Feedback Capacity for a Discrete Non-Binary Noise Channel with Memory

by

Nevroz Şen

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Master of Engineering

Queen's University
Kingston, Ontario, Canada
May 2009

Abstract

In this project we study the channel capacity for communication systems when a feedback exists from the channel output to the encoder. More specifically, we study the feedback capacity of a discrete binary-input non-binary output channel with memory recently introduced in [15] to model soft-decision demodulated time-correlated fading channels. The channel, whose output process can be explicitly expressed in terms of its binary input process and a non-binary noise process, encompasses modulo-additive noise binary channels as a special case (realized when hard-decision demodulation is used on the underlying fading channel). We show that, even though the channel has memory, feedback does not increase its capacity when the noise process is stationary ergodic. We also note the validity of the result for arbitrary noise processes.

Acknowledgments

I am grateful to my supervisors Prof. Fady Alajaji and Prof. Serdar Yüksel for their trust and sincere interest in my education, insightful guidance, continuous support and for being so understanding to every difficulty that I had faced while completing my degree.

I am also grateful to my family. I cannot imagine myself going through with this degree without their unconditional love and support.

Table of Contents

Chapter 1

Introduction

In this chapter, we present some prior results on the feedback problem for channel capacity. We then specify the main contribution of this project. Following that, we give the outline of the project.

1.1 Literature Overview

The effects of feedback on the channel capacity where the channel encapsulates memory has taken a lot of attention especially in the last several years. Therefore, the literature on the feedback capacity is vast. In this project, we only state some of these results that are more closely related to our research. In earlier works, Shannon [18] showed that feedback does not increase the capacity of discrete memoryless channels. Cover and Pombra [9] and others considered additive channels with Gaussian noise and showed that feedback can increase the capacity at most half a bit and later it has been shown that [9] feedback at most doubles the capacity of a nonwhite Gaussian channel (the later result is originally due to Pinsker [16] and Ebert [10]). Alajaji

[1] showed that feedback does not increase the capacity of discrete modulo additive channels with arbitrary noise.

1.2 Contributions

Based on these available results, it is known that for some types of channels, e.g., symmetric channels with identical input and output alphabet sizes, feedback does not increase capacity [1, 4]. Inspired by this result, we investigate the feedback capacity of a discrete binary-input 2^q-ary output communication channel, which has recently been proposed in [15] to model soft-decision demodulated fading channels with memory. The channel, which we refer to by the non-binary noise discrete channel (NBNDC), is explicitly described in terms of a non-binary noise process that is independent of the channel input. We show that, in spite of the NBNDC's memory structure, feedback does not help to increase its capacity. It should be noted that the non-binary channel is still symmetric [12], however in contrast with the additive noise channel model the cardinality of the channel output is not the same as that of the input. Moreover, a uniform input does not yield a uniform output which brings a hope that by using feedback capacity can be increased. This is mainly an indication that there is still some room for the capacity to be increased as capacity without feedback is smaller than $\log_2 2^q = q$. However, as we show later, even though the output distribution is not uniform, it is still not possible to get a higher capacity via feedback. Although the result is proved under the assumption of stationary ergodic non-binary noise, we remark that it also holds for arbitrary (not necessarily stationary ergodic) noise.

This result generalizes in some sense the work in [1], where it is also shown that feedback does not increase capacity for discrete modulo-k additive channels with arbitrary noise with memory. Furthermore, when $q = 1$, the result intersects exactly with the result for $k = 2$ in [1], since the NBNDC reduces to the modulo-2 additive noise channel.

1.3 Organization of Thesis

We proceed by introducing some basic notations and definitions for memoryless channels. We continue with discussing information theoretic concepts when the channel encompasses memory. We follow this by introducing the channel model proposed in [15]. We make a deeper investigation on the non-feedback capacity for this channel. Next, we discuss the feedback capacity and state our main results. In the last chapter, we present a summary of the project.

Chapter 2

Memoryless Channels

In this section we give some basic definitions and theorems mainly on capacity for channels without memory.

2.1 Definitions

In the most general sense, a communication system consists of three parts: (1) The source, which generates messages at the transmitting end of the system, (2) The destination, which tries to estimate the message within a certain accuracy, and (3) The channel which consists of a noisy (in general) transmission medium to transfer the signal from the source to the destination.

In parallel to this definition, let us define what a discrete channel is.

Definition 2.1.1. *A discrete communication channel, denoted by* $(\mathcal{X}, p(y|x), \mathcal{Y})$ *is a system consisting of two finite sets* $\mathcal{X}$ *and* $\mathcal{Y}$ *and a collection of probability mass functions,* $p(y|x)$*, one for each* $x \in \mathcal{X}$*, such that for every* x *and* y*,* $p(y|x) \geq 0$ *and for every* x*,* $\sum_y p(y|x) = 1$*, where* $\mathcal{X}$ *is the input alphabet and* $\mathcal{Y}$ *is the output alphabet.*

Definition 2.1.2. *The n^{th} extension of a discrete memoryless channel (DMC) is the channel which is denoted by $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$ where*

$$p(y_k|x_k, x^{k-1}) = p(y_k|x_k) \quad k = 1, \cdots, n. \tag{2.1}$$

We should note that, when there is no feedback in the channel, i.e., if the input symbols are independent of past output symbols, then

$$p(y|x) = \prod_{i=1}^{n} p(y_i|x_i). \tag{2.2}$$

Definition 2.1.3. *An (M, n) block code for the channel given by $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following;*

- *An index set $\{1, 2, \cdots, M\}$.*
- *An encoding function $X^n : \{1, 2, \cdots, M\} \to \mathcal{X}^n$.*

Definition 2.1.4. *One important definition in communication theory is the Conditional Probability of Error which is given as follows:*

$$\lambda_i = Pr(g(Y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} p(y^n|x^n(i)) I(g(y^n) \neq i) \tag{2.3}$$

where $I(\cdot)$ is the indicator function and the definition stands for the conditional probability of error given that the index i was sent.

Definition 2.1.5. *The Average Probability of Error P_e^n for an (M, n) code is given as follows;*

$$P_e^n = \frac{1}{M} \sum_{i=1}^{M} \lambda_i. \tag{2.4}$$

Definition 2.1.6. *The rate R of an (M, n) code is*

$$R = \frac{\log_2 M}{n} \quad \textit{bits per transmission.} \tag{2.5}$$

Definition 2.1.7. *A rate R is called achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the average probability of error, P_e^n, tends to 0 as $n \to \infty$.*

Definition 2.1.8. *The capacity of a channel is the supremum of all achievable rates.*

In other words, *capacity* characterizes the maximum amount of reliable information that the channel can transmit.

To further study the channel capacity, we need to define two related concepts. We first introduce the concept of entropy, which is a measure of the uncertainty of a random variable.

Definition 2.1.9. *The entropy $H(X)$ of a discrete random variable X is given by*

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2.6}$$

Similarly, the joint entropy and conditional entropy of two random variables are defined as follows:

Definition 2.1.10. *The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is given by*

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{2.7}$$

Definition 2.1.11. *The conditional entropy $H(Y|X)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is given as*

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X) \end{aligned}$$

where E denotes the expectation.

We now define mutual information, which is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Definition 2.1.12. *For a pair of discrete random variables X, Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$, the mutual information, $I(X; Y)$, is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:*

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= E_{p(x,y)}\log\frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

Definition 2.1.13. (Information Channel Capacity) *For a DMC, the information channel capacity is given by*

$$C = \max_{p(x)} I(X;Y) \tag{2.8}$$

where the maximization is taken over all possible source distributions $p(x)$ and $I(X; Y)$ is the mutual information between the input and the channel output.

Additionally, the operational meaning of channel capacity can be given as the highest rate in bits per channel use that the information can be transmitted with arbitrary low probability of error. However, as proven in Shannon's Second Coding theorem, the operational capacity and the information capacity are equal.

We can now state some of the properties of channel capacity:

(a) $C \geq 0$, since $I(X; Y) \geq 0$.

(b) $C \leq \log_2 |\mathcal{X}|$, since $C = \max_{p(x)} I(X;Y) \leq H(X) = \log_2 |\mathcal{X}|$.

(c) $C \leq \log_2 |\mathcal{Y}|$.

(d) $I(X;Y)$ is a continuous function of $p(x)$.

(e) $I(X;Y)$ is a concave function of $p(x)$.

