently, we generate the codebook in this way throughout this chapter.

2. The codebook is then revealed to both sender and receiver. Equivalently, both the source and destination now know which sequences in $\mathcal{A}^n$ are chosen as codewords to form the codebook so as to be potential to send.

3. The sender X picks a message W uniformly, i.e., $Pr(W = w) = 2^{-nR}$ for any $w \in \{1, \cdots, 2^{nR}\}$, and transmits the corresponding codeword $x^n(W)$.

4. Over the channel, the transmitted codeword is "mapped" into another sequence Y^n in $\mathcal{A}^n$, according to the distribution:

$$p(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)). \tag{3.4}$$

5. Upon receiving Y^n, generally, decoder Y forms the *ϵ-jointly typical set*,

$$A_\epsilon(Y^n) := \left\{w : x^n(w) \in \mathcal{C}^{(n,R)}, (x^n(w), Y^n) \text{ are } \epsilon\text{-jointly typical}\right\}. \tag{3.5}$$

If the jointly typical set has essentially only one codeword in the sense that

$$\lim_{n\to\infty} E_{\mathcal{C}^{(n,R)}}(Pr(|A_\epsilon(Y^n)| = 1|\mathcal{C}^{(n,R)})) = 1 \tag{3.6}$$

for any small ϵ, then with high probability Y can decode correctly when the block length is sufficiently large and the corresponding rate R is said to be achievable. The capacity is defined as the supremum of all achievable rates.

Particularly, under the case of BSC, it can be shown that the decoder can form the jointly typical set geometrically by searching the codewords whose *Hamming Distance* with Y^n is between $n(p-\epsilon)$ and $n(p+\epsilon)$. Specifically, we have the following proposition.

Proposition 3.1.1. *For the Binary Symmetric Channel, decoder Y can form ϵ-jointly typical set based on its observation and codebook by*

$$A_\epsilon(Y^n) := \left\{w : x^n(w) \in \mathcal{C}^{(n,R)}, n(p-\epsilon_1) \le d_H(x^n(w), Y^n) \le n(p+\epsilon_1)\right\} \tag{3.7}$$

where $d_H(\cdot,\cdot)$ denotes the Hamming Distance, which is the number of bits where two sequences differ from each other, and $\epsilon_1 \to 0$ as $\epsilon \to 0$

Proof. According to (3.5), the ϵ-jointly typical set consists of every codeword x^n satisfying the constraint

$$\left|\frac{1}{n}\log p(Y^n|x^n) + H(Y|X)\right| \le \epsilon \tag{3.8}$$

$$\overset{(a)}{\iff} \left|\frac{1}{n}\log p^{d_H(x^n(w),Y^n)}(1-p)^{n-d_H(x^n(w),Y^n)} + H(Y|X)\right| \le \epsilon \tag{3.9}$$

$$\overset{(b)}{\iff} \left|\frac{1}{n}\log p^{d_H(x^n(w),Y^n)}(1-p)^{n-d_H(x^n(w),Y^n)} + H(p)\right| \le \epsilon \tag{3.10}$$

$$\overset{(c)}{\iff} \left|\frac{1}{n}\left[d_H \log p + (n-d_H)\log(1-p)\right] - p\log p - (1-p)\log(1-p)\right| \le \epsilon \tag{3.11}$$

$$\overset{(d)}{\iff} \left|\left(\frac{d_H}{n} - p\right)\log\frac{p}{1-p}\right| \le \epsilon \tag{3.12}$$

$$\overset{(e)}{\iff} n\left(p - \frac{\epsilon}{|\log\frac{p}{1-p}|}\right) \le d_H \le n\left(p + \frac{\epsilon}{|\log\frac{p}{1-p}|}\right) \tag{3.13}$$

where
"(a)" holds since $p(y^n|x^n(w)) = \prod_{i=1}^n p(y_i|x_i(w))$ and we calculate this product of probabilities by introducing Hamming distance and categorizing the n terms in the product into two classes: flipped bits(corresponding to $p^{d_H(x^n(w),Y^n)}$) and unflipped bits(corresponding to $(1-p)^{n-d_H(x^n(w),Y^n)}$);
"(b)" follows from the fact that $H(Y|X) = H(Y \oplus X|X) = H(p)$;

"(c)" follows from extending terms according to the definition of entropy;
"(d)" follows after some simple calculations;
"(e)" gives the explicit constraint to the Hamming distance.
Let $\epsilon_1 = \frac{\epsilon}{|\log \frac{p}{1-p}|}$ and notice that $\epsilon_1 \to 0$ as $\epsilon \to 0$, which completes the proof. □

Now, we are interested in the following conditional probabilities,

$$Pr(Y^n = a^n(i)|\mathcal{C}^{(n,R)}), \text{for any } i \in \{1, 2, \cdots, 2^n\},$$

especially when coding rate R is greater than C. These probabilities are random variables depending on the random codebook $\mathcal{C}^{(n,R)}$ and they are of pivotal importance in the sense that when n is sufficiently large, if there exists some kinds of codebooks with non-negligible probability, such that given these codebooks the above conditional probabilities of the 2^n sequences in $\mathcal{A}^n$ to be output is not uniformly distributed, then we may do better than the compress-and-forward strategy! Equivalently, if we utilize the knowledge which codebook is used at sender to figure out that the conditional distribution of channel output is not uniform, then we should use this *posterior* conditional distribution to do the quantization and Wyner-Ziv coding. However, if the posterior conditional distribution is uniform for sufficiently large n, there is no gain to exploit the codebook information and the compress-and-forward strategy is asymptotically optimal in this sense. In fact, when R is greater than C, we have the following surprising result.

For simplicity, we say $f(n)$ asymptotically equivalent to $g(n)$, denoted by $f(n) \sim g(n)$ if for any small $\epsilon > 0$,

$$g(n)(1-\epsilon) < f(n) < g(n)(1+\epsilon) \tag{3.14}$$

for sufficiently large n, i.e.,

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1 \tag{3.15}$$

or

$$f(n) = g(n) + o(g(n)). \tag{3.16}$$

Let U_i denote the event that $Pr(Y^n = a^n(i)|\mathcal{C}^{(n,R)}) \sim \frac{1}{2^n}$, then obviously $\bigcap_{i=1}^{2^n} U_i$ represents that the conditional distribution of channel output given codebook information is asymptotically uniform.

Theorem 3.1.1. *For any small $\delta > 0$,*

$$Pr\left(\bigcap_{i=1}^{2^n} U_i\right) > 1 - \delta, \tag{3.17}$$

when R is greater than C and n is sufficiently large.

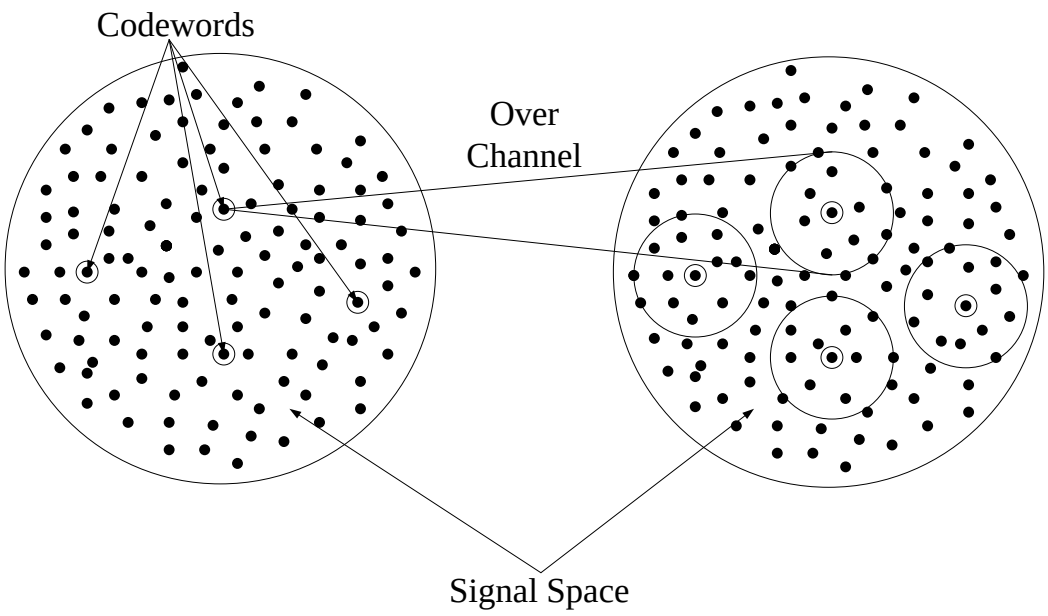

Figure 3.1: Low density of codewords and nonuniform distribution of output.

3.2 Proof via a Geometric Approach

At a first glance, theorem 3.1.1 seems contradictory to our intuition. Intuitively, over the channel, different codebook structures might result in different distributions of output sequences, and seemingly the distribution could never be always uniform for sufficiently large n. Indeed the intuition is right under some certain circumstance, and just consider the case where $R < C$. Obviously, for any $\epsilon > 0$, the set of sequences having Hamming distance less than $n(p + \epsilon)$ with a codeword is mostly likely mapped into from that codeword (precisely with high probability when n is sufficiency large). So intuitively, when $R < C$, as shown in Figure 3.1, since the density of codewords scattered in $\mathcal{A}^n$ is so low that many sequences can hardly be mapped into from these codewords and hence the posterior conditional distribution is not uniform.

However, when $R > C$, the intuition is not true any longer and theorem 3.1.1 holds now. Note here the capacity C plays a role of good threshold again, as in the classical channel coding theorem [1]. We show theorem 3.1.1 via a geometric approach by two steps:

- Firstly, we define a class of codebooks as *typical codebooks*, where the 2^{nR} codewords are approximately uniformly scattered over the n-dimensional signal space $\mathcal{A}^n$ with density $\frac{2^{nR}}{2^n}$. Using the powerful tool Vapnik-Chervonekis Theorem [16], [17], we show the typical codebook appears with high probability for sufficiently large n.

- Then, the asymptotically uniform distribution of output is shown when typical

codebook is used at the sender, and this is sufficient as a proof of theorem 3.1.1 since the typical codebooks are with high probability when n is sufficiently large.

3.2.1 Typical Codebooks for BSC

To define the typical codebook and show it accounts for high probability, we need to recall some relevant definitions and Vapnik-Chervonekis Theorem:

Definition 3.2.1. *A Range Space is a pair $(X, \mathcal{F})$, where X is a set and $\mathcal{F}$ is a family of subsets of X.*

Definition 3.2.2. *For any $A \subseteq X$, we define $P_{\mathcal{F}}(A)$, the projection of $\mathcal{F}$ on A, as $\{F \cap A : F \in \mathcal{F}\}$.*

Definition 3.2.3. *We say that A is* shattered *by $\mathcal{F}$ if $P_{\mathcal{F}}(A) = 2^A$, i.e., if the projection of $\mathcal{F}$ on A includes all possible subsets of A.*

Definition 3.2.4. *The VC-dimension of $\mathcal{F}$, denoted by VC-d($\mathcal{F}$) is the cardinality of the largest set A that $\mathcal{F}$ shatters. If arbitrarily large finite sets are shattered, the VC dimension of $\mathcal{F}$ is infinite.*

Theorem 3.2.1 (The Vapnik-Chervonekis Theorem)**.** *If $\mathcal{F}$ is a set of finite VC-dimension and $\{Y_j\}$ is a sequence of n i.i.d. random variables with common probability distribution P, then for every ϵ, $\delta > 0$*

$$Prob\left\{\sup_{F\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n} I(Y_j \in F) - P(F)\right| \leq \epsilon\right\} > 1 - \delta \tag{3.18}$$

whenever

$$n > \max\left\{\frac{8\,VC\text{-}d(\mathcal{F})}{\epsilon}\log_2\frac{16e}{\epsilon}, \frac{4}{\epsilon}\log_2\frac{2}{\delta}\right\} \tag{3.19}$$

Proof. The proof of theorem 3.2.1 is referred to [16], [17] and omitted here, and some applications can be found in [18], [19] and [20]. □

The following notations are useful for us to utilize the above tool and define the typical codebooks:

Notation 3.2.1.

$$F_s(i) := \{a^n \in \mathcal{A}^n : d_H(a^n, a^n(i)) = s\} \tag{3.20}$$

where $i \in \{1, \cdots, 2^n\}, s \in \{0, \cdots, n\}$.