Throughout the project, we frequently refer to some specific class of channels where they mainly carry some sense of symmetry. This symmetry is characterized by looking at the channel transition matrix and it is very helpful in calculating the channel capacity. Therefore, before discussing channels with memory, we first make definitions for these symmetric channels.

2.2 Symmetric Channels

Typically, a discrete channel which is defined above, is characterized by a matrix, called channel transition matrix, which is a $|\mathcal{X}| \times |\mathcal{Y}|$ matrix whose entries are composed of $p(y|x)$ values. In some situations, the structure of this matrix is quite helpful to compute the channel capacity. We now define some of these structures and following them we state how we can compute the capacity for these channels.

Definition 2.2.1. *A channel is said to be strongly symmetric if the rows of its transition matrix $Q = [p(y|x)]$ are permutations of each other and also the columns are permutations of each other.*

A channel is said to be weakly symmetric if the rows of its transition matrix $Q = [p(y|x)]$ are permutations of each other and all the column sums $\sum_x p(y|x)$ are equal for every y [8].

For example, the channel with transition matrix

$$Q = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}$$

is weakly symmetric but not strongly symmetric.

In computing the channel capacity these symmetry conditions are quite helpful and as such we can state a theorem on the channel capacity of weakly symmetric channels.

Theorem 2.2.1. (Capacity for Weakly-Symmetric Channels) *[8] For a weakly symmetric channel*

$$C = \log|\mathcal{Y}| - H(\textit{row of transition matrix}) \tag{2.9}$$

where $H(\cdot)$ denotes entropy and the capacity is achieved by a uniform input distribution.

Proof of Theorem 2.2.1.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (2.10)\\ &= H(Y) - \sum_x p(x)H(Y|X=x) && (2.11) \end{aligned}$$

where $H(Y|X=x) = \sum_y p(y|x)\log(p(y|x))$.

Since every row of Q is a permutation of every other row, then $H(Y|X=x)$ is independent of x. Therefore,

$$H(Y|X=x) = H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \quad \forall x \in \mathcal{X} \tag{2.12}$$

where $(q_1, q_2, \cdots, q_Y)$ is any row in Q. This implies that

$$\begin{aligned} I(X;Y) &= H(Y) - H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) && (2.13) \\ &\leq \log|\mathcal{Y}| - H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}). && (2.14) \end{aligned}$$

A uniform input distribution yields a uniform distribution for the output Y since

$$\begin{aligned} p(y) &= \sum_{x \in \ \mathcal{X}} p(y|x)p(x) && (2.15) \\ &= \frac{1}{|\mathcal{X}|} \sum_{x \in \ \mathcal{X}} p(y|x) && (2.16) \\ &= \frac{K}{|\mathcal{X}|} \quad \forall y \in \mathcal{Y}, && (2.17) \end{aligned}$$

where $K = \sum_x p(y|x)$ is a constant. Since $\sum_x p(y) = 1$, we obtain that $K = \frac{|\mathcal{X}|}{|\mathcal{Y}|}$ and thus $p(y) = \frac{1}{|\mathcal{Y}|}, \forall y \in \mathcal{Y}$. □

Although this notion of symmetry includes many channel types such as the binary symmetric channel (BSC), the binary erasure channel, the modulo addition channel etc., it is possible to define a more general class of symmetric class which is called the quasi-symmetric channel.

Definition 2.2.2. *A DMC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and channel transition matrix $Q = [p(y|x)]$ is quasi-symmetric if Q can be partitioned along its columns into weakly-symmetric* sub-arrays $Q_1, Q_2, \ldots, Q_n$, with each Q_i having size $|\mathcal{X}| \times |\mathcal{Y}_i|$ where $\mathcal{Y}_1 \cup \cdots \cup \mathcal{Y}_n = \mathcal{Y}$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \ \forall i \neq j$ [3].

We should note that, the class of quasi-symmetric channels includes the classes of strongly and weakly symmetric channels as well as the class of symmetric channels defined by Gallager [12].

Theorem 2.2.2. (Capacity for Quasi-Symmetric Channels)*[3] The capacity C for a quasi-symmetric channel is given by*

$$C = \sum_{i=1}^{n} a_i C_i \tag{2.18}$$

where

$$a_i \;\; = \;\; \sum_{y \in \; \mathcal{Y}_i} p(y|x) = \textit{sum of any row in} \quad Q_i \tag{2.19}$$

and

$$C_i = \log|\mathcal{Y}_i| - H\Bigg(\textit{any row in the matrix} \quad \frac{1}{a_i} Q_i\Bigg)$$

for $i = 1, \cdots, n$

Proof of Theorem 2.2.2. We first observe that for each $i = 1, \cdots, n$, a_i is independent of the input value x, since sub-array i is weakly symmetric (so any row in Q_i is a permutation of any other row); and hence a_i is the sum of any row in Q_i.

For each $i = 1, \cdots, n$, define

$$p_i(y|x) = \begin{cases} \frac{p(y|x)}{a_i} & y \in \mathcal{Y}_i \\ 0 & \text{otherwise} \end{cases}$$

It can be easily verified that $p_i(y|x)$ is a legitimate conditional distribution. Thus $[p_i(y|x)] = \frac{1}{a_i} Q_i$ is the transition matrix of the weakly-symmetric sub-channel i with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}_i$. Let $I_i(X;Y)$ denote its mutual information. Since each such sub-channel i is weakly-symmetric, we know that its capacity C_i is given by

$$C_i = \max_{p(x)} I_i(X;Y) = \log|\mathcal{Y}_i| - \mathrm{H}\Bigg(\text{any row in the matrix} \frac{1}{a_i} Q_i\Bigg) \tag{2.20}$$

where the maximum is achieved by a uniform input distribution.

Now, the mutual information between the input and the output of channel Q can be written as

$$\begin{aligned} I(X;Y) &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x)p(y|x)\log_2 \frac{p(y|x)}{\sum_{x'\in\mathcal{X}} p(y|x')p(x')} && (2.21) \\ &= \sum_{i=1}^{n}\sum_{y\in\mathcal{Y}_i}\sum_{x\in\mathcal{X}} a_i p(x)\frac{p(y|x)}{a_i}\log_2 \frac{\frac{p(y|x)}{a_i}}{\sum_{x'\in\mathcal{X}} \frac{p(y|x')}{a_i}p(x')} && (2.22) \\ &= \sum_{i=1}^{n} a_i \sum_{y\in\mathcal{Y}_i}\sum_{x\in\mathcal{X}} p(x)p_i(y|x)\log_2 \frac{p_i(y|x)}{\sum_{x'\in\mathcal{X}} p_i(y|x')p(x')} && (2.23) \\ &= \sum_{i=1}^{n} a_i I_i(X;Y). && (2.24) \end{aligned}$$

Therefore, the channel capacity of channel Q is

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} \sum_{i=1}^{n} a_i I_i(X;Y) \\ &= \sum_{i=1}^{n} a_i \max_{p(x)} I_i(X;Y) \quad \text{(since each } I_i(X;Y) \text{ is maximized by the same uniform p(x))} \\ &= \sum_{i=1}^{n} a_i C_i. && (2.25) \end{aligned}$$

□

The definitions and theorems that we stated so far are mainly for communication systems when there is no memory in the channel. In the next chapter, we extend the notions that we have been discussing to communication systems with memory.

Chapter 3

Channels With Memory

In the previous chapter we discussed main concepts on the channel capacity for memoryless channels. Channels with memory are however more interesting as their noise process exhibits statistical dependency. Furthermore, for feedback capacity problems, memory is crucial since Shannon already showed that feedback does not increase capacity of memoryless channels [18]. Therefore, in this chapter we present some further information theoretic aspects of channels with memory.

3.1 Information Sources

We begin by a classification of sources with memory. In the rest of this chapter, by a "random source" we mean a stochastic process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$. In general we say that a source has a memory if there exists a dependence between the random variables of the source. Consider a discrete source $\{X_i\}_{i=1}^{\infty}$ with finite alphabet $\mathcal{X}$ characterized by the joint n-dimensional probability mass functions (pmfs); $P[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] := p(x_1, x_2, \cdots, x_n)$ for all $x^n \in \mathcal{X}^n$ where $p(x^n)$ satisfies the compatibility condition;

$$\sum_{x_n \in \mathcal{X}} p(x_1, x_2, \cdots, x_n) = p(x_1, x_2, \cdots, x_{n-1})$$

and it should be noted that $p(x^n) = p(x_1)\prod_{i=2}^{n} p(x_i|x_{i-1}, \cdots, x_1)$. Therefore, if $p(x_1)$ and conditional distribution $p(x_i|x_{i-1}, \cdots, x_1)$ are given $p(x^n)$ can be recursively determined.