Notation 3.2.2.

$$\mathcal{F}_s := \{F_s(i),\ i \in \{1, \cdots, 2^n\}\} \tag{3.21}$$

where $s \in \{0, \cdots, n\}$.

Geometrically, $F_s(i)$ consists of all the sequences on the surface of the *Hamming sphere*, whose center is $a^n(i)$ and radius is Hamming distance s, and $\mathcal{F}_s$ is the set of all $F_s(i)$, for $i \in \{1, \cdots, 2^n\}$.

Note that by definition 3.2.1, the pair $(\mathcal{A}^n, \mathcal{F}_s)$ is a range space, for any $s \in \{0, \cdots, n\}$, and we have the following lemma.

Lemma 3.2.1. *For a fixed dimension n, VC-d$(\mathcal{F}_s) \leq n$, for any $s \in \{0, \cdots, n\}$.*

Proof. Assume VC-d$(\mathcal{F}_s) > n$, then there exists a set $A' \subseteq \mathcal{A}^n$ such that $|\{F_s \cap A', F_s \in \mathcal{F}_s\}| = 2^{\text{VC-d}(\mathcal{F}_s)} > 2^n$. However, since $|\mathcal{F}_s| = 2^n$, obviously $|\{F_s \cap A, F_s \in \mathcal{F}_s\}| \leq 2^n$, for any $A \subseteq \mathcal{A}^n$, which is contradictory with the assumption and hence lemma 3.2.1 is proved. □

Remark 3.2.1. *Although the bound in lemma 3.2.1 seems loose, it is a sufficiently good upper bound for our needs, as we will show soon.*

Now, we are in a position to define the typical codebook set and show it is with high probability when n is sufficiently large. The typical codebook for the n-used binary symmetric channel is defined as the codebook in the way such that from the view of any specific sequence $a^n(i)$ in $\mathcal{A}^n$, the number of codewords which has Hamming distance s with $a^n(i)$ in a typical codebook is approximately proportional to the total number of all the sequences with Hamming distance s far away from $a^n(i)$. To accurately describe it, let $N(i, s|\mathcal{C}^{(n,R)})$ denote the number of codewords which have Hamming distance s with sequence $a^n(i)$ given the codebook $\mathcal{C}^{(n,R)}$.

Definition 3.2.5.

$$T(\mathcal{C}^{(n,R)}) := \left\{ \mathcal{C}^{(n,R)} : \sup_{F_s(i) \in \mathcal{F}_s} \left| \frac{N(i, s|\mathcal{C}^{(n,R)})}{2^{nR}} - \frac{\binom{n}{s}}{2^n} \right| \leq \frac{\Delta_s nR}{2^{nR}}, \textit{for any } s \in \{0, \cdots, n\} \right\} \tag{3.22}$$

where $\Delta_s = \max\{8\,\textit{VC-d}(\mathcal{F}_s), 16e\}$.

Theorem 3.2.2. *For a binary symmetric channel, generate the codebook $\mathcal{C}^{(n,R)}$ at random according to the distribution $p_X(0) = p_X(1) = \frac{1}{2}$, then*

$$Pr(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \to 1 \textit{ as } n \to \infty. \tag{3.23}$$

Note that the typical codebook implicates a good uniform geometric distribution of codewords from the view of any sequence in $\mathcal{A}^n$, therefore, to prove Theorem 3.2.2, we need to consider the probability of a class of 2^n simultaneously happening events, which invokes a sort of uniform convergence (in probability) of law of large numbers over the entire class. To achieve this, we resort to the Vapnik-Chervonekis Theorem. And fortunately, as lemma 3.2.1 shows, for a fixed dimension n, the VC-d$(\mathcal{F}_s)$ is finite, which is a sufficient condition for the uniform convergence in the weak law of large numbers. The formal proof is as following:

Proof. Consider all the codewords in a random codebook $\{X^n(w), w \in \{1, \cdots, 2^{nR}\}\}$. Notice that the 2^{nR} codewords are actually a sequence of random variables with the common distribution $P(X^n = a^n(i)) = 1/2^n$, for any $i \in \{1, \cdots, 2^n\}$.

Now consider the range space $(\mathcal{A}^n, \mathcal{F}_s)$. According to the definition of $F_s(i)$, we easily get $P(X^n \in F_s(i)) = \frac{\binom{n}{s}}{2^n}$. Recall that $F_s(i)$ represents the surface of the Hamming sphere with center at $a^n(i)$ and radius s, and then the intuition is clear: since we choose every codeword uniformly and randomly over the signal space, the probability that a codeword is on a specific Hamming sphere's surface is proportional to its volume $\binom{n}{s}$.

Since the VC-d($\mathcal{F}_s$) is finite for a fixed dimension n, we are justified to use the Vapnik-Chervonekis theorem. To satisfy (3.19), let $\epsilon = \delta = \frac{\Delta_s nR}{2^{nR}}$, then the Vapnik-Chervonekis theorem states that

$$P\left\{\sup_{F_s(i)\in\mathcal{F}_s}\left|\frac{\sum_{w=1}^{2^{nR}} I(X^n(w) \in F_s(i), X^n(w) \in \mathcal{C}^{(n,R)})}{2^{nR}} - \frac{\binom{n}{s}}{2^n}\right| \leq \frac{\Delta_s nR}{2^{nR}}\right\} > 1 - \frac{\Delta_s nR}{2^{nR}} \tag{3.24}$$

where $\Delta_s = \max\{8\text{VC-d}(\mathcal{F}_s), 16e\}$.

Observing that

$$N(i, s|\mathcal{C}^{(n,R)}) = \sum_{w=1}^{2^{nR}} I(X^n(w) \in F_s(i), X^n(w) \in \mathcal{C}^{(n,R)}), \tag{3.25}$$

we have

$$Pr\left\{\sup_{F_s(i)\in\mathcal{F}_s}\left|\frac{N(i, s|\mathcal{C}^{(n,R)})}{2^{nR}} - \frac{\binom{n}{s}}{2^n}\right| \leq \frac{\Delta_s nR}{2^{nR}}\right\} > 1 - \frac{\Delta_s nR}{2^{nR}} \tag{3.26}$$

where $\Delta_s = \max\{8\text{VC-d}(\mathcal{F}_s), 16e\}$.

Define

$$E_s := \left\{\sup_{F_s(i)\in\mathcal{F}_s}\left|\frac{N(i, s|\mathcal{C}^{(n,R)})}{2^{nR}} - \frac{\binom{n}{s}}{2^n}\right| > \frac{\Delta_s nR}{2^{nR}}\right\}, \forall s \in \{0, \cdots, n\}, \tag{3.27}$$

and then we have

$$Pr\{\mathcal{C}^{(n,R)} \notin T(\mathcal{C}^{(n,R)})\} \tag{3.28}$$

$$= Pr\left(\bigcup_{s=0}^{n} E_s\right) \tag{3.29}$$

$$\leq \sum_{s=0}^{n} Pr(E_s) \tag{3.30}$$

$$\leq \sum_{s=0}^{n} \frac{\Delta_s nR}{2^{nR}} \tag{3.31}$$

$$\leq \frac{(n+1)\Delta nR}{2^{nR}} \tag{3.32}$$

where $\Delta := \max\{8n, 16e\}$ according to lemma 3.2.1.

Hence,

$$Pr\{\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\} \to 1, \text{ as } n \to \infty, \tag{3.33}$$

and this completes the proof of theorem 3.2.2. □

3.2.2 Asymptotically Uniform CDOCI for BSC

To prove theorem 3.1.1, the following two lemmas are useful.

Lemma 3.2.2. *Letting x^n be any given channel input for the BSC, and Y^n be the output, we have, for any small $\epsilon > 0$, $Pr(d_H(x^n, Y^n) \in [n(p-\epsilon), n(p+\epsilon)]) \to 1$ as $n \to \infty$.*

Proof. Note that the binary symmetric channel can be represented as

$$Y = X \oplus N, \tag{3.34}$$

where N is the noise random variable with distribution $p_N(1) = p$ and $p_N(0) = 1-p$. Readily, we get $\mathbf{E}[N] = p$ and $Var(N) = p(1-p)$.

For any small $\epsilon > 0$,

$$Pr(d_H(x^n, Y^n) \in [n(p-\epsilon), n(p+\epsilon)]) \tag{3.35}$$

$$= Pr\left(\sum_{i=1}^{n} N_i \in [n(p-\epsilon), n(p+\epsilon)]\right) \tag{3.36}$$

$$= Pr\left(\left|\frac{\sum_{i=1}^{n} N_i}{n} - p\right| \leq \epsilon\right) \tag{3.37}$$

$$\overset{(a)}{\geq} 1 - \frac{Var(N)}{n\epsilon^2} \tag{3.38}$$

$$\to 1, \text{ as } n \to \infty, \tag{3.39}$$

where "(a)" follows from Chebyshev's Inequality and this finishes the proof. □

Remark 3.2.2. *For simplicity, from now on, we use H_ϵ to denote the event that $d_H(x^n, Y^n) \in [n(p-\epsilon), n(p+\epsilon)]$ for any given x^n. Note that although the $d_H(x^n, Y^n)$ is a quantity related to x^n, it essentially describes the effect of channel noise, and whether the event H_ϵ happens is independent of which specific sequence x^n is. This observation is important for the calculation later.*

Lemma 3.2.3.

$$Pr\left(\bigcap_{i=1}^{2^n} U_i \middle| H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) = 1 \tag{3.40}$$

when R is greater than C and n is sufficiently large.

Proof. Firstly, consider $Pr(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}))$ for any $i \in \{1, \cdots, 2^n\}$. We will try to give both the tight lower bound and upper bound to this probability.

$$Pr(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{3.41}$$

$$\overset{(a)}{=} \sum_{w=1}^{2^{nR}} P(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), x^n(w) \text{ is sent})$$

$$\cdot P(x^n(w) \text{ is sent}|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{3.42}$$

$$\overset{(b)}{=} \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} P(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), x^n(w) \text{ is sent}) \tag{3.43}$$

$$\overset{(c)}{=} \frac{1}{2^{nR}} \sum_{d_H(x^n(w),Y^n)\in[n(p-\epsilon),n(p+\epsilon)]} P(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), x^n(w) \text{ is sent}) \tag{3.44}$$

$$\overset{(d)}{=} \frac{1}{2^{nR}} \sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \frac{N(i,s|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \cdot p^s(1-p)^{n-s}}{\sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \binom{n}{s} p^s(1-p)^{n-s}} \tag{3.45}$$

$$\overset{(e)}{\geq} \frac{1}{2^{nR}} \sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \frac{\left[\frac{2^{nR}\binom{n}{s}}{2^n} - \Delta_s nR\right] p^s(1-p)^{n-s}}{\sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \binom{n}{s} p^s(1-p)^{n-s}} \tag{3.46}$$

$$= \frac{1}{2^n} - \frac{\phi_2}{\phi_1}, \tag{3.47}$$

where

$$\phi_1 := \sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \binom{n}{s} p^s(1-p)^{n-s}, \tag{3.48}$$

$$\phi_2 := \frac{nR}{2^{nR}} \sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \Delta_s p^s(1-p)^{n-s}. \tag{3.49}$$

"(a)" follows from the *Law of Total Probability*. We do this kind of transformation since we want to accumulate the contributions from all the codewords in the codebook to the probability for $a^n(i)$ to be channel output.

"(b)" follows from the fact that W is uniformly distributed and the events H_ϵ and $\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})$ are independent of the choice of index W. Actually, the

event $\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})$ is just a description of the codebook and hence has nothing to do with which index to be sent. H_ϵ is also an independent event according to remark 3.2.2.

"(c)" holds because of the condition H_ϵ. Having known that $d_H(x^n, Y^n) \in [n(p-\epsilon), n(p+\epsilon)]$, it is justified to only count in the contributions from the effective codewords, which are those codewords have Hamming distance between $n(p-\epsilon)$ and $n(p+\epsilon)$ with Y^n.

"(d)" allows us to accumulate the contributions from the effective codewords *shell-by-shell*.

"(e)" follows from the definition of typical codebook.