Definition 3.1.1. *A stochastic process is stationary if*

$$P[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] = P[X_{1+\tau} = x_1, X_{2+\tau} = x_2, \cdots, X_{n+\tau} = x_n]$$

$\forall n, \tau$ *and* $\forall x^n \in \mathcal{X}^n$.

The main idea in this definition is that joint distribution is invariant under time shifts. In practice, many sources are well modeled using stationary sources.

Another important point that should be noted is that stationarity implies identical distribution for a source and we can easily show that an (i.i.d) discrete memoryless source is stationary since;

$$\begin{aligned} P[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] &= \prod_{i=1}^{n} P[X_i = x_i] \text{ by independence} \\ &= \prod_{i=1}^{n} P[X_{i+\tau} = x_i] \text{ by identical distribution} \\ &= P[X_{1+\tau} = x_1, X_{2+\tau} = x_2, \cdots, X_{n+\tau} = x_n] \end{aligned}$$

One of the interesting sources that embeds memory is the **Markov** source.

Definition 3.1.2. *A discrete process* $\{X_i\}_{i=1}^{\infty}$ *with finite alphabet* $\mathcal{X}$ *is said to be a Markov chain (MC) (or Markov source) if for* $n = 1, 2, \cdots$

$$P[X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_1 = x_1] = P[X_n = x_n | X_{n-1} = x_{n-1}], \ \forall x^n \in \mathcal{X}^n.$$

In this case, $p(x^n) = p(x_1)\prod_{i=2}^{n} p(x_i|x_{i-1})$.

Furthermore, a process is a Markov chain of memory order M if

$$P[X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_1 = x_1] = P[X_n = x_n | X_{n-1} = x_{n-1}, \cdots X_{n-M} = x_{n-M}]$$

$\forall n \geq M, x^n \in \mathcal{X}^n$.

Definition 3.1.3. *A Markov chain is time-invariant or homogenous if $p(x_n|x_{n-1})$ does not change with n, i.e., if*

$$P[X_n = a | X_{n-1} = b] = P[X_2 = a | X_1 = b], \ \ \forall n, \forall a, b \in \mathcal{X}.$$

The widest class of sources that we have defined so far is the class of stationary sources and this property implies that two source sequences with the same pattern, even if they are far away from each other in time, occur with the same probability. In addition to stationarity, another property (which we do not explicitly define) that is important in information theory is the following:

Stationary sources that cannot be separated into different persisting (asymptotic) modes of behavior are known as ergodic sources. A stationary ergodic source has the property that the statistical average of a function defined on its random variable sequence is arbitrarily close to its time average with probability close one as the sequence length approaches infinity.

Although throughout this thesis we mainly consider stationary-ergodic sources it is worth to note that any nonergodic stationary stochastic process can be decomposed into ergodic components (the *ergodic decomposition of a stationary source*) [13]. The following theorem concerning stationary ergodic sources is called the *individual ergodic theorem* by G.D Birkhoff [14].

Theorem 3.1.1. *Let $\boldsymbol{X} = \{X_i\}_{i=1}^{\infty}$ be an arbitrary stationary source. Then, $\boldsymbol{X}$ is a stationary ergodic if and only if, for any natural number k and any integrable function*

f on $\mathcal{X}^k$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} f(X_{i+1}, X_{i+2}, \cdots, X_{i+k}) = E[f(X_1, X_2, \cdots, X_k)] \tag{3.1}$$

with probability one, where E denotes expectation.

The left- and right-hand side of equation (3.1) are called the *time-average* and *ensemble average* of f respectively.

To study the characteristics of a stationary stochastic process $\mathbf{X} = (X_1, X_2, \cdots)$ as a model of an information source, it is necessary to know how the entropy of its finite blocks, $X^n = (X_1, X_2, \cdots, X_n)$, grows with length n. In the previous chapter, we showed that it is enough to know the entropy of input, output and noise processes to find the capacity of memoryless channels. However, while working with channels with memory, we need to find the *entropy-rate* of these processes.

Definition 3.1.4. *The entropy rate of a source $\{X_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{X}$ is denoted by $H(\mathcal{X})$ or $H_\infty(X)$ and defined by;*

$$H(\mathcal{X}) := \lim_{n\to\infty}\frac{1}{n}H(X_1, X_2, \cdots, X_n)$$

provided the limit exists.

By the definition, one can see that for a DMS $H(\mathcal{X}) = H(X_1)$. For a stationary source, the limit actually exists and it coincides with the limit of conditional entropy conditioned on the previous data, i.e., we have the following result

Theorem 3.1.2. *[8] For a stationary discrete source $\boldsymbol{X} = (X_1, X_2, \cdots)$ satisfying $H(X_1) < \infty$, the entropy rate exists and is expressed by*

$$H(\mathcal{X}) \equiv \lim_{n\to\infty} H^n(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}). \tag{3.2}$$

We state Cesàro's theorem which we use in the proof of the above theorem.

Theorem 3.1.3. *If a sequence of numbers (α_n) converges to α as $n \to \infty$, the sequence*

$$(\beta_n = \frac{1}{n}\sum_{i=1}^{n} \alpha_i)$$

converges to the same value α as $n \to \infty$.

Proof of Theorem 3.1.2. The sequence of conditional entropies α_n=$H(X_n|X_1, \cdots, X_{n-1})$ is non-increasing since

$$\begin{aligned} H(X_n|X_1, X_2, \cdots, X_{n-1}) &= H(X_{n+1}|X_2, \cdots, X_n) && (3.3) \\ &\geq H(X_{n+1}|X_1, X_2, \cdots, X_n) && (3.4) \end{aligned}$$

where (3.3) follows from stationarity and (3.4) is valid since conditioning reduces entropy and $H(X_{n+1}|X_1, X_2, \cdots, X_n) \geq 0$. Thus by the monotone convergence theorem, the sequence (α_n) converges to a value α. From the chain rule of entropy, we have

$$\begin{aligned} \beta_n &= \frac{1}{n}H(X_1, X_2, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1, X_2, \cdots, X_{i-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \alpha_i \end{aligned}$$

Therefore, from Cesàro's theorem, β_n converges to α. The common limit α of these two sequences α_n and β_n converges to $H(\mathcal{X})$. □

In the rest of this chapter, we derive a more general capacity formulation [21] that covers arbitrary classes of channels with memory. To achieve this objective, we need to define so-called *information spectrum measures* which will be playing a key role in the general channel capacity theorem. The material described in the remainder of this chapter is synthesized from [7, 6, 21, 20].

3.2 Information Spectrum Measures

Consider a general source with memory (not necessarily stationary, ergodic) taking values in a finite alphabet $\mathcal{X}$. This general source may exhibit *distinct statistics* for each block length n:

$$\begin{aligned} n = 1 &\ :\ X_1^1 \\ n = 2 &\ :\ X_1^2, X_2^2 \\ n = 3 &\ :\ X_1^3, X_2^3, X_3^3 \\ &\ \vdots \end{aligned}$$

In other words, the source consists of a *triangular array* of random variables. Let us denote it by

$$\mathbf{X} := \{X^n = (X_1^{(n)}, X_2^{(n)}, \cdots, X_n^{(n)})\}_{n=1}^{\infty}.$$

This general source, which models a wide class of real-time varying sources, does not need to satisfy the *consistency* condition which is defined as follows.

From a physical point of view, the most fundamental characteristic of a random process is the set

$$F_{X_{t_1},X_{t_2},\cdots,X_{t_n}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n}) = P(X_{t_1} \leq x_{t_1}, X_{t_2} \leq x_{t_2}, \cdots, X_{t_n} \leq x_{t_n}) \qquad (3.5)$$

defined for all sets $t_1, \cdots, t_n$ such that $t_1 < t_2 < \cdots < t_n$. We see from (3.5) that for each set $t_1, \cdots, t_n$ with $t_1 < t_2 < \cdots < t_n$, the functions $F_{X_{t_1},X_{t_2},\cdots,X_{t_n}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n})$ are n-dimensional distribution functions and that the collection

$$\{F_{X_{t_1},X_{t_2},\cdots,X_{t_n}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n})\}$$

is said to be consistent if the following condition is satisfied

$$F_{X_{t_1},X_{t_2},\cdots,X_{t_n}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n}) = F_{X_{t_1},X_{t_2},\cdots,X_{t_n},X_{t_{n+1}}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n}, \infty).$$

Sources satisfying this consistency condition are usually called *processes* and let us denote them $\{X^n = (X_1, X_2, \cdots, X_n)\}_{n=1}^{\infty}$. However, for the rest of this section we consider general, not necessarily consistent, sources.

Definition 3.2.1. ***Liminf in probability***: *For an arbitrary real-valued sequence of random variables* $\{A^n = (A_1^{(n)}, A_2^{(n)}, \cdots, A_n^{(n)})\}_{n=1}^{\infty}$*, the liminf in probability* $\underline{U}$ *of a sequence of random variables* $\{A_n\}$ *is defined as the largest extended real number* $(u \in \boldsymbol{R} \cup \{-\infty, +\infty\})$ *such that*

$$\forall \epsilon > 0, \lim_{n\to\infty} P[A_n \le \underline{U} - \epsilon] = 0.$$

Equivalently

$$\underline{U} := p - \liminf_{n\to\infty} A_n := \sup\{\beta : \lim_{n\to\infty} P[A_n < \beta] = 0\}.$$

Limsup in probability: *Similarly, the limsup in probability* $\overline{U}$ *of a sequence of random variables* $\{A_n\}$ *is defined as the smallest extended real number such that*

$$\forall \epsilon > 0, \lim_{n\to\infty} P[A_n \ge \overline{U} + \epsilon] = 0.$$

Equivalently,

$$\overline{U} := p - \limsup_{n\to\infty} A_n := \inf\{\alpha : \lim_{n\to\infty} P[A_n > \alpha] = 0\}.$$

Let us now look at some properties of lim inf/lim sup in probability.

- $\underline{U} := p - \liminf_{n\to\infty} A_n$ and $\overline{U} := p - \limsup_{n\to\infty} A_n$ always exists. Furthermore,

$$p - \liminf_{n\to\infty} A_n = p - \limsup_{n\to\infty} A_n = C \Leftrightarrow p - \lim_{n\to\infty} A_n = C$$

means that A_n converges in probability to a constant C since

$$\begin{aligned} & A_n \quad \xrightarrow{n \to \infty} \quad C \text{ in probability} \\ & \Leftrightarrow \lim_{n\to\infty} P[|A_n - C| > \epsilon] = 0 \forall \epsilon > 0 \\ & \Leftrightarrow \lim_{n\to\infty} P[A_n > C + \epsilon] = 0 \text{ and} \lim_{n\to\infty} P[A_n < C - \epsilon] = 0 \forall \ \epsilon > 0. \end{aligned}$$

- $p-liminf$ and $p-limsup$ are extended notions of $\liminf$ and $\limsup$ when A_n is a deterministic real-valued sequence. They indeed have properties that are similar to $\liminf$ and $\limsup$ operations as follows.

  (a)

$$\begin{aligned} p - \liminf_{n\to\infty} A_n + p - \liminf_{n\to\infty} B_n &\leq p - \liminf_{n\to\infty}(A_n + B_n) \\ &\leq p - \liminf_{n\to\infty} A_n + p - \limsup_{n\to\infty} B_n \\ &\leq p - \limsup_{n\to\infty}(A_n + B_n) \\ &\leq p - \limsup_{n\to\infty} A_n + p - \limsup_{n\to\infty} B_n. \end{aligned}$$

  (b) $p - \lim_{n\to\infty} \sup(-A_n) = -p - \lim_{n\to\infty} \inf(A_n)$.

These quantities can be better understood by examining two related definitions.

Definition 3.2.2. *If $\{A_n\}_{n=1}^{\infty}$ is a sequence of random variables, then its inf-spectrum $\underline{u}(\cdot)$ and its sup-spectrum $\overline{u}(\cdot)$ are defined by*

$$\underline{u}(\theta) := \liminf_{n\to\infty} P(A_n \leq \theta),$$

and

$$\overline{u}(\theta) := \limsup_{n\to\infty} P(A_n \leq \theta),$$

where $\theta \in R$. In other words, $\underline{u}(\cdot)$ and $\overline{u}(\cdot)$ are respectively the liminf and the limsup of the cumulative distribution function (CDF) of A_n [6].

It should be noted that by the definition of CDF both $\underline{u}(.)$ and $\overline{u}(.)$ are non-decreasing functions. From the definition of $\underline{U}$, we have

$$\begin{aligned} \underline{U} &:= \sup\{\beta : \lim_{n\to\infty} P[A_n < \beta] = 0\} \\ &= \sup\{\beta : \lim_{n\to\infty} \sup P[A_n < \beta] = 0\}. \end{aligned}$$

However, it can be shown that

$$\sup\{\beta : \limsup_{n\to\infty} P[A_n < \beta] = 0\} = \sup\{\beta : \limsup_{n\to\infty} P[A_n \leq \beta] = 0\},$$

therefore, we obtain that $\underline{U} = \sup\{\beta : \underline{u}(\beta) = 0\}$. In other words, $\underline{U}$ is the largest extended real number for which the sup-spectrum of A_n vanishes. Furthermore, from the definition of $\overline{U}$, we have

$$\begin{aligned} \overline{U} &:= \inf\{\alpha : \lim_{n\to\infty} P[A_n > \alpha] = 0\} \\ &= \inf\{\alpha : \limsup_{n\to\infty} P[A_n > \alpha] = 0\} \\ &= \inf\{\alpha : \liminf_{n\to\infty} P[A_n \leq \alpha] = 1\} \\ &= \inf\{\alpha : \underline{u}(\alpha) = 1\} \\ &= \sup\{\alpha : \underline{u}(\alpha) < 1\} \qquad (3.6) \end{aligned}$$

where (3.6) is due to that $\underline{u}$ is non-decreasing. In other words,

$$\overline{U} = \inf\{\alpha : \underline{u}(\alpha) = 1\} = \sup\{\alpha : \underline{u}(\alpha) < 1\}.$$

The above Han and Verdú quantities given in [21] and were generalized by Chen and Alajaji in [6] in terms of ”quantiles” of information spectrum which enabled the latter authors to establish ”optimistic” source and channel coding operational quantities [7].

Definition 3.2.3. *Consider a general source* $\boldsymbol{X} := \{X^n = (X_1^{(n)}, X_2^{(n)}, \cdots, X_n^{(n)})\}$ *with alphabet* $\mathcal{X}$. *Then, the random variable* $\frac{-1}{n}\log P_{X^n}(X^n)$ *is called the normalized entropy density of the source and is usually denoted by* $\frac{1}{n}h_{X^n}(X^n)$. *Note that the* *expectation* $E_{X^n}[\frac{1}{n}h_{X^n}(X^n)] = \frac{1}{n}H(X^n)$, *which is the normalized entropy of* X^n.

Definition 3.2.4. *The inf-entropy rate (or the spectral inf-entropy rate) of the source, denoted by* $\underline{H}(\mathcal{X})$, *is defined as,*

$$\begin{aligned}\underline{H}(X) &:= p-\lim_{n\to\infty}\inf\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)}\\ &= \sup\left\{\beta : \limsup_{n\to\infty} P\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} \le \beta\right] = 0\right\}\end{aligned}$$

Similarly, the sup-entropy rate of the source, denoted by $\overline{H}(\mathcal{X})$, *is defined as,*

$$\begin{aligned}\overline{H}(X) &:= p-\lim_{n\to\infty}\sup\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)}\\ &= \sup\left\{\beta : \liminf_{n\to\infty} P\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} \le \beta\right] < 1\right\}\end{aligned}$$

Another important definition given in [21] which plays a key role in proving generalized source/channel coding theorems is so called inf/sup information rate. Before defining this, we first define a related information density quantity.