For ϕ_1, consider a *Bernoulli* Random Variable Z with probability p to be 0 and $1-p$ to be 1. In this specific case, after some simple calculation, we can get the ϵ_1-typical sequence set

$$A^n_{\epsilon_1}(Z) = \{z^n : N(0|z^n) \in [n(p-\epsilon), n(p+\epsilon)]\}, \tag{3.50}$$

where $N(0|z^n)$ denotes the number of 0 in sequence z^n and $\epsilon \to 0$, as $\epsilon_1 \to 0$. Besides, according to AEP, for any small $\epsilon_1 > 0$,

$$Pr(Z^n \in A^n_{\epsilon_1}(Z)) \to 1, \text{ as } n \to 0. \tag{3.51}$$

Therefore, for any small $\epsilon > 0$

$$\sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} \binom{n}{s} p^s (1-p)^{n-s} = \sum_{z^n \in A^n_{\epsilon_1}(Z)} p(z^n) \to 1, \text{ as } n \to 0, \tag{3.52}$$

and $\phi_1 = 1 - o(1)$.

For ϕ_2, consider the following series of inequalities (or equations):

$$\begin{aligned}
\phi_2 &\overset{(a)}{\leq} \frac{nR}{2^{nR}} \Delta \sum_{s=n(p-\epsilon)}^{n(p+\epsilon)} p^s (1-p)^{n-s} && (3.53)\\
&\overset{(b)}{\leq} \frac{nR}{2^{nR}} \Delta \cdot 2n\epsilon \cdot p^{n(p-\epsilon)} (1-p)^{n-n(p-\epsilon)} && (3.54)\\
&\overset{(c)}{=} \frac{nR}{2^{nR}} \Delta \cdot 2n\epsilon \cdot p^{np} (1-p)^{n-np} \cdot p^{-n\epsilon} (1-p)^{n\epsilon} && (3.55)\\
&\overset{(d)}{=} \frac{nR}{2^{nR}} \Delta \cdot 2n\epsilon \cdot 2^{-nH(p)} \cdot 2^{n\epsilon \log \frac{1-p}{p}} && (3.56)\\
&\overset{(e)}{=} \frac{nR \cdot \Delta \cdot 2n\epsilon}{2^{n(R+H(p)-\epsilon \log \frac{1-p}{p})}}, && (3.57)
\end{aligned}$$

where
"(a)" follows from lemma 3.2.1 and $\Delta = \max\{8n, 16e\}$;
"(b)" gives a upper bound to the summation by taking $s = n(p-\epsilon)$;

"(c)" follows from extending the terms;
"(d)" follows from the definition of entropy;
"(e)" gives an explicit expression to analyze the exponent.

Combining all the above, for any $i \in \{1, \cdots, 2^n\}$ and for any small $\epsilon > 0$,

$$Pr(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{3.58}$$

$$\geq \frac{1}{2^n} - \frac{\phi_2}{\phi_1} \tag{3.59}$$

$$\geq \frac{1}{2^n} - \frac{nR \cdot \Delta \cdot 2n\epsilon}{(1-o(1))2^{n(R+H(p)-\epsilon \log \frac{1-p}{p})}} \tag{3.60}$$

$$\geq \frac{1}{2^n}(1-o(1)) \tag{3.61}$$

as long as $R+H(p)-\epsilon \log \frac{1-p}{p} > 1$, i.e., $R > 1-H(p)+\epsilon'$, where $\epsilon' := \epsilon \log \frac{1-p}{p} > 0$. Note that ϵ' can be driven to 0 by letting $\epsilon \to 0$, and therefore, (3.61) holds as long as $R > C$ and n is sufficiently large.

Similarly, we can show that when $R > C$ and n is sufficiently large, for any $i \in \{1, \cdots, 2^n\}$,

$$Pr(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \leq \frac{1}{2^n}(1+o(1)) \tag{3.62}$$

and hence

$$Pr(Y^n = a^n(i)|H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \sim \frac{1}{2^n}. \tag{3.63}$$

Therefore, lemma 3.2.3 is proved. □

Until now, the proof to theorem 3.1.1 is obvious:

Proof of theorem 3.1.1. For any small $\delta > 0$, when $R > C$ and n is sufficiently large,

$$Pr\left(\bigcap_{i=1}^{2^n} U_i\right) \tag{3.64}$$

$$= Pr\left(\bigcap_{i=1}^{2^n} U_i \middle| \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) Pr\left(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right)$$

$$+ Pr\left(\bigcap_{i=1}^{2^n} U_i \middle| \mathcal{C}^{(n,R)} \notin T(\mathcal{C}^{(n,R)})\right) Pr\left(\mathcal{C}^{(n,R)} \notin T(\mathcal{C}^{(n,R)})\right) \tag{3.65}$$

$$\geq Pr\left(\bigcap_{i=1}^{2^n} U_i \middle| \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) Pr\left(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) \tag{3.66}$$

$$
\begin{aligned}
=&P\left(\bigcap_{i=1}^{2^n} U_i \middle| H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) P(H_\epsilon) Pr\left(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) \\
+&P\left(\bigcap_{i=1}^{2^n} U_i \middle| H_\epsilon^c, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) P(H_\epsilon^c) Pr\left(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) && (3.67) \\
\geq&P\left(\bigcap_{i=1}^{2^n} U_i \middle| H_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) P(H_\epsilon) P\left(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) && (3.68) \\
>&1 \cdot (1-\delta) \cdot (1-\delta) && (3.69) \\
=&1 - \delta_1 && (3.70)
\end{aligned}
$$

where $\delta_1 := 2\delta - \delta^2$. Noting that δ_1 can be driven to 0 by letting $\delta \to 0$, this completes the proof. □

Chapter 4

A General Characterization of the CDOCI for DMC and A Proof of the Strong Converse to Channel Coding Theorem

In this chapter, we focus on the Discrete Memoryless Channel. We give a general characterization of the CDOCI, calculate the rate needed for the relay to forward its observation, and present a strong converse to the channel coding theorem under the random coding framework. In section 4.1, we formally formulate the problem, and point out for those strongly typical output sequences, the CDOCI is asymptotically uniform and approaching the unconditional distribution with high probability when $R > C$ and n is sufficiently large. Then, we prove this result in section 4.2. In section 4.3 we confirm that there is no gain to utilize the codebook information in the sense that the rate needed for the relay to forward its observation doesn't decrease. Finally, in section 4.4 we show by the random coding technique, the error probability goes to 1 with high probability when $R > C$.

4.1 Problem Formulation for Discrete Memoryless Channel and Results

Consider a discrete memoryless channel defined by a conditional distribution $p(y|x)$ for $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, as depicted in Figure 1.2. To transmit at a rate R, consider a random codebook generated by selecting a distribution on the input alphabet, $p(x)$, and generating 2^{nR} i.i.d random codewords. Paralleling the content in Chapter 3, we use the following definitions and notations for DMC.

Notation 4.1.1. *The n-dimensional input signal space and output signal space for the n-used discrete memoryless channel are denoted as $\mathcal{X}^n$ and $\mathcal{Y}^n$ respectively.*

Definition 4.1.1. *The nth extension of the discrete memoryless channel (without feedback) is the channel* $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$*, where*

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i) \tag{4.1}$$

Notation 4.1.2. *The codebook corresponding to rate* R *for the nth extension of the discrete memoryless channel is defined as*

$$\mathcal{C}^{(n,R)} := \{X^n(w) \in \mathcal{X}^n,\ w = 1, \cdots, 2^{nR}\}, \tag{4.2}$$

where each of the 2^{nR} *sequences in* $\mathcal{C}^{(n,R)}$ *represents a codeword of length* n*, randomly generated according to the distribution,*

$$p(x^n) = \prod_{i=1}^{n} p(x_i). \tag{4.3}$$

For a fixed input distribution $p(x)$, we can get the output distribution

$$p(y) = \sum_{x \in \mathcal{X}} p(x, y) = \sum_{x \in \mathcal{X}} p(x)p(y|x) \tag{4.4}$$

and then we have the ϵ-strongly typical input and output sequence set for the n-used discrete memoryless channel, $A_\epsilon^{*(n)}(X)$ and $A_\epsilon^{*(n)}(Y)$, respectively. To facilitate the analysis, we arbitrarily order the elements in both ϵ-strongly typical sets and denote them by $x_\epsilon^n(i)$, for $i \in \{1, \cdots, L_\epsilon^{(n)}\}$ and $y_\epsilon^n(i)$, for $i \in \{1, \cdots, M_\epsilon^{(n)}\}$, where $L_\epsilon^{(n)} = |A_\epsilon^{*(n)}(X)|$ and $M_\epsilon^{(n)} = |A_\epsilon^{*(n)}(Y)|$.

Our result is that given $p(x)$, the CDOCI for the strongly typical output sequences is asymptotically uniform and approaching the unconditional distribution with high probability, when $R > C$ and the block length is sufficiently large.

We now define the same notation as that in [5] to express equality to the first order in the exponent.

Notation 4.1.3. *We say* a_n *and* b_n *are equal to the first order in the exponent, denoted by* $a_n \doteq b_n$*, if*

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0. \tag{4.5}$$

Using this notation, we can now restate our results as for $R > C$ and sufficiently large n,

$$p(y_\epsilon^n(i)|\mathcal{C}^{(n,R)}) \doteq p(y_\epsilon^n(i)) \doteq 2^{-nH(Y)}, \forall\ i \in \{1, \ldots, M_\epsilon^{(n)}\} \tag{4.6}$$

with high probability.

Formally, given $p(x)$, letting

$$U_i = \left\{p(y_\epsilon^n(i)|\mathcal{C}^{(n,R)}) \doteq 2^{-nH(Y)}\right\}, \text{for any } i \in \{1, \ldots, M_\epsilon^{(n)}\}, \tag{4.7}$$

we have the following theorem:

Theorem 4.1.1. *For any small $\delta > 0$*

$$Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} U_i\right) > 1 - \delta, \tag{4.8}$$

when n is sufficiently large and $R > C$.

4.2 A General Proof of the Uniform CDOCI

Similarly in Chapter 3, we prove theorem 4.1.1 by defining the typical codebooks and focus on the CDOCI given the typical codebooks. In 4.2.1, we define the typical codebook, and then show the typical codebooks are with high probability, while in 4.2.2, we finish the proof of theorem 4.1.1. Moreover, the further discussion on the uniform CDOCI is given in 4.2.3.

4.2.1 Typical Codebooks for DMC

To define the typical codebook set for the discrete memoryless channel and show it contains most of the probability, we need a fundamental result on the size of conditionally typical set, which is the the following lemma 4.2.1.

Lemma 4.2.1. *Given a joint distribution $p(x, y)$, for any y^n such that there exsits at least one pair $(x^n, y^n) \in A_\epsilon^{*(n)}$, the set of sequences x^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}$ satisfies*

$$2^{n(H(X|Y)-\epsilon_1)} \leq |\{x^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq 2^{n(H(X|Y)+\epsilon_1)}, \tag{4.9}$$

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

Proof. This lemma corresponds to equation (13.168) on page 372 in the Cover's book [5] except exchanging the position of x^n and y^n. See the outline of proof in [5] and the details are omitted here. □

Due to the above lemma, we have the following theorem, which bounds the probability that a randomly chosen sequence is jointly typical with a fixed typical sequence.

Theorem 4.2.1. *Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim p(x)$. For $y^n \in A_\epsilon^{*(n)}(Y)$, the probability that $(X^n, y^n) \in A_\epsilon^{*(n)}$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon')} \leq Pr((X^n, y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon')}, \tag{4.10}$$

where ϵ' goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Proof.

$$
\begin{aligned}
& Pr((X^n, y^n) \in A_\epsilon^{*(n)}) && (4.11)\\
= & \sum_{x^n \in \{x^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}} p(x^n) && (4.12)\\
\leq & \sum_{x^n \in \{x^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}} 2^{-n(H(X)-\epsilon_1)} && (4.13)\\
\leq & 2^{n(H(X|Y)+\epsilon_2)} 2^{-n(H(X)-\epsilon_1)} && (4.14)\\
= & 2^{-n(I(X;Y)-\epsilon^{'})} && (4.15)
\end{aligned}
$$

where ϵ_1, ϵ_2 and $\epsilon^{'}$ all go to 0 as $n \to \infty$. Similarly,

$$Pr((X^n, y^n) \in A_\epsilon^{*(n)}) \geq 2^{-n(I(X;Y)+\epsilon^{'})} \tag{4.16}$$

and hence

$$2^{-n(I(X;Y)+\epsilon^{'})} \leq Pr((X^n, y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon^{'})}. \tag{4.17}$$

□

To formally define the typical codebook, we use the following notation:

Notation 4.2.1.

$$
\begin{aligned}
F_\epsilon(i) &:= \left\{x^n \in A_\epsilon^{*(n)}(X) : y_\epsilon^n(i) \in A_\epsilon^{*(n)}(Y|x^n)\right\}, \forall\ i \in \{1, \cdots, M_\epsilon^{(n)}\}; && (4.18)\\
\mathcal{F}_\epsilon &:= \{F_\epsilon(i), i \in \{1, \cdots, M_\epsilon^{(n)}\}; && (4.19)\\
P_\epsilon(i) &:= P(X^n \in F_\epsilon(i)); && (4.20)\\
N_\epsilon(i|\mathcal{C}^{(n,R)}) &:= \sum_{w=1}^{2^{nR}} I(X^n(w) \in F_\epsilon(i), X^n(w) \in \mathcal{C}^{(n,R)}); && (4.21)\\
N_\epsilon^*(\mathcal{C}^{(n,R)}) &= \sum_{w=1}^{2^{nR}} I(X^n(w) \in A_\epsilon^{*(n)}, X^n(w) \in \mathcal{C}^{(n,R)}). && (4.22)
\end{aligned}
$$

Noting that $(\mathcal{X}^n, \mathcal{F}_\epsilon)$ forms a range space, we have the following definition of typical codebooks for discrete memoryless channel.

Definition 4.2.1.

$$T(\mathcal{C}^{(n,R)}) = \left\{ \mathcal{C}^{(n,R)} : \begin{array}{c} \sup\limits_{F_\epsilon(i) \in \mathcal{F}_\epsilon} \left| \dfrac{N_\epsilon(i|\mathcal{C}^{(n,R)})}{2^{nR}} - P_\epsilon(i) \right| \leq \dfrac{\Delta_\epsilon nR}{2^{nR}} \\ \left| \dfrac{N_\epsilon^*(\mathcal{C}^{(n,R)})}{2^{nR}} - P(X^n \in A_\epsilon^{*(n)}(X)) \right| \leq \dfrac{n}{2^{nR/2}} \end{array} \right\} \tag{4.23}$$

where $\Delta_\epsilon := \max\{8\, VC\text{-}d(\mathcal{F}_\epsilon), 16e\}$.

Theorem 4.2.2. *For a discrete memoryless channel, generate the codebook $\mathcal{C}^{(n,R)}$ at random according to the distribution $p(x)$, then*

$$Pr(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \to 1 \text{ as } n \to \infty. \tag{4.24}$$

Similar to the proof of Theorem 3.2.2, we need to employ the Vapnik-Chervonekis Theorem to show Theorem 4.2.2 holds and hence a finite VC dimension of $\mathcal{F}_\epsilon$ is desired. For this reason, we introduce a lemma first.

Lemma 4.2.2. *For a fixed dimension n, VC-d$(\mathcal{F}_\epsilon) \leq n(H(Y) + \epsilon^{'})$, where $\epsilon^{'}$ goes to 0 as $\epsilon \to 0$ and $n \to \infty$.*

Proof. According to the definition of $\mathcal{F}_\epsilon$ and the property of typical sequence set, we have

$$2^{n(H(Y)-\epsilon^{'})} \leq |\mathcal{F}_\epsilon| = |A_\epsilon^{*(n)}(Y)| \leq 2^{n(H(Y)+\epsilon^{'})}, \tag{4.25}$$

where $\epsilon^{'} \to 0$ as $\epsilon \to 0$ and $n \to \infty$. Using the same idea as the proof of Lemma 3.2.1, we can readily get this lemma. Note for a fixed n, VC-d$(\mathcal{F}_\epsilon)$ is finite, which ensures us be able to utilize the Vapnik-Chervonekis Theorem safely. □

Proof of Theorem 4.2.2. Noticing that a typical codebooks satisfies two constraints in Definition 4.2.1, we will prove Theorem 4.2.2 by showing a random codebook satisfies each constraint with high probability.

For the given $p(x)$, consider all the codewords in a random codebook $X^n(w)$, $w \in \{1, \cdots, 2^{nR}\}$, which are a sequence of random variables with the common distribution $p(x^n) = \Pi_{i=1}^n p(x_i)$. Employing the Vapnik-Chervonekis theorem under the range space $(\mathcal{X}^n, \mathcal{F}_\epsilon)$ and observing that

$$N_\epsilon(i|\mathcal{C}^{(n,R)}) = \sum_{w=1}^{2^{nR}} I(X^n(w) \in F_\epsilon(i), X^n(w) \in \mathcal{C}^{(n,R)}), \tag{4.26}$$

we have

$$Pr\left\{ \sup_{F_\epsilon(i) \in \mathcal{F}_\epsilon} \left| \frac{N_\epsilon(i|\mathcal{C}^{(n,R)})}{2^{nR}} - P_\epsilon(i) \right| \leq \frac{\Delta_\epsilon nR}{2^{nR}} \right\} > 1 - \frac{\Delta_\epsilon nR}{2^{nR}} \tag{4.27}$$

where $\Delta_\epsilon = \max\{8\text{VC-d}(\mathcal{F}_\epsilon), 16e\}$.

Therefore, letting $\Delta = \max\{8n(H(Y) + \epsilon^{'}), 16e\}$ according to lemma 4.2.2, it follows that

$$Pr\left\{ \sup_{F_\epsilon(i) \in \mathcal{F}_\epsilon} \left| \frac{N_\epsilon(i|\mathcal{C}^{(n,R)})}{2^{nR}} - P_\epsilon(i) \right| \leq \frac{\Delta_\epsilon nR}{2^{nR}} \right\} \tag{4.28}$$

$$\geq 1 - \frac{\Delta_\epsilon nR}{2^{nR}} \tag{4.29}$$

$$\geq 1 - \frac{\Delta nR}{2^{nR}} \tag{4.30}$$

$$\to 1 \text{ as } n \to \infty. \tag{4.31}$$

Until now, we prove a random codebook satisfies the first constraint of typical codebook with high probability and we consider the second constraint below. Let $B_\epsilon(w) = I(X^n(w) \in A_\epsilon^{*(n)}, X^n(w) \in \mathcal{C}^{(n,R)})$ for $w \in \{1, \cdots, 2^{nR}\}$, then $B_\epsilon(1), B_\epsilon(2), \cdots, B_\epsilon(2^{nR})$ are a sequence of i.i.d. random variables with common distribution:

$$B_\epsilon = \begin{cases} 1 & \text{if} \quad X^n \in A_\epsilon^{*(n)} \\ 0 & \text{if} \quad X^n \notin A_\epsilon^{*(n)} \end{cases} \tag{4.32}$$

Readily, we have $\mathbf{E}[B_\epsilon] = P(X^n \in A_\epsilon^{*(n)})$ and $Var(B_\epsilon) = P(X^n \in A_\epsilon^{*(n)})(1 - P(X^n \in A_\epsilon^{*(n)})) \leq 1$.

By Chebyshev's Inequality, we have for any $\delta > 0$,

$$P\left\{\left|\frac{\sum_{w=1}^{2^{nR}} B_\epsilon(w)}{2^{nR}} - P(X^n \in A_\epsilon^{*(n)})\right| \geq \delta\right\} \leq \frac{Var(B)}{2^{nR}\delta^2} \leq \frac{1}{2^{nR}\delta^2}. \tag{4.33}$$

Noting $N_\epsilon^*(\mathcal{C}^{(n,R)}) = \sum_{w=1}^{2^{nR}} B_\epsilon(w)$ and replacing the δ in (4.33) by $\frac{n}{2^{\frac{nR}{2}}}$, we have

$$P\left\{\left|N_\epsilon^*(\mathcal{C}^{(n,R)}) - P(X^n \in A_\epsilon^{*(n)})\right| \geq \frac{n}{2^{\frac{nR}{2}}}\right\} \leq \quad \frac{1}{2^{nR}(\frac{n}{2^{nR/2}})^2} = \frac{1}{n^2} \to 0, \text{ as } n \to \infty. \tag{4.34}$$

Combining (4.31) and (4.34), by union bound, we have

$$Pr(\mathcal{C}^{(n,R)} \notin T(\mathcal{C}^{(n,R)})) \tag{4.35}$$

$$\leq Pr\left\{\sup_{F_\epsilon(i) \in \mathcal{F}_\epsilon} \left|\frac{N_\epsilon(i|\mathcal{C}^{(n,R)})}{2^{nR}} - P_\epsilon(i)\right| \geq \frac{\Delta_\epsilon nR}{2^{nR}}\right\}$$

$$+ Pr\left\{\left|\frac{N_\epsilon^*(\mathcal{C}^{(n,R)})}{2^{nR}} - P(X^n \in A_\epsilon^{*(n)}(X))\right| \geq \frac{n}{2^{nR/2}}\right\} \tag{4.36}$$

$$\to 0 \quad \text{as} \quad n \to \infty, \tag{4.37}$$

which completes the proof of theorem 4.2.2. □

4.2.2 Asymptotically Uniform CDOCI for DMC

Notation 4.2.2. *We say a_n is greater than or equal to b_n to the first order in the exponent, denoted by $a_n \dot{\geq} b_n$, if*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} \geq 0; \tag{4.38}$$

and a_n is less than or equal to b_n to the first order in the exponent, denoted by $a_n \dot{\leq} b_n$, if

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0. \tag{4.39}$$

Remark 4.2.1. *By the above notation, we see that if $a_n \dot{\geq} b_n$ and $a_n \dot{\leq} b_n$, then $a_n \doteq b_n$, since*

$$0 \leq \lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0. \tag{4.40}$$

Theorem 4.2.3. *Given a typical codebook is used, assuming the codeword sent is ϵ-strongly typical $\sim p(x)$ and the noise introduced by channel is also "typical", i.e.,*

$$X^n \in A_\epsilon^{*(n)}(X) \text{ and } Y^n \in A_\epsilon^{*(n)}(Y|x^n) \text{ for any given } x^n, \tag{4.41}$$

then

$$Pr(Y^n = y_\epsilon^n(i)) \doteq 2^{-nH(Y)}, \forall i \in \{1, \ldots, M_\epsilon^{(n)}\}, \tag{4.42}$$

when $R > I(X;Y)$ and n is sufficiently large.

Remark 4.2.2. *In fact, the assumption that the codeword sent is a typical sequence with respect to $p(x)$ is not necessary to give rise to the uniform CDOCI. As we will see in 4.2.3, we can still get the uniform CDOCI with this assumption eliminated, but the range of R should be also modified accordingly.*

Proof. Let

$$C_\epsilon = \{Y^n \in A_\epsilon^{*(n)}(Y|x^n) \text{ for any given } x^n\}, \tag{4.43}$$

which can be interpreted as the event that the noise is "typical" and hence the output lies in the conditionally typical set given input x^n.

Consider $Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, X^n \in A_\epsilon^{*(n)}(X), \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}))$ for any $i \in \{1, \cdots, M_\epsilon\}$. We will try to give the lower bound to this probability.

To accumulate the contributions from all the codewords in the codebook to the probability for $y_\epsilon^n(i)$ to be channel output, we employ the Law of Total Probability and have the following equations(or inequalities):

$$Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, X^n \in A_\epsilon^{*(n)}(X), \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{4.44}$$

$$\overset{(a)}{=} \sum_{w=1}^{2^{nR}} P(Y^n = y_\epsilon^n(i)|C_\epsilon, X^n \in A_\epsilon^{*(n)}(X), \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), X^n = x^n(w))$$

$$\cdot P(X^n = x^n(w)|C_\epsilon, X^n \in A_\epsilon^{*(n)}(X), \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{4.45}$$

$$\overset{(b)}{=} \frac{1+o(1)}{2^{nR}} \sum_{x^n(w)\in A_\epsilon^{*(n)}} P(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), X^n = x^n(w)) \tag{4.46}$$

$$\overset{(c)}{\geq} \frac{1}{2^{nR}} \sum_{x^n(w)\in F_\epsilon(i)} P(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), X^n = x^n(w)) \tag{4.47}$$

$$\overset{(d)}{\geq}\frac{1}{2^{nR}}N_{\epsilon}(i|\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))\cdot 2^{-n[H(Y|X)+\epsilon^{''}]} \tag{4.48}$$

$$\overset{(e)}{\geq}\frac{1}{2^{nR}}(2^{nR}\cdot 2^{-n[I(X;Y)+\epsilon^{'}]}-\Delta_{\epsilon}nR)\cdot 2^{-n[H(Y|X)+\epsilon^{''}]} \tag{4.49}$$

$$\overset{(f)}{=}2^{-n[I(X;Y)+\epsilon^{'}]}\cdot 2^{-n[H(Y|X)+\epsilon^{''}]}-\frac{\Delta_{\epsilon}nR}{2^{nR}}\cdot 2^{-n[H(Y|X)+\epsilon^{''}]} \tag{4.50}$$

$$\overset{(g)}{=}2^{-n[H(Y)+\epsilon^{'}+\epsilon^{''}]}-\frac{\Delta_{\epsilon}nR}{2^{nR}}\cdot 2^{-n[H(Y|X)+\epsilon^{''}]} \tag{4.51}$$

$$\overset{(h)}{=}2^{-n[H(Y)+\epsilon^{'}+\epsilon^{''}]}\cdot\left\{1-\frac{\Delta_{\epsilon}nR}{2^{nR}}\cdot 2^{n[I(X;Y)+\epsilon^{'}]}\right\} \tag{4.52}$$

where

"(a)" follows from the Law of Total Probability.