Definition 3.2.5. *Let* $\boldsymbol{X} := \{X^n\}_{n=1}^{\infty}$ *be a general input source with finite alphabet* $\mathcal{X}$ *and let*

$$\boldsymbol{Y} := \{Y^n := (Y_1^{(n)}, Y_2^{(n)}, \cdots, Y_n^{(n)})\}$$

be the corresponding output source with alphabet $\mathcal{Y}$ *induced by source* $\boldsymbol{X}$ *via the channel*

$$\boldsymbol{W} := \{W^n = P_{Y^n|X^n} : \mathcal{X}^n \to \mathcal{Y}^n\}_{n=1}^{\infty}$$

which is an arbitrary sequence of n*-dimensional conditional distributions from* $\mathcal{X}^n$ *to* $\mathcal{Y}^n$ *satisfying*

$$\sum_{y^n\in\mathcal{Y}^n} W^n(y^n|x^n) = 1, \ \ \forall x^n \in \mathcal{X}^n, \forall n = 1, 2\cdots.$$

Then the random variable

$$\frac{1}{n}\log\frac{W^n(Y^n|X^n)}{P_{Y^n}(Y^n)} = \frac{1}{n}\log\frac{P_{X^n,Y^n}(X^n,Y^n)}{P_{X^n}(X^n)P_{Y^n}(Y^n)}$$

is called the normalized information density of the channel and is usually denoted by $\frac{1}{n} i_{X^nW^n}(X^n, Y^n)$. *It should be noted that*

$$E_{X^nW^n}[\frac{1}{n} i_{X^nW^n}(X^n, Y^n)] = \frac{1}{n} I(X^n; Y^n),$$

which is the normalized mutual information between X^n *and* Y^n.

Definition 3.2.6. *The inf-information rate (or spectral inf-information rate), denoted by* $\underline{I}(\boldsymbol{X}; \boldsymbol{Y})$ *is defined as*

$$\begin{aligned} \underline{I}(\boldsymbol{X}; \boldsymbol{Y}) &:= p - \lim_{n\to\infty} \inf \frac{1}{n} i_{X^nW^n}(X^n, Y^n) \\ &= \sup\left\{\beta : \limsup_{n\to\infty} P\left[\frac{1}{n} i_{X^nW^n}(X^n, Y^n) \leq \beta\right] = 0\right\}. \end{aligned}$$

The sup-information rate (or spectral sup-information rate), denoted by $\overline{I}(\boldsymbol{X}; \boldsymbol{Y})$ *is defined as*

$$\overline{I}(\boldsymbol{X}; \boldsymbol{Y}) := p - \lim_{n\to\infty} \sup \frac{1}{n} \log \frac{W^n(Y^n|X^n)}{P_{Y^n}(Y^n)}.$$

Before stating the generalized channel capacity theorem, we first discuss some important properties of inf-information rate. Many of the familiar properties that mutual information satisfies turn out to be inherited by the inf-information rate. Those properties are particularly useful in the computation of $\sup_X \underline{I}(\mathbf{X}; \mathbf{Y})$ for some specific channels. In deriving these properties one of the frequently used properties is the non-negativity of *divergence*. Therefore, let us first define the divergence.

Definition 3.2.7. *The divergence (also called relative entropy) between two pmfs* $p(.)$ *and* $q(.)$ *given over the same alphabet* $\mathcal{X}$ *is defined as*

$$D(p\|q) := \sum_{x\in\mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} = E_p\left[\log_2 \frac{p(x)}{q(x)}\right].$$

Divergence is a measure of "distance" between distributions p and q; it is a measure of the inefficiency of assuming that the distribution of a random variable X *is* $q(.)$

when its true distribution is $p(.)$

Lemma 3.2.1. *$D(p\|q) \geq 0$ with equality if and only if $p = q \quad (p(x) = q(x), \forall x \in \mathcal{X})$.*

The proof of this lemma directly follows the definition. In a similar way, we can also define inf-divergence rate for two arbitrary processes **U** and **V**.

Definition 3.2.8. *The inf-divergence rate for two arbitrary processes $\boldsymbol{U}$ and $\boldsymbol{V}$, $\underline{D}(\boldsymbol{U}\|\boldsymbol{V})$, is given as the liminf in probability of the sequence of the log-likelihood ratios $\frac{1}{n}\log\frac{P_{U^n}(U^n)}{P_{V^n}(V^n)}$ [21].*

In the next theorem, we state the properties of inf-information rate. Their proof is available in [21].

Theorem 3.2.1. *An arbitrary sequence of joint distributions $(\boldsymbol{X}, \boldsymbol{Y})$ satisfies*

(a) $\underline{D}(\boldsymbol{X}\|\boldsymbol{Y}) \geq 0$.

(b) $\underline{I}(\boldsymbol{X}; \boldsymbol{Y}) = \underline{I}(\boldsymbol{Y}; \boldsymbol{X})$.

(c) $\underline{I}(\boldsymbol{X}; \boldsymbol{Y}) \geq 0$.

(d)

$$\begin{aligned}
\underline{I}(\boldsymbol{X}; \boldsymbol{Y}) &\leq \underline{H}(\boldsymbol{Y}) - \underline{H}(\boldsymbol{Y}|\boldsymbol{X}) \\
\underline{I}(\boldsymbol{X}; \boldsymbol{Y}) &\leq \overline{H}(\boldsymbol{Y}) - \overline{H}(\boldsymbol{Y}|\boldsymbol{X}) \\
\underline{I}(\boldsymbol{X}; \boldsymbol{Y}) &\geq \underline{H}(\boldsymbol{Y}) - \overline{H}(\boldsymbol{Y}|\boldsymbol{X}).
\end{aligned}$$

(e) $0 \leq \overline{H}(\boldsymbol{Y}) < \log|\mathcal{Y}|$.

(f) $\underline{I}(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{Z}) \geq \underline{I}(\boldsymbol{X}; \boldsymbol{Z})$.

(g) If $\overline{I}(\boldsymbol{X}; \boldsymbol{Y}) = \underline{I}(\boldsymbol{X}; \boldsymbol{Y})$ *and the input alphabet is finite, then* $\underline{I}(\boldsymbol{X}; \boldsymbol{Y}) = \lim_{n\to\infty}\frac{1}{n}I(X^n; Y^n)$.

(h) $\underline{I}(\boldsymbol{X};\boldsymbol{Y}) \leq \liminf_{n\to\infty} \frac{1}{n} I(X^n;Y^n)$.

Using these properties, we can define an important theorem the analogue of which is also defined on standard mutual information.

Theorem 3.2.2 (Data Processing Theorem for Inf-Information Rate)**.** *Suppose that for every n, X_1^n and X_3^n are conditionally independent given X_2^n. Then,*

$$\underline{I}(\boldsymbol{X}_1;\boldsymbol{X}_3) \leq \underline{I}(\boldsymbol{X}_1;\boldsymbol{X}_2). \tag{3.7}$$

Proof. By Theorem 3.2.1, we get

$$\begin{aligned}\underline{I}(\mathbf{X}_1;\mathbf{X}_3) &\leq \underline{I}(\mathbf{X}_1;\mathbf{X}_2,\mathbf{X}_3) \\ &= \underline{I}(\mathbf{X}_1;\mathbf{X}_2) \qquad (3.8)\end{aligned}$$

where the equality holds because $\underline{I}(\mathbf{X}_1;\mathbf{X}_2,\mathbf{X}_3)$ is the liminf in probability of

$$\begin{aligned}\frac{1}{n}\log\frac{P_{X_1^n|X_2^nX_3^n}(X_1^n|X_2^n,X_3^n)}{P_{X_1^n}(X_1^n)} &= \frac{1}{n}\log\frac{P_{X_1^n|X_2^n}(X_1^n|X_2^n)}{P_{X_1^n}(X_1^n)} + \frac{1}{n}\log\frac{P_{X_1^n|X_2^nX_3^n}(X_1^n|X_2^n,X_3^n)}{P_{X_1^n|X_2^n}(X_1^n|X_2^n)} \\ &= \frac{1}{n}\log\frac{P_{X_1^n|X_2^n}(X_1^n|X_2^n)}{P_{X_1^n}(X_1^n)}. \qquad (3.9)\end{aligned}$$

□

3.3 General Channel Coding Theorem

Let us consider a general channel with memory described by $\mathbf{W} := \{W^n = P_{Y^n|X^n} : \mathcal{X}^n \to \mathcal{Y}^n\}_{n=1}^{\infty}$ which is an arbitrary sequence of n-dimensional distributions from $\mathcal{X}^n$ to $\mathcal{Y}^n$, where both $\mathcal{X}$ and $\mathcal{Y}$ are finite alphabet, such that

$$\sum_{y^n\in\mathcal{Y}^n} W^n(y^n|x^n) = 1, \;\; \forall x^n \in \mathcal{X}^n, \forall n = 1,2\cdots.$$

Let $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ and $\mathbf{Y} = \{Y_i\}_{i=1}^{\infty}$ denote the input and output sources of the channel, respectively.