"(b)" follows from the fact

$$\begin{aligned}&Pr(X^n=x^n(w)|X^n\in A_{\epsilon}^{*(n)}(X),C_{\epsilon},\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))\\&=\begin{cases}\dfrac{1}{N_{\epsilon}^{*}(\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))} & \text{for } x^n(w)\in A_{\epsilon}^{*(n)}(X)\\ 0 & \text{for } x^n(w)\notin A_{\epsilon}^{*(n)}(X)\end{cases}.\end{aligned} \tag{4.53}$$

Due to the definition of typical codebooks, we have

$$N_{\epsilon}^{*}(\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))=2^{nR}(1-o(1)) \tag{4.54}$$

and then (4.46) follows.

"(c)" allows us to only accumulate the contributions from those codewords in $F_{\epsilon}(i)$, which we call the codewords of interest. Actually, the codewords of interest are strongly jointly typical with $y_{\epsilon}^{n}(i)$ with respect to $p(x,y)$. This can be seen by observing that the codewords in $F_{\epsilon}(i)$ are strongly typical with respect to $p(x)$ and $y_{\epsilon}^{n}(i)$ is conditionally strongly typical with them. We will see that the contributions from the codewords of interest are sufficient to result in a tight lower bound of $p(y_{\epsilon}^{n}(i)|C_{\epsilon},X^n\in A_{\epsilon}^{*(n)},\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))$.

"(d)" follows from the definition of typical codebooks and gives a lower bound to the probability contribution from the codeword of interest. Specifically, This lower bound directly follows from the definition of weakly joint typicality, implied by the strongly joint typicality, and $\epsilon^{''}\to 0$, as $\epsilon\to 0$ and $n\to\infty$.

"(e)" follows from the definition of typical codebook and gives a lower bound to the number of codewords of interest, where $\epsilon^{'}\to 0$, as $\epsilon\to 0$ and $n\to\infty$.

"(f)" follows from the expansion of terms.

"(g)" follows from the basic definitions and relation of entropy and mutual information.

Therefore, for any $i\in\{1,\cdots,M_{\epsilon}^{n}\}$,

$$Pr(Y^n=y_{\epsilon}^{n}(i)|C_{\epsilon},X^n\in A_{\epsilon}^{*(n)},\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))\geq 2^{-n[H(Y)+\epsilon^{'}+\epsilon^{''}]}(1-o(1)) \tag{4.55}$$

as long as $R > I(X;Y) + \epsilon^{'}$. Since $\epsilon^{'} \to 0$, we have for $R > I(X;Y)$ and sufficiently large n,

$$Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \dot{\geq} 2^{-nH(Y)}, \forall\ i \in \{1, \ldots, M_\epsilon^{(n)}\}. \tag{4.56}$$

Similarly, we can derive a tight upper bound, i.e., for $R > I(X;Y)$ and sufficiently large n,

$$Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \dot{\leq} 2^{-nH(Y)}, \forall\ i \in \{1, \ldots, M_\epsilon^{(n)}\}, \tag{4.57}$$

and therefore

$$Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \doteq 2^{-nH(Y)}, \forall\ i \in \{1, \ldots, M_\epsilon^{(n)}\}, \tag{4.58}$$

which completes the proof of theorem 4.2.3. □

Now, we are in a position to prove Theorem 4.1.1.

Proof of Theorem 4.1.1. According to theorem 4.2.3, we readily have

$$Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} U_i \middle| C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) = 1. \tag{4.59}$$

Therefore,

$$\begin{aligned}
&Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} U_i\right) && (4.60)\\
\geq &Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} U_i, C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) && (4.61)\\
= &Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} U_i \middle| C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right)\\
&\cdot Pr\left(C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) && (4.62)\\
= &Pr\left(C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) && (4.63)\\
= &1 - Pr\left(\left(C_\epsilon, X^n \in A_\epsilon^{*(n)}, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right)^c\right) && (4.64)\\
\geq &1 - Pr\left(C_\epsilon^c\right) - Pr(X^n \in A_\epsilon^{*(n)}) - Pr\left(\mathcal{C}^{(n,R)} \notin T(\mathcal{C}^{(n,R)})\right) && (4.65)\\
\to &1 && (4.66)
\end{aligned}$$

as $n \to \infty$, which finishes the proof of Theorem 4.1.1. □

4.2.3 Further Discussion on the Uniform CDOCI

Note in the above analysis, the uniform CDOCI is obtained under the assumption that the codeword sent is strongly typical $\sim p(x)$. Actually, this assumption is a *technical* one in order to prove Theorem 4.1.1, because in reality, the randomly generated codebook may contain some nontypical sequences and hence the possibility that a nontypical codeword is sent always exists. Therefore, it is natural to ask: what if we eliminate the technical assumption and count in the effect of nontypical sequences? In fact, the answer is also positive, i.e., we can still get the uniform CDOCI even if all the sequences in the codebook are treated, except that a higher rate is needed.

Theorem 4.2.4. *Given a typical codebook is used, only assuming the noise introduced by channel is "typical", i.e.,*

$$Y^n \in A_\epsilon^{*(n)}(Y|x^n) \text{ for any given } x^n, \tag{4.67}$$

then

$$Pr(Y^n = y_\epsilon^n(i)) \doteq 2^{-nH(Y)}, \forall i \in \{1, \ldots, M_\epsilon^{(n)}\}, \tag{4.68}$$

when $R > C$ and n is sufficiently large.

Methods of Types and a Useful Lemma

To prove theorem 4.2.4, we need to slightly modify the definition of typical codebook. For this purpose, we use the method of types.

Definition 4.2.2. *The type P_{x^n} or (empirical probability distribution) of a sequence x^n is the relative proportion of occurrences of each symbol of $\mathcal{X}$, i.e., $P_{x^n}(a) = N(a|x^n)/n$, where $N(a|x^n)$ is the number of times the symbol a occurs in the sequence x^n. It is a probability mass function on $\mathcal{X}$.*

Definition 4.2.3. *Let $\mathcal{P}_n$ denote the set of types with denominator n. If $P \in \mathcal{P}_n$, then the set of sequences of length n and type P is called the type class of P, denoted as $T(P)$, i.e.,*

$$T(P) = \{x^n : P_{x^n} = P\}. \tag{4.69}$$

The power of the method of types essentially arises from the following theorem, which indicates that the number of types is at most polynomial with n.

Theorem 4.2.5.

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|} \tag{4.70}$$

Proof. Notice that each type is actually a vector consisting of $|\mathcal{X}|$ components and each component can at most take $n+1$ values, i.e., $0, 1/n, 2/n, \cdots, 1$. Therefore, there are at most $(n+1)^{|\mathcal{X}|}$ choices for the type vector, which gives a little loose but good enough upper bound for our needs, as we will show later. □

To redefine the typical codebook, we need the following additional lemma.

Lemma 4.2.3. *Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim p(x) = \Sigma_y p(x,y)$ and $A_{\epsilon,1}^{*(n)}$ denote the strongly typical set with respect to another joint distribution $p_1(x,y)$ having the marginal distribution $p(y)$, i.e., $\Sigma_x p_1(x,y) = p(y)$. For $y^n \in A_\epsilon^{*(n)}(Y)$, the probability that $(X^n, y^n) \in A_{\epsilon,1}^{*(n)}$ is bounded by*

$$Pr((X^n, y^n) \in A_{\epsilon,1}^{*(n)}) \leq 2^{-n(I_1(X;Y)-\epsilon')}, \tag{4.71}$$

where $I_1(X;Y)$ is calculated according to joint distribution $p_1(x,y)$ and ϵ' goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Proof. Let us first consider the conditional probability. For $y^n \in A_{\epsilon,1}^{*(n)}(Y)$, we have

$$\begin{aligned}
& Pr((X^n, y^n) \in A_{\epsilon,1}^{*(n)} | X^n \in A_{\epsilon,1}^{*(n)}) && (4.72)\\
= & \sum_{x^n \in \{x^n : (x^n, y^n) \in A_{\epsilon,1}^{*(n)}\}} p(X^n = x^n | X^n \in A_{\epsilon,1}^{*(n)}) && (4.73)\\
\leq & \sum_{x^n \in \{x^n : (x^n, y^n) \in A_{\epsilon,1}^{*(n)}\}} 2^{-n(H_1(X)-\epsilon_1)} && (4.74)\\
\leq & 2^{n(H_1(X|Y)+\epsilon_2)} 2^{-n(H_1(X)-\epsilon_1)} && (4.75)\\
= & 2^{-n(I_1(X;Y)-\epsilon')} && (4.76)
\end{aligned}$$

where $H_1(X), H_1(X|Y), I_1(X;Y)$ are all information measures calculated according to joint distribution $p_1(x,y)$.

Notice that $A_{\epsilon,1}^{*(n)}(Y) = A_\epsilon^{*(n)}(Y)$ due to the same marginal distribution $p(y)$ of $p(x,y)$ and $p_1(x,y)$, so for any $y^n \in A_\epsilon^{*(n)}(Y)$,

$$\begin{aligned}
& Pr((X^n, y^n) \in A_{\epsilon,1}^{*(n)}) && (4.77)\\
= & Pr((X^n, y^n) \in A_{\epsilon,1}^{*(n)}, X^n \in A_{\epsilon,1}^{*(n)}) && (4.78)\\
= & Pr((X^n, y^n) \in A_{\epsilon,1}^{*(n)} | X^n \in A_{\epsilon,1}^{*(n)}) \cdot Pr(X^n \in A_{\epsilon,1}^{*(n)}) && (4.79)\\
\leq & Pr((X^n, y^n) \in A_{\epsilon,1}^{*(n)} | X^n \in A_{\epsilon,1}^{*(n)}) && (4.80)\\
\leq & 2^{-n(I_1(X;Y)-\epsilon')}. && (4.81)
\end{aligned}$$

□

Categorization of Sequences in $\mathcal{X}^n$

Notation 4.2.3. *Let J_ϵ^n denote the number of types whose type class is not contained in the ϵ-typical set with respect to $p(x)$, i.e.,*

$$J_\epsilon^n := \left| \left\{ P \in \mathcal{P}_n : T(P) \nsubseteq A_\epsilon^{*(n)} \right\} \right|, \tag{4.82}$$

where $J_\epsilon^n \leq |\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$ due to Theorem 4.2.5. Regard these types as probability mass functions and index them as $p_j(x)$, where j ranges from 1 to J_ϵ^n.

Now, we categorize the sequences in the whole input signal space $\mathcal{X}^n$ as follows:

Firstly, $\mathcal{X}^n$ can be divided into two classes, the strongly typical set with respect to $p(x)$ and the atypical set, denoted by $A_\epsilon^{*(n)}$ and $A_\epsilon^{*(n)^c}$ respectively.

Secondly, in the atypical set $A_\epsilon^{*(n)^c}$, the sequences can actually be regarded as typical with respect to the other distributions. Using notation 4.2.3, we can express the atypical set as

$$A_\epsilon^{*(n)^c} \approx \bigcup_{j=1}^{J_\epsilon^n} A_{\epsilon,j}^{*(n)}, \tag{4.83}$$

where $A_{\epsilon,j}^{*(n)}$ is the strongly typical set with respect to the type $p_j(x)$, for any $j \in \{1, \cdots, J_\epsilon^n\}$. Note that, due to the effect introduced by ϵ, these ϵ-typical sets may not be necessarily disjoint with each other but possibly overlap. However, the union of them is always approximately equal to $A_\epsilon^{*(n)^c}$. Now, clearly, we have the following partition for $\mathcal{X}^n$:

$$\mathcal{X}^n = A_\epsilon^{*(n)} \bigcup \cup_{j=1}^{J_\epsilon^n} A_{\epsilon,j}^{*(n)}, \tag{4.84}$$

Finally, given the channel $p(y|x)$, one can calculate the joint distribution $p(x,y)$ and $p_j(x,y)$ associated with $p(x)$ and $p_j(x)$, and then the marginal distribution $p(y)$ and $p_j(y)$ can also be obtained, where $j \in \{1, \cdots, J_\epsilon^n\}$. For simplicity and consistency with $p_j(\cdot)$, where $j \in \{1, \cdots, J_\epsilon^n\}$, we also use $p_0(\cdot)$ to denote $p(\cdot)$, i.e., we will not distinguish the use of $p_0(\cdot)$ and $p(\cdot)$ in the later discussion and calculation. To use $p_0(\cdot)$ or $p(\cdot)$ is based on the context to make the discussion or calculation easy and clear. Fix an $y^n \in A_\epsilon^{*(n)}(Y)$, then the sequences in $A_\epsilon^{*(n)}(X)$ and $A_{\epsilon,j}^{*(n)}(X)$, can be further divided into two classes: those ϵ-jointly typical with y^n and those not jointly typical with y^n, with respect to $p(x,y)$ or $p_j(x,y)$, where $j \in \{1, \cdots, J_\epsilon^n\}$.

The above categorization is illustrated in Figure 4.1.

Redefining the Typical Codebook

To formally redefine the typical codebook, we use the following notation:

$$F_\epsilon(i) := \left\{x^n \in \mathcal{X}^n : y_\epsilon^n(i) \in A_\epsilon^{*(n)}(Y|x^n)\right\}, \forall\ i \in \{1, \cdots, M_\epsilon^{(n)}\}, \tag{4.85}$$

where $A_\epsilon^{*(n)}(Y|x^n)$ is with respect to the channel transition probability $p(y|x)$. The interpretation is that $y_\epsilon^n(i)$ lies in the conditionally typical set of each sequence in $F_\epsilon(i)$. In other words, each sequence x^n in $F_\epsilon(i)$ can reach $y_\epsilon^n(i)$ over a channel with typical noise, i.e., $x^n \in F_\epsilon(i)$ is at a typical distance with $y_\epsilon^n(i)$.

Based on the above analysis of $\mathcal{X}^n$, we can also classify the sequences in $F_\epsilon(i)$. The first class is denoted by

$$F_{\epsilon,0}(i) := \left\{x^n \in A_\epsilon^{*(n)} : y_\epsilon^n(i) \in A_\epsilon^{*(n)}(Y|x^n), \right\}, \forall\ i \in \{1, \cdots, M_\epsilon^{(n)}\}. \tag{4.86}$$

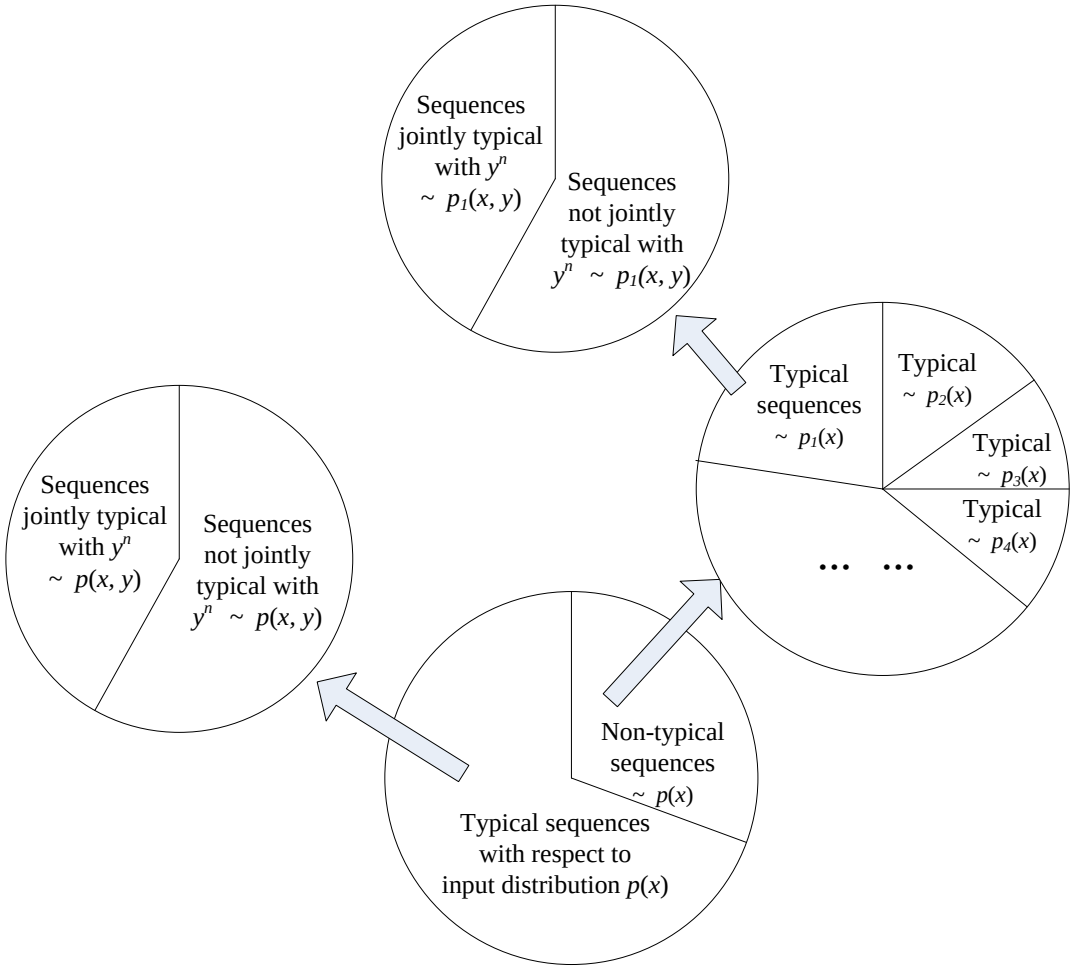

Figure 4.1: The categorization of sequences in $\mathcal{X}^n$.

Besides, $F_\epsilon(i)$ also includes the sequences not typical $\sim p(x)$ but typical with respect to other types. Label these types as $p_{i,k}(x)$, where k ranges from 1 to $K_\epsilon^n(i)$ and obviously

$$K_\epsilon^n(i) \leq J_\epsilon^n \leq (n+1)^{|\mathcal{X}|}. \tag{4.87}$$

Letting

$$F_{\epsilon,k}(i) := \left\{ x^n \in A_{\epsilon,k}^{*(n)} : y_\epsilon^n(i) \in A_\epsilon^{*(n)}(Y|x^n) \right\}, \tag{4.88}$$

where $k \in \{1, \cdots, K_\epsilon^{(n)}(i)\}, i \in \{1, \cdots, M_\epsilon^{(n)}\}$, we can express $F_\epsilon(i)$ as

$$F_\epsilon(i) = \bigcup_{k=0}^{K_\epsilon^n(i)} F_{\epsilon,k}(i). \tag{4.89}$$

Furthermore, the following notation is useful for the concise definition of typical codebooks.

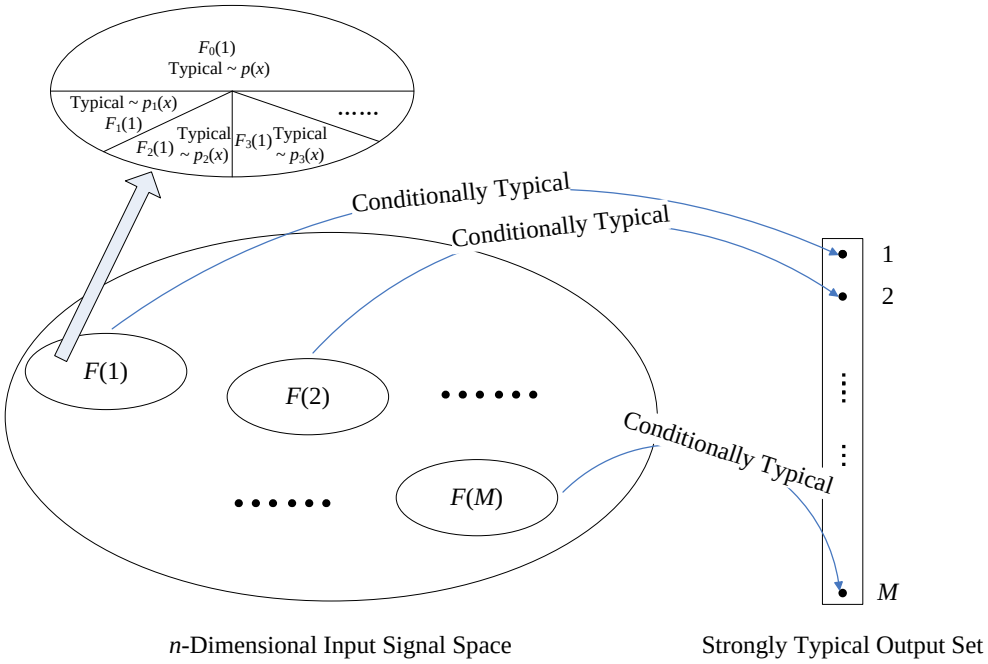

Figure 4.2: The way to redefine a typical codebook.

Notation 4.2.4.

$$\mathcal{F}_\epsilon := \{F_{\epsilon,k}(i), i \in \{1, \cdots, M_\epsilon^{(n)}\}, k \in \{0, \cdots, K_\epsilon^n(i)\}\}; \tag{4.90}$$

$$P_\epsilon(i,k) := Pr(X^n \in F_{\epsilon,k}(i)); \tag{4.91}$$

$$N_\epsilon(i,k|\mathcal{C}^{(n,R)}) := \sum_{w=1}^{2^{nR}} I(X^n(w) \in F_{\epsilon,k}(i), X^n(w) \in \mathcal{C}^{(n,R)}). \tag{4.92}$$

The relation among $A_\epsilon^{*(n)}(Y)$, $\mathcal{X}^n$ and F_ϵ and the partition of F_ϵ is depicted in Figure 4.2, where all the ϵ's and n's are omitted for simplicity. Each sequence $y_\epsilon^n(i) \in A_\epsilon^{*(n)}(Y)$ is a ϵ-strongly typical output sequence and corresponding to a set $F_\epsilon(i)$ in $\mathcal{X}^n$. Since $A_\epsilon^{*(n)}(Y)$ consists of $M_\epsilon^{(n)}$ elements, there are correspondingly $M_\epsilon^{(n)}$ sets. Further do the partition within those sets and combine all the sets together to form the $\mathcal{F}_\epsilon$. Moreover, for any $i \in \{1, \cdots, M_\epsilon^{(n)}\}$,

$$2^{-n(I(X;Y)+\epsilon_0)} \leq P_\epsilon(i,0) \leq 2^{-n(I(X;Y)-\epsilon_0)}, \tag{4.93}$$

and

$$P_\epsilon(i,k) \leq 2^{-n(I_k(X;Y)-\epsilon_k)}, \tag{4.94}$$

where ϵ_0 and ϵ_k go to 0 as $\epsilon \to 0$ and $n \to \infty$. Also note that each $F_{\epsilon,k}(i)$ is actually a subset of $\mathcal{X}^n$ and therefore $(\mathcal{X}^n, \mathcal{F}_\epsilon)$ forms a range space.

Based on all the above discussions, we redefine the typical codebooks for discrete memoryless channel as following:

Definition 4.2.4.

$$T(\mathcal{C}^{(n,R)}) = \left\{ \mathcal{C}^{(n,R)} : \sup_{F_{\epsilon,k}(i)\in\mathcal{F}_\epsilon} \left| \frac{N_\epsilon(i,k|\mathcal{C}^{(n,R)})}{2^{nR}} - P_\epsilon(i,k) \right| \leq \frac{\Delta_\epsilon nR}{2^{nR}} \right\} \tag{4.95}$$

where $\Delta_\epsilon := \max\{8\,VC\text{-}d(\mathcal{F}_\epsilon), 16e\}$.