Definition 3.3.1 (Channel Block Code). *An (M, n) block code for the channel $\boldsymbol{W} := \{W^n = P_{Y^n|X^n}\}_{n=1}^{\infty}$ consists of:*

- *Encoder:* $f_n : \mathcal{M} := \{1, 2, \cdots, M\} \to \mathcal{X}^n$.
- *Decoder:* $g_n : \mathcal{Y}^n \to \mathcal{M}$.

The codebook is given by

$$C_n = \{f_n(1), \cdots, f_n(M)\} = \{x^n(1), \cdots, x^n(M)\} \subseteq \mathcal{X}^n.$$

The code rate $R(C_n) = \frac{1}{n}\log_2 M$ message bits/channel symbols and the average probability of error is

$$\begin{aligned} P_e(C_n) &:= \frac{1}{M}\sum_{i=1}^{M}\sum_{y^n : g_n(y^n)\neq i} P_{Y^n|X^n}(y^n|f_n(i)) \\ &= P[I \neq \hat{I}], \end{aligned}$$

where $I \in \mathcal{M}$ is uniform and $\hat{I}$ is the decoder output.

Although we defined channel capacity in the previous chapter we herein restate it for the channels with memory.

Definition 3.3.2 ((Operational) Channel Capacity). *$R \geq 0$ is said to be achievable channel coding rate for the channel $\boldsymbol{W}$ if there exists a sequence of $\{C_n = (M, n)\}_{n=1}^{\infty}$ block codes for $\boldsymbol{W}$ such that*

$$\liminf_{n\to\infty} \frac{1}{n}\log M \geq R \quad \textit{and} \quad \limsup_{n\to\infty} P_e(C_n) = 0.$$

The supremum of all achievable channel coding rates for $\boldsymbol{W}$ is denoted by C and called the (operational) channel capacity:

$$C := \sup\{R \geq 0 : R \quad \textit{is achievable for channel} \quad \boldsymbol{W}\}.$$

Next, we state two lemma's on the average probability of error.

Lemma 3.3.1 (Feinstein's Lemma)**.** *Fix a positive integer n. For every $\gamma > 0$ and input distribution P_{X^n} on $\mathcal{X}^n$, there exists an (M, n) block code C_n for $\boldsymbol{W^n} = P_{Y^n|X^n}$ such that its average error probability satisfies*

$$P_e(C_n) < P[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) < \frac{1}{n}\log M + \gamma] + \exp(-n\gamma)$$

where $i_{X^n W^n}(X^n; Y^n) := \frac{1}{n}\log\frac{P_{Y^n|X^n}(Y^n|x^n)}{P_{Y^n}(Y^n)}$ is the normalized information density [11].

Lemma 3.3.2 (Verdú - Poor Channel Coding Lemma)**.** *Every (M_n, n) block code C_n for channel $\boldsymbol{W^n} = P_{Y^n|X^n}$ satisfies*

$$P_e(C_n) \geq (1 - \exp(-n\gamma))P[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) < \frac{1}{n}\log M - \gamma]$$

for every $\gamma > 0$, where X^n places probability $\frac{1}{M}$ on each codeword of C_n [22].

Lemmas 3.3.1 and 3.3.2 gives a lower and upper bound on the average probability of error in terms of the normalized density function. We can now state the general channel coding theorem of Verdú and Han.

Theorem 3.3.1 (General Channel Coding Theorem)**.** *[21] For* ***any*** *channel given by $\boldsymbol{W} := \{W^n = P_{Y^n|X^n}\}_{n=1}^{\infty}$,*

$$C = \sup_{\boldsymbol{X}} \underline{I}(\boldsymbol{X}; \boldsymbol{Y})$$

In other words, for any arbitrary channel $\boldsymbol{W}$, the capacity is given by the supremum over all input sources of the inf-information rate $\underline{I}(\boldsymbol{X}; \boldsymbol{Y})$ [21].

3.4 Application to Information Stable Channels

Let us start with a modification on the definition of information stability from sources to channels [2].

Definition 3.4.1 (Information Stable Channel)**.** *A channel $\boldsymbol{W} := \{W^n = P_{Y^n|X^n}\}_{n=1}^{\infty}$ is said to be information stable if there exists an input source $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ such that $O < C^n < \infty$ for n sufficiently large and*

$$\limsup_{n\to\infty} P\left[|\frac{\frac{1}{n}i_{X^nW^n}(X^n;Y^n)}{C^n} - 1| > \gamma\right] = 0, \quad \forall\gamma > 0$$

where $C^n := \sup_{P(X^n)} \frac{1}{n}I(X^n;Y^n)$.

Remark.

- *DMC's are information stable.*
- *More generally, stationary ergodic channels are information stable. It should be noted that, a channel is called stationary (respectively ergodic) if for every stationary input source (respectively ergodic), the resulting joint input-output process is stationary (respectively ergodic).*
- *A channel with (modulo) additive stationary ergodic noise is information stable.*
- *A channel with non-stationary independent (modulo) additive noise is information stable.*

Theorem 3.4.1. *[21] Every information stable channel $\boldsymbol{W} := \{W^n = P_{Y^n|X^n}\}_{n=1}^{\infty}$ satisfies*

$$C = \liminf_{n\to\infty} C^n = \liminf_{n\to\infty} \sup_{P(X^n)} \frac{1}{n}I(X^n;Y^n).$$

In the next two section, we will be using this theorem while computing the channel capacity.

Chapter 4

A Discrete Non-Binary Noise Channel

In this chapter, we consider a new binary-input non-binary output channel with memory recently introduced in [15] to model soft-decision demodulated time-correlated fading channels. We first study the non-feedback capacity of this channel.

4.1 Channel Model

In this section, we first define the communication system model considered in [15] and next we state an equivalent discrete channel model to this fading channel.

4.1.1 A Discrete Fading Channel with Soft-Decision Information

Wireless communication channels undergo time-varying fading which can be modeled as a time-correlated random process. Moreover, since each fading statistically depends on the previous one, this stochastic process exhibits memory. Considering this memory embedded in the process, a discrete binary-input 2^q-ary output communication channel with memory is introduced [15] where the objective is to capture both the statistical memory and the soft-decision information of time-correlated fading channels modulated by binary phase-shift keying (BPSK) and coherently demodulated with an output quantizer of resolution q. The main motivation of this channel model is that it may be used in designing new coding/decoding schemes for soft-decision demodulated channels with memory that result in superior performance over systems that ignore the channel's memory (via interleaving) and/or soft-decision information (via hard demodulation)[15]. Additionally, the receivers operating with 1-3 bit quantization have potential applications in ultrawideband and millimeter wave communication.

The discrete fading channel (DFC) is composed of a BPSK modulator, a time-correlated flat fading channel and a q-bit soft-quantized coherent demodulator [15].

The complex envelope of the fading process, $\tilde{G}(t)$, is a zero-mean stationary Gaussian noise process with known covariance. Let, $\{X_k\} \in \mathcal{X}, k = 1, 2, \cdots$, be the input process to the discrete channel. The sample of the fading envelope at the kth interval, $A_k = |\tilde{G}(kT)|$, where T is symbol interval, has the Rayleigh density function with a unit second moment. At the kth signaling interval, the symbol received at the output of the matched filter is written as;

$$R_k = \sqrt{E_s} A_k S_k + N_k \quad k = 1, 2, \cdots ,$$

where $S_k = 2X_k - 1$, E_s is the energy of the transmitted signal, N_k is a sequence of i.i.d zero-mean Gaussian random variables with variance $N_0/2$ and A_k is a stationary time-correlated Rayleigh process. The processes $\{A_k\}$ and $\{N_k\}$ are independent of each other and also of the input process. The channel output,$Y_k \in \mathcal{Y}$ is obtained with demodulating the random variable R_k via a q-bit uniform scalar quantizer as follows;

$$Y_k = j \quad \text{if} \quad R_k \in (T^{'}_{j-1}, T^{'}_{j})$$

for $j \in \mathcal{Y}$. The thresholds $T^{'}_{j}$ are uniformly spaced with step size Δ, satisfying [5]

$$T^{'}_{j} = \begin{cases} -\infty & \text{if} \quad j = -1 \\ (j + 1 - 2^{q-1})\triangle & \text{if} \quad j = 0, 1, \cdots, 2^q - 1 \\ \infty & \text{if} \quad j = 2^q - 1 \end{cases}$$

To normalize step size and thresholds let $\delta = \Delta\sqrt{E_s}$ and $T_j = T^{'}_{j}/\sqrt{E_s}$. Then, $T_j = (j + 1 - 2^{q-1})\delta$ for $j = 0, 1, \cdots, 2^q - 1$.