Theorem 4.2.6. *For a discrete memoryless channel, generate the codebook* $\mathcal{C}^{(n,R)}$ *at random according to the distribution* $p(x)$*, then*

$$Pr(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \to 1 \text{ as } n \to \infty. \tag{4.96}$$

Proof. The cardinality of $\mathcal{F}_\epsilon$ can be bounded as follows:

$$|\mathcal{F}_\epsilon| \leq ((n+1)^{|\mathcal{X}|}+1)\cdot 2^{n(H(Y)+\epsilon_1)} := 2^{n(H(Y)+\epsilon^{'})} \tag{4.97}$$

where $\epsilon^{'}$ goes to 0 as $\epsilon \to 0$ and $n \to \infty$, and therefore, for a fixed dimension n, VC-d$(\mathcal{F}_\epsilon) \leq n(H(Y)+\epsilon^{'})$. Due to the finite VC Dimension, along the same line as before, it follows the proof of Theorem 4.2.6. □

We are now in a position to prove Theorem 4.2.4.

Proof of Theorem 4.2.4. Consider $Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}))$ for any $i \in \{1,\cdots,M_\epsilon\}$. Along the same line as in the proof of theorem 4.2.3, we can readily get a tight lower bound,

$$Pr(Y^n = y_\epsilon^n(i)) \dot{\geq} 2^{-nH(Y)}, \forall i \in \{1,\ldots,M_\epsilon^{(n)}\}, \tag{4.98}$$

when $R > C$ and n is sufficiently large.

Below, we will try to give a tight upper bound.

$$Pr(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{4.99}$$

$$\overset{(a)}{=} \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} P(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), X^n = x^n(w)) \tag{4.100}$$

$$\overset{(b)}{=} \frac{1}{2^{nR}} \sum_{x^n(w)\in F_\epsilon(i)} P(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), X^n = x^n(w)) \tag{4.101}$$

$$\overset{(c)}{\leq} \frac{1}{2^{nR}} \sum_{k=0}^{K_\epsilon^n(i)} \sum_{x^n(w)\in F_{\epsilon,k}(i)} P(Y^n = y_\epsilon^n(i)|C_\epsilon, \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), X^n = x^n(w)) \tag{4.102}$$

$$\overset{(d)}{\leq} \frac{1}{2^{nR}} \sum_{k=0}^{K_\epsilon^n(i)} N_\epsilon(i,k|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \cdot \frac{2^{-n[H_k(Y|X)-\epsilon_k^{'}]}}{(1-o(1))} \tag{4.103}$$

$$\overset{(e)}{\leq} \frac{(1+o(1))}{2^{nR}} \sum_{k=0}^{K_\epsilon^n(i)} N_\epsilon(i,k|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \cdot 2^{-n[H_k(Y|X)-\epsilon_k^{'}]} \tag{4.104}$$

$$\overset{(f)}{\leq}\frac{(1+o(1))}{2^{nR}}\sum_{k=0}^{K_\epsilon^n(i)}(2^{nR}\cdot 2^{-n[I_k(X;Y)-\epsilon_k]}+\Delta_\epsilon nR)\cdot 2^{-n[H_k(Y|X)-\epsilon_k^{'}]} \tag{4.105}$$

$$\overset{(g)}{=}(1+o(1))\sum_{k=0}^{K_\epsilon^n(i)}\left\{2^{-n[H(Y)-\epsilon_k-\epsilon_k^{'}]}+\frac{\Delta_\epsilon nR}{2^{nR}}\cdot 2^{-n[H_k(Y|X)-\epsilon_k^{'}]}\right\} \tag{4.106}$$

$$\overset{(h)}{=}(1+o(1))\sum_{k=0}^{K_\epsilon^n(i)}2^{-n[H(Y)-\epsilon_k-\epsilon_k^{'}]}\left\{1+\frac{\Delta_\epsilon nR}{2^{nR}}\cdot 2^{n[I_k(Y|X)-\epsilon_k]}\right\} \tag{4.107}$$

where

"(a)" follows along the similar line with that in the proof of theorem 4.2.3.

"(b)" follows from the constraint C_ϵ, and excludes the contributions from those invalid codewords, i.e., the codewords outside the set $F_\epsilon(i)$.

"(c)" follows from equation (4.89) and accumulate the contributions from all the codewords in $F_\epsilon(i)$.

"(d)" follows from upper bounding $p(y_\epsilon^n(i)|C_\epsilon,\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}),X^n=x^n(w))$, where $x^n(w)\in F_{\epsilon,k}(i)$. Specifically, given that $x^n\in F_{\epsilon,k}(i)$, we know x^n is strongly typical $\sim p_k(x)$. Moreover, according to the condition C_ϵ, $y_\epsilon^n(i)$ is conditionally strongly typical with x^n and hence $(x^n,y_\epsilon^n(i))$ are jointly strongly typical $\sim p_k(x,y)$. Therefore, it follows that $(x^n,y_\epsilon^n(i))$ are jointly weakly typical where $x^n\in F_{\epsilon,k}(i)$, and hence we can easily get the bound for $p(y_\epsilon^n(i)|C_\epsilon,\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}),X^n=x^n(w))$, where $\epsilon_k^{'}\to 0$, as $\epsilon\to 0$ and $n\to\infty$.

"(f)" follows from the definition of typical codebook and gives a upper bound to the number of valid codewords, where $\epsilon_k\to 0$, as $\epsilon\to 0$ and $n\to\infty$.

"(g)" and "(h)" follow from the calculations similar with those in the proof of theorem 4.2.3.

Now, let us investigate equation (4.107). If $R>C=\max_{p(x)}I(X;Y)$, then $\forall k\in\{0,\cdots,K_\epsilon^n(i)\},i\in\{1,\ldots,M_\epsilon^{(n)}\}$, we have

$$\frac{\Delta_\epsilon nR}{2^{nR}}\cdot 2^{n[I_k(X;Y)-\epsilon_k]}\to 0,\ \text{as } n\to\infty. \tag{4.108}$$

Recall that $K_\epsilon^n(i)$ is at most polynomial with n. Therefore, when $R>C$,

$$Pr(Y^n=y_\epsilon^n(i)|C_\epsilon,\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))\leq 2^{-n[H(Y)-\epsilon^{'}]} \tag{4.109}$$

where $\epsilon^{'}\to 0$ as $n\to\infty$, for any $i\in\{1,\ldots,M_\epsilon^{(n)}\}$.

Therefore, when $R>C$ and n is sufficiently large,

$$Pr(Y^n=y_\epsilon^n(i)|C_\epsilon,\mathcal{C}^{(n,R)}\in T(\mathcal{C}^{(n,R)}))\dot{\leq}2^{-nH(Y)}, \tag{4.110}$$

for any $i\in\{1,\ldots,M_\epsilon^{(n)}\}$ and this completes the proof of Theorem 4.2.4. □

4.3 Rate Needed for the Relay to forward its Observation

To strengthen our judgement that relay cannot do better compression even if it tries to utilizes the source's codebook, we study the rate needed for the relay to forward its observation under two scenarios: source's codebook used and not used. For this purpose, we calculate the conditional entropy of the channel output given codebook $H(Y^n|\mathcal{C}^{(n,R)})$ when $R > C$. It is shown that the average rate $\frac{1}{n}H(Y^n|\mathcal{C}^{(n,R)})$ is asymptotically equal to $H(Y)$, which is the traditional rate needed for relay to simply ignore the codebook information and forward its observation.

Formally, we have the following theorem:

Theorem 4.3.1. *For the n-th extension of discrete memoryless channel, if the coding rate is greater than channel capacity, the average rate needed to forward(store) the channel output with codebook information employed is asymptotically equal to the rate with codebook information unemployed, i.e., when $R > I(X;Y)$,*

$$\lim_{n\to\infty}\frac{1}{n}H(Y^n|\mathcal{C}^{(n,R)}) = H(Y). \tag{4.111}$$

Proof. We use definition 4.2.1 and the results shown in section 4.2. First we define a indicator random variables, i.e., let

$$I := I(E_\epsilon) = I(X^n \in A_\epsilon^{*(n)}, C_\epsilon). \tag{4.112}$$

Then, we consider the conditional entropy of the channel output given codebook $H(Y^n|\mathcal{C}^{(n,R)})$ when $R > I(X;Y)$. To distinguish the difference between a random codebook $\mathcal{C}^{(n,R)}$ and one of its realizations, we use $\mathcal{C}$ to represent the latter.

$$H(Y^n|\mathcal{C}^{(n,R)}) \tag{4.113}$$
$$\overset{(a)}{\geq} H(Y^n|I,\mathcal{C}^{(n,R)}) \tag{4.114}$$
$$= Pr(I=1)\cdot H(Y^n|I=1,\mathcal{C}^{(n,R)}) + Pr(I=0)H(Y^n|I=0,\mathcal{C}^{(n,R)}) \tag{4.115}$$
$$\geq Pr(I=1)\cdot H(Y^n|I=1,\mathcal{C}^{(n,R)}) \tag{4.116}$$
$$\overset{(b)}{=} (1-o(1))\cdot H(Y^n|I=1,\mathcal{C}^{(n,R)}) \tag{4.117}$$
$$= (1-o(1))\cdot\sum_{\mathcal{C}} p(\mathcal{C})\cdot H(Y^n|I=1,\mathcal{C}^{(n,R)}=\mathcal{C}) \tag{4.118}$$
$$\geq (1-o(1))\cdot\sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C})\cdot H(Y^n|I=1,\mathcal{C}^{(n,R)}=\mathcal{C}) \tag{4.119}$$
$$\overset{(c)}{=} (1-o(1))\cdot\sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C})\cdot\left(\sum_{y^n} p(y^n|E_\epsilon,\mathcal{C}^{(n,R)}=\mathcal{C})\log\frac{1}{p(y^n|E_\epsilon,\mathcal{C}^{(n,R)}=\mathcal{C})}\right) \tag{4.120}$$

$$\geq \quad (1-o(1))$$
$$\cdot \sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C}) \cdot \left(\sum_{y^n \in A_\epsilon^{*(n)}(Y)} p(y^n|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C}) \log \frac{1}{p(y^n|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C})} \right) \tag{4.121}$$
$$= \quad (1-o(1))$$
$$\cdot \sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C}) \cdot \left(\sum_{i\in\{1,\ldots,M_\epsilon^{(n)}\}} p(y_\epsilon^n(i)|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C}) \log \frac{1}{p(y_\epsilon^n(i)|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C})} \right) \tag{4.122}$$
$$\overset{(d)}{\geq} (1-o(1)) \cdot \sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C}) \cdot \left(\sum_{i\in\{1,\ldots,M_\epsilon^{(n)}\}} p(y_\epsilon^n(i)|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C}) \log 2^{n[H(Y)-\epsilon^*]} \right) \tag{4.123}$$
$$= n[H(Y)-\epsilon^*] \cdot (1-o(1)) \cdot \sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C}) \cdot \left(\sum_{i\in\{1,\ldots,M_\epsilon^{(n)}\}} p(y_\epsilon^n(i)|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C}) \right) \tag{4.124}$$
$$= n[H(Y)-\epsilon^*] \cdot (1-o(1)) \cdot \sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C}) \cdot Pr(Y^n \in A_\epsilon^{*(n)}(Y)|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C}) \tag{4.125}$$
$$\overset{(e)}{\geq} n[H(Y)-\epsilon^*] \cdot (1-o(1)) \cdot \sum_{\mathcal{C}\in T(\mathcal{C}^{(n,R)})} p(\mathcal{C}) \tag{4.126}$$
$$\overset{(f)}{=} n[H(Y)-\epsilon^*] \cdot (1-o(1)) \cdot (1-o(1)) \tag{4.127}$$
$$= n[H(Y)-\epsilon^*] \cdot (1-o(1)) \tag{4.128}$$

where

"(a)" follows from the fact that conditioning reduces entropy;

"(b)" follows since $Pr(E_\epsilon) \to 1$ as $n \to \infty$, for any $\epsilon > 0$;

"(c)" follows from the definition of conditional entropy;

"(d)" involves upper bounding $p(y_\epsilon^n(i)|E_\epsilon, \mathcal{C}^{(n,R)} = \mathcal{C})$ by $2^{-n[H(Y)-\epsilon^*]}$, where $\epsilon^* \to 0$ as $n \to \infty$;

"(e)" follows the fact that

$$Pr(Y^n \in A_\epsilon^{*(n)}(Y)|E_\epsilon, \mathcal{C}^{(n,R)}=\mathcal{C}) = 1-o(1). \tag{4.129}$$

"(f)" follows from $Pr(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \to 1$ as $n \to \infty$.

Therefore, when $R > I(X;Y)$,

$$\lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathcal{C}^{(n,R)}) \tag{4.130}$$

$$= \lim_{n\to\infty} \frac{1}{n} \left(n[H(Y) - \epsilon^*] \cdot (1 - o(1))\right) \tag{4.131}$$

$$= \lim_{n\to\infty} [H(Y) - \epsilon^*] \cdot (1 - o(1)) \tag{4.132}$$

$$= H(Y), \tag{4.133}$$

which finishes the proof. □

4.4 A Proof of the Strong Converse to Channel Coding Theorem under Random Coding Framework

From the above discussions, we have seen that the proposal of typical codebooks and our analysis method provide a new perspective to consider the fundamental problems, especially those needing to be treated under a specific codebook. This undoubtedly adds more insight into the random coding schema, compared to the classical analysis method based on averaging over all codes and ignoring the structure of a specific code.

As a by-product of studying the relay problem and CDOCI, in this section, we show that under the now standard random codebook construction shema, we cannot achieve rate larger than capacity in the sense that with high probability the error probability associated with a random code is 1 for rate greater than capacity. Although we cannot claim this eliminates all possibilities to achieve $R > C$ (not only under the random coding framework) and hence may not be a strict converse to channel coding theorem, this result, however, again demonstrates the power of our analysis method to consider the converse part of the information theoretic problems and may find other future applications.

Let $P_e^{(n)}(\mathcal{C}^{(n,R)})$ be the average error probability associated with a random codebook $\mathcal{C}^{(n,R)}$. Then it is obvious that $P_e^{(n)}(\mathcal{C}^{(n,R)})$ is a random variable depending on the random codebook.

Theorem 4.4.1. *The rate greater than channel capacity is not achievable under the standard random codebook construction shema in the sense that*

$$Pr(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1) \to 1, \ \textit{as } n \to \infty, \tag{4.134}$$

when $R > C$.

To prove this theorem, let us introduce some notations and a lemma first.

Let

$$E_i := \{|A_{\epsilon_1}(y_\epsilon^n(i))| > 1\}, \ \forall i \in \{1, \ldots, M_\epsilon^{(n)}\}. \tag{4.135}$$

Obviously, if E_i happens, the size of ϵ_1-jointly typical set formed by decoder for $y_\epsilon^n(i)$ is greater than 1 and hence an error should be declared. The event $\cap_{i=1}^{M_\epsilon^{(n)}} E_i$ indicates that if an strongly typical sequence is received, then there is an error.

Lemma 4.4.1. *When the coding rate is larger than channel capacity and the block length is sufficiently large,*

$$Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} E_i \middle| \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) = 1. \tag{4.136}$$

Proof. Recall the interpretation of $N_\epsilon(i|\mathcal{C}^{(n,R)})$. Briefly, $N_\epsilon(i|\mathcal{C}^{(n,R)})$ is the number of codeword in $\mathcal{C}^{(n,R)}$, which are jointly typical with $y_\epsilon^n(i)$. Property choose ϵ_1 such that

$$|A_{\epsilon_1}(y_\epsilon^n(i))| \geq N_\epsilon(i|\mathcal{C}^{(n,R)}),\ \forall i \in \{1,\ldots,M_\epsilon^{(n)}\}. \tag{4.137}$$

where both ϵ and ϵ_1 go to 0 as $n \to \infty$.

According to the definition of typical codebook, we have when $R > C$,

$$N_\epsilon(i|\mathcal{C}^{(n,R)}) \geq 2^{nR}2^{-n[I(X;Y)+\epsilon^{'}]} - \Delta_\epsilon nR \to \infty,\ \text{as } n \to \infty. \tag{4.138}$$

Therefore, for rate larger than capacity and sufficiently large n,

$$|A_{\epsilon_1}(y_\epsilon^n(i))| > 1\ \forall i \in \{1,\ldots,M_\epsilon^{(n)}\}, \tag{4.139}$$

and hence

$$Pr\left(\bigcap_{i=1}^{M_\epsilon^{(n)}} E_i \middle| \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})\right) = 1. \tag{4.140}$$

□

Proof of 4.4.1. According to the definition of $P_e^{(n)}(\mathcal{C}^{(n,R)})$,

$$\begin{aligned}
&P_e^{(n)}(\mathcal{C}^{(n,R)}) && (4.141)\\
=&Pr(\text{an error is declared}|\mathcal{C}^{(n,R)}) && (4.142)\\
\geq&Pr(|A_{\epsilon_1}(Y^n)| > 1|\mathcal{C}^{(n,R)}), && (4.143)
\end{aligned}$$

and thus we have the following relationship between two events,

$$\{Pr(|A_{\epsilon_1}(Y^n)| > 1|\mathcal{C}^{(n,R)}) = 1\} \subseteq \{P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1\}. \tag{4.144}$$

When $R > C$, we have

$$\begin{aligned}
&Pr(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1) && (4.145)\\
\geq&P(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \cdot P(\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) && (4.146)\\
=&(1-o(1))P(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) && (4.147)\\
\geq&(1-o(1))P(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1, Y^n \in A_\epsilon^{*(n)}(Y))|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) && (4.148)\\
=&(1-o(1))P(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), Y^n \in A_\epsilon^{*(n)}(Y))
\end{aligned}$$

$$\cdot \underbrace{P(Y^n \in A_\epsilon^{*(n)}(Y))|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})}_{1-o(1)} \tag{4.149}$$

$$\overset{(a)}{=}(1-o(1))Pr(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), Y^n \in A_\epsilon^{*(n)}(Y)) \tag{4.150}$$

$$\geq(1-o(1))Pr(Pr(|A_{\epsilon_1}(Y^n)| > 1|\mathcal{C}^{(n,R)}) = 1|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), Y^n \in A_\epsilon^{*(n)}(Y)) \tag{4.151}$$

$$\overset{(b)}{=}1-o(1) \tag{4.152}$$

where

"(a)" holds since

$$P(Y^n \in A_\epsilon^{*(n)}(Y)|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{4.153}$$

$$\geq P(Y^n \in A_\epsilon^{*(n)}(Y), X^n \in A_{\epsilon_1}^{*(n)}(X)|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \tag{4.154}$$

$$=P(Y^n \in A_\epsilon^{*(n)}(Y)|X^n \in A_{\epsilon_1}^{*(n)}(X), \mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)})) \cdot \underbrace{P(X^n \in A_{\epsilon_1}^{*(n)}(X)|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}))}_{1-o(1)} \tag{4.155}$$

$$\to 1, \text{ as } n \to \infty. \tag{4.156}$$

"(b)" follows from that

$$Pr(Pr(|A_{\epsilon_1}(Y^n)| > 1|\mathcal{C}^{(n,R)}) = 1|\mathcal{C}^{(n,R)} \in T(\mathcal{C}^{(n,R)}), Y^n \in A_\epsilon^{*(n)}(Y)) = 1 \tag{4.157}$$

due to Lemma 4.4.1.

Therefore, when $R > C$,

$$Pr(P_e^{(n)}(\mathcal{C}^{(n,R)}) = 1) \to 1, \text{ as } n \to \infty, \tag{4.158}$$

showing that the rate greater than channel capacity is not achievable under the standard random codebook construction shema. This completes the proof of Theorem 4.4.1. □

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Motivated by the problem of cooperation in networking shown in Figure 1.3, say a relay channel with orthogonal link between relay and destination, we try to characterize the posterior conditional probability of channel output given the codebook information when coding rate is larger than channel capacity, for both the BSC and DMC channel.

It is shown that under both scenarios, with high probability, the conditional distribution is asymptotically uniform and approaching to the unconditional distribution, when $R > C$ and the block length n is sufficiently large. This implicates that for this relay problem, if source's coding rate is greater than what relay can handle, the compress-and-forward strategy is kind of optimal under the random coding framework.

Moreover, the above judgement on the optimality of compress-and-forward strategy is further confirmed by showing that the rates needed for the relay to forward its observation are asymptotically the same no matter whether the relay utilize the codebook information or not.

Finally, as a by-product of studying the relay problem, we give a strong converse to the channel coding theorem under the now standard random codebook construction shema, using the techniques developed in this thesis.

5.2 Future Work

Today, with the pervasion of Internet and the massive demand of wireless broadband access, network information theory is experiencing a renaissance. Several significant progresses have been made in this field over the past decade. For example, the *throughput and transport capacity* is introduced to analyze the capacity of wireless networks [18] [19], and a new notion of *network coding* is proposed and the related

theory is well developed [23], and so on. However, disappointingly, compared to the thousands of nodes in the Internet, the size of networks that are completely understood in network information theory is merely no more than 3. Precisely, even in the three-node scenario shown in Figure 1.3, much work remains to be done.

In the short run, we hope to extend our results on BSC and DMC to the Additive White Gaussian Noise (AWGN) channel, i.e., to show that if the input distribution of the AWGN channel is gaussian, then for those typical output sequences, with high probability, the CDOCI is asymptotically uniform and approaching to the unconditional distribution, for rate above capacity and sufficiently large block length. Here, the conditional distribution refers to the conditional *probability density function*, rather than probability mass function in the DMC scenario.

In the long run, either to improve the current upper bound (converse part) or lower bound (forward part) to the capacity of relay channels is still a challenge. For the converse part, even if the compress-and-forward strategy can be shown to be optimal under random coding framework, whether the standard random coding is optimal for the relay channel is still not clear. For the forward part, to use the current techniques and obtain the better performance seems to be a tough task and a new coding technique may be called for.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pt. I, pp. 379–423, 1948; pt. II, pp. 623–656, 1948. 1, 10, 23

[2] T. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 572–584, Sep. 1979. 2, 15

[3] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–11, Jan. 1976. 3, 12

[4] F. Xue, P. R. Kumar and L.-L. Xie, "The conditional entropy of the jointly typical set when coding rate is larger than Shannon Capacity," *Unpublished Manuscript*. 3

[5] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991. 6, 10, 14, 20, 33, 34

[6] D. Slepian and J. K.Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. IT-19, pp. 471–480, Jul. 1973. 11

[7] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Info. Theory*, vol. 21, pp. 226–228, Mar. 1975. 12

[8] R. Ahlswede, "Multi-way communication channels," in *Proc. 2nd Int. Symp. Information Theory*, pp. 23–52, Budapest, Hungary: Hungarian Acad. Sci., 1971. 13

[9] H. Liao, "Multiple access channels," Ph.D. dissertation, Dept. Elec. Eng., Univ. of Hawaii, Honolulu, 1972. 13

[10] T. M. Cover, "Broadcast channels," *IEEE Trans. Info. Theory*, vol. IT-18, pp. 2–14, Jan. 1972. 13, 14

[11] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Info. Theory*, vol. IT-19, pp. 197–207, Mar. 1973. 13, 14

[12] R. G. Gallager, "Capacity and coding for degraded broadcast channels," *Problemy Peredacy Informaccii*, vol. 10, no. 3, pp. 3–14, Jul.–Sep. 1974. 13, 14

[13] P. P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Info. Theory*, vol. IT-20, pp. 279–280, Mar. 1974. 13, 14

[14] E. C. van der Meulen, "Three-terminal communication channels," *Adv. Appl. Prob.*, vol. 3, pp. 120–154, 1971. 14

[15] E. C. van der Meulen, "Transmission of information in a t-terminal discrete memoryless channel," Ph.D. dissertation, Dept. of Statistics, Univ. of California, Berkeley, 1968. 14

[16] V. N. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264280, Jan. 1971. 23, 24

[17] V. N. Vapnik, *Estimation of dependences based on empirical data*. New York: Springer-Verlag, 1982. 23, 24

[18] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 388–404, March. 2000. 24, 53

[19] L.-L. Xie and P. R. Kumar, "A Network Information Theory for Wireless Communication: Scaling Laws and Optimal Operation," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 748–767, Feb. 2004. 24, 53

[20] J. Ghaderi, L.-L. Xie, and X. Shen, "Hierarchical cooperation in Ad Hoc networks: optimal clustering and achievable throughput," submitted to *IEEE Trans. Inf. Theory*, February 2008. 24

[21] I. Csiszár and J. Körner, *Information Theory: Coding Theorem for Discrete Memoryless Systems*. Academic Press, New York, 1981. 7, 9

[22] R. G. Gallager, *Information Theory and Reliable Communication*. New York, Wiley, 1968. 10

[23] R. W. Yeung, N. Cai, S.-Y. R. Li, and Z. Zhang, "Theory of Network Coding," *Foundations and Trends in Communications and Information Theory*, 2006. 54