We can now determine the conditional probability, $q_{i,j}(a_k) = Pr(Y_k = j|X_k = i, A_k = a_k)$, where $i \in \mathcal{X}$, $j \in \mathcal{Y}$ and $a_k \in [0, \infty)$, as follows;

$$\begin{aligned} q_{i,j}(a_k) &= Pr(T^{'}_{j-1} < R_k < T^{'}_{j}|X_k = i, A_k = a_k) \\ &= Pr(T_{j-1} - (2i-1)a_k < \frac{N_k}{\sqrt{E_s}} < T_j - (2i-1)a_k) \\ &= Q(\sqrt{2\gamma}(T_{j-1} - (2i-1)a_k)) - Q(\sqrt{2\gamma}(T_j - (2i-1)a_k)) \qquad (4.1) \end{aligned}$$

where $\gamma = E_s/N_0$ is the signal-to-noise ratio (SNR) and $Q(x) = 1/\sqrt{2\pi}\int_x^\infty \exp{(-t^2/2)}dt$ is the Gaussian Q-function. Due to the symmetry of the BPSK constellation and the quantizer thresholds, we observe from (4.1) that $q_{i,j}(a_k) = q_{1-i,2^q-1-j}(a_k)$. This can also be written as;

$$q_{i,j}(a_k) = q_{0,\frac{j-(2^q-1)}{(-1)^i}}(a_k)$$

for $i \in \mathcal{X}$ and $j \in \mathcal{Y}$. For integer $n \geq 1$, let $Pr(y^n|x^n, a^n)$ denote the n-fold probability distribution. Then,

$$Pr(y^n|x^n, a^n) = \prod_{k=1}^{n} q_{x_k,y_k}(a_k) = \prod_{k=1}^{n} q_{0,\frac{y_k-(2^q-1)x_k}{(-1)_k^x}}(a_k) \tag{4.2}$$

Thus, the DFC is specified in terms of the channel block conditional probability

$$\begin{aligned} P_{DFC}^{(n)}(y^n|x^n) &= Pr(Y^n = y^n|X^n = x^n) \\ &= \mathbf{E}_{A_1\ldots A_n}\left[\prod_{k=1}^{n} q_{0,\frac{y_k-(2^q-1)x_k}{(-1)^{x_k}}}(A_k)\right] \end{aligned} \tag{4.3}$$

where $y^n = (y_1, \cdots, y_n)$ and $E_X[.]$ denotes the expectation over X. For $n = 1$, a closed form expression for $P_{DFC}^{(j)}$, $j \in \mathcal{Y}$, is given by [19]

$$P_{DFC}^{(j)} = m(-T_{j-1}) - m(-T_j) \tag{4.4}$$

where

$$m(T_j) = 1 - Q(T_j\sqrt{2\gamma}) - \frac{[1 - Q(\frac{T_j\sqrt{2\gamma}}{\sqrt{\frac{1}{\gamma}+1}})]e^{-\frac{T_j^2}{(\frac{1}{\gamma}+1)}}}{\sqrt{\frac{1}{\gamma}+1}} \tag{4.5}$$

The expected value in (4.3) can be directly calculated for $n \leq 3$ and for $n > 3$ it can be determined via simulations.

4.2 An Alternative Model to DFC

In general, it is convenient to express the channel output process as an explicit function of input and noise processes. Pimentel and Alajaji in [15] developed an alternative model to the above soft-demodulated discrete fading channel. In this subsection, we state this equivalent model and in the next section we consider this model with feedback and show that feedback does not increase the capacity for this channel.

Consider the following non-binary noise discrete channel (NBNDC)

$$Y_k = (2^q - 1)X_k + (-1)^{X_k} Z_k \tag{4.6}$$

for $k = 1, 2, \cdots$, where $\{X_k\}$ is the input process, $\{Y_k\}$ is the output process and $\{Z_k\}$ is the noise process. Here the input $X_k \in \mathcal{X} = \{0, 1\}$ is binary, and both noise and output symbols, Z_k and Y_k, take values from the same 2^q-ary alphabet given by $\mathcal{Z} = \mathcal{Y} = \{0, 1, \cdots, 2^q - 1\}$. It is also assumed that the noise and input processes are independent of each other.

The noise process is governed by the n-fold distribution

$$P^n_{NBNDC}(z^n) := P^n_{NBNDC}(z_1, \cdots, z_n)$$

where $z_k \in \mathcal{Y}$. Since the input and noise processes are independent of each other, looking at (4.6) it can be seen that

$$\begin{aligned} P^n_{NBNDC}(y^n|x^n) &= P^n_{NBNDC}(z^n) && (4.7) \\ \text{where } z_k &= \frac{y_k - (2^q - 1)x_k}{(-1)^{x_k}}, \quad k = 1, \cdots, n. && (4.8) \end{aligned}$$

Now, it should be noted that if the distribution of noise process $\{Z_k\}$ in (4.8) is given by (4.3) for each n, then the discrete fading channel and NBNDC have the same channel block conditional probability. Thus, NBNDC provides an alternative representation of the DFC. It can also be seen that when $q = 1$ (hard-decision demodulation), then the NBNDC expression in (4.6) gives us a familiar expression

$$Y_k = X_k \oplus Z_k$$

where $\oplus$ denotes modulo-2 addition. In other words, when $q = 1$, the NBNDC reduces to the binary (modulo-2) additive noise discrete channel with memory. Furthermore, when $\{Z_k\}$ is memoryless, we obtain the memoryless BSC which fully represents the fully interleaved discrete fading channel.

We now state some properties of channel which will be used frequently. The NBNDC, as described by $Y_k = f(X_k, Z_k)$, where $f(\cdot, \cdot)$ is given in (4.6), satisfies the following “invertibility” properties:

(a) For any fixed input $x \in \mathcal{X}$, $f(x, \cdot) : \mathcal{Z} \to \mathcal{Y}$ is invertible.

(b) Every output symbol is the image of exactly two distinct input-noise pairs; i.e., for any $y \in \mathcal{Y}$, there are exactly two pairs (x_1, z_1) and (x_2, z_2) in $\mathcal{X} \times \mathcal{Z}$ such that $x_1 \neq x_2$, $z_1 \neq z_2$ and $y = f(x_1, z_1) = f(x_2, z_2)$.

It should be noted that, when the input alphabet is binary property (b) implies property (a). We continue our analysis by computing the channel capacity for this model when the noise process is stationary ergodic.

4.3 Capacity Without Feedback

Consider the NBNDC given by (4.6), where the noise process is stationary ergodic. For this information stable channel, its non-feedback capacity, in bits per channel use, is given by (see Theorem 3.3.1)

$$C = \liminf_{n\to\infty} C^{(n)} = \lim_{n\to\infty} C^{(n)} \tag{4.9}$$

where

$$C^{(n)} = \max_{p(x^n)} \frac{1}{n} I(X^n; Y^n)$$

where maximum is taken with respect to all input distributions and $I(X^n; Y^n)$ is the block mutual information. Since $\{X_k\}$ and $\{Z_k\}$ are independent of each other, the block mutual information can be rewritten as;

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) = H(Y^n) - H(Z^n)$$

Therefore,

$$C^{(n)} = \frac{1}{n}\left(\max_{p(x_n)}[H(Y^n) - H(Z^n)]\right) \tag{4.10}$$

At this point, to find the capacity it is only required to find the distribution maximizing $H(Y^n)$. Let us look at this distribution.

Definition 4.3.1. *Let $\mathcal{W} = \{0, 1, \cdots, 2^{q-1} - 1\}$ and let $\{W_k\}$, $W_k \in \mathcal{W}$, be a process with n-fold probability distribution*

$$Pr(W^n = w^n) = \sum_{x^n \in \ \mathcal{X}^n} Pr\left(Z^n = \frac{w^n - (2^q - 1)x^n}{(-1)^{x^n}}\right) \tag{4.11}$$

where $Z^n = \frac{(w^n - (2^q-1)x^n)}{(-1)^{x^n}}$ denotes the tuple obtained from component-wise operations, i.e., $(Z_1 = \frac{(w_1-(2^q-1)x_1)}{(-1)^{x_1}}, \cdots, Z_n = \frac{(w_n-(2^q-1)x_n)}{(-1)^{x_n}})$.

It should be noted that, the mapping $g : \mathcal{W} \times \mathcal{X} \rightarrow \mathcal{Y}$ given by

$$z = g(w, x) := \frac{w - (2^q - 1)x}{(-1)^x}$$

is invertible.

We can easily check that the probability assignment in (4.11) is valid since

$$\begin{aligned} 1 &= \sum_{z^n \in \ \mathcal{Z}^n} Pr(Z^n = z^n) \\ &= \sum_{w^n \in \ \mathcal{W}^n} \sum_{x^n \in \ \mathcal{X}^n} Pr\left(Z^n = \frac{w^n - (2^q - 1)x^n}{(-1)^{x^n}}\right) \\ &= \sum_{w^n \in \ \mathcal{W}^n} Pr(W^n = w^n) \end{aligned} \tag{4.12}$$

The process $\{W_k\}$ is stationary since $\{Z_k\}$ is stationary when $\{X_k\}$ is stationary: for any integer $m > 0$ $w^n \in \mathcal{W}^n$,

$$\begin{aligned}
&Pr(W_{1+m} = w_1, \cdots, W_{n+m} = w_n)\\
&\quad = \sum_{x^n \in \ \mathcal{X}^n} Pr\left(Z_{1+m} = \frac{(w_1 + (2^q - 1)x_1)}{(-1)^{x_1}}, \cdots, Z_{n+m} = \frac{(w_n - (2^q - 1)x_n)}{(-1)^{x_n}}\right)\\
&\quad = \sum_{x^n \in \ \mathcal{X}^n} Pr\left(Z_1 = \frac{(w_1 + (2^q - 1)x_1)}{(-1)^{x_1}}, \cdots, Z_n = \frac{(w_n - (2^q - 1)x_n)}{(-1)^{x_n}}\right)\\
&\quad = Pr(W_1 = w_1, \cdots, W_n = w_n).
\end{aligned}$$

Proposition 4.3.1. *Consider the $2^n \times 2^{qn}$ channel transition probability matrix $\boldsymbol{Q}^n = [P^n_{NBNDC}(y^n|x^n)]$ corresponding to n channel uses, where each row (respectively column) of $\boldsymbol{Q}^n$ is indexed by a sequence x^n (respectively y^n). Then, $\boldsymbol{Q}^n$ is quasi-symmetric.*

Proof. During the proof, we will be using the term "weight" to mean that the input, output and noise tuples are expressed in decimal form. Thus, $x^n = (x_1, \cdots, x_n)$, $y^n = (y_1, \cdots, y_n)$ and $z^n = (z_1, \cdots, z_n)$ can be expressed (in a one-to-one correspondence) in terms of the decimal scalars

$$\begin{aligned}
\tilde{x} &= x_1 + x_2 2 + \cdots + x_n 2^{n-1}\\
\tilde{y} &= y_1 + y_2 2^q + \cdots + y_n 2^{q(n-1)}\\
\tilde{z} &= z_1 + z_2 2^q + \cdots + z_n 2^{q(n-1)}
\end{aligned}$$

respectively.

Remark. *Let $\tilde{\boldsymbol{Q}}^n$ be a matrix such that its entries are composed of the weight of noise tuples that is given by (4.8). It should be noted that, the entries of $\boldsymbol{Q}^n$ are $P(z^n)$ values and the entries of $\tilde{\boldsymbol{Q}}^n$ are the weights of noise tuples z^n. However, since for each z^n the weight is unique, to show that $\boldsymbol{Q}^n$ is quasi-symmetric, it is sufficient to show that $\tilde{\boldsymbol{Q}}^n$ is quasi-symmetric. In the rest of the proof, we show that $\tilde{\boldsymbol{Q}}^n$ is* *quasi-symmetric.*

We observe that, from (4.6), there are exactly two pairs of (x, y), (x_i, y_i) and (x_j, y_j), with $x_i \neq x_j$ and $y_i \neq y_j$, that satisfy (4.6). We refer to this property by property (c). Therefore, for the n-fold noise tuple z^n, there are 2^n possible combinations of such (x^n, y^n) pairs that satisfy (4.6) component-wise. Moreover, considering $\tilde{\mathbf{Q}}^n$, each specific weight of $z^n, \tilde{z} \in \{0, 1, 2, \ldots, 2^{qn} - 1\}$, appears exactly once in each row by the property (a). In the rest of the proof, we will show the following:

(i) Pick any column from $\tilde{\mathbf{Q}}^n$ and choose a specific entry in this column.

(ii) By the fact described above, this selected weight appears in another column (in fact in 2^n other columns).

(iii) Let us denote these two columns by y_i^n and y_j^n, respectively. Then we claim that, these two columns are permutations of each other.

(iv) By extending this idea to the other $2^{q(n-1)} - 2$ columns, we obtain a $2^n \times 2^n$ array such that its columns are permutations of each other. Furthermore, by property (a), the rows of this array are also permutations of each other.

$$\tilde{\mathbf{Q}}^n = \begin{array}{cc} & \begin{array}{ccccc} & & y_i^n & \cdots & y_j^n \end{array} \\ \begin{array}{c} \\ x_t^n \\ \\ \\ x_s^n \end{array} & \left(\begin{array}{ccccc} & & \tilde{z}_{mi} & & \\ \cdots & & & & \tilde{z}_{tj} \\ & \uparrow & \vdots & \downarrow & \vdots \\ & & & & \tilde{z}_{mj} \\ \cdots & & \tilde{z}_{si} & & \end{array} \right) \end{array} \tag{4.13}$$

The idea of the proof can be seen better in the matrix (4.13). First we select column $y_i^n = (y_{i_1}, y_{i_2}, \dots, y_{i_n})$ and select an entry at row $x_s^n = (x_{s_1}, x_{s_2}, \dots, x_{s_n})$ in this column. Let the selected weight be $\tilde{z}_{si}$ and by (ii) we know that $\tilde{z}_{si}$ appears in some other column $y_j^n = (y_{j_1}, y_{j_2}, \dots, y_{j_n})$ such that $\tilde{z}_{si} = \tilde{z}_{tj}$ and t denotes the row position of this weight. Then, we show that $\tilde{z}_{mi}$, which is an another entry in the column y_i^n also appears in column y_j^n. Let us denote this equivalent weight in column y_j^n by $\tilde{z}_{mj}$.

Let $z_{si}^n = (z_{si_1}, z_{si_2}, \dots, z_{si_n})$ and $z_{mj}^n = (z_{mj_1}, z_{mj_2}, \dots, z_{mj_n})$ be the noise tuple corresponding to weights $\tilde{z}_{si}$ and $\tilde{z}_{mj}$, respectively. Let us also assume that, there are k bits differences between x_s^n and the row corresponding to the entry $\tilde{z}_{mi}$. Let us denote the positions of these bits by $c_1, c_2, \dots, c_k$. Then,

- if the bit $x_{s_{c_l}}$ is toggled from 0 to 1, then

$$\tilde{z}_{si} - \tilde{z}^{'} = \left(2y_{i_{c_l}} - (2^q - 1)\right) 2^{q(c_l - 1)}$$

- if the bit $x_{s_{c_l}}$ is toggled from 1 to 0, then

$$\tilde{z}_{si} - \tilde{z}^{'} = \left((2^q - 1) - 2y_{i_{c_l}}\right) 2^{q(c_l - 1)}$$

where $l = 1, \ldots, k$ and $\tilde{z}^{'}$ is the new noise weight due to toggling the bit $x_{s_{c_l}}$. Thus the total difference in noise weight due toggling k bits in x_s^n is,

$$\tilde{z}_{si} - \tilde{z}_{mi} = \sum_{l=1}^{k} (-1)^{x_{s_{c_l}}} \left(2y_{i_{c_l}} - (2^q - 1)\right) 2^{q(c_l - 1)}. \tag{4.14}$$

In the rest of the proof, we show that this new weight $\tilde{z}_{mi}$ also appears in column y_j^n.

Since $\tilde{z}_{si}$ also appears in (x_t^n, y_j^n) as $\tilde{z}_{tj}$, we have that

$$\begin{aligned} \tilde{z}_{si} &= \tilde{z}_{tj} && (4.15) \\ \frac{y_{i_l} - (2^q - 1)x_{s_l}}{(-1)^{x_{s_l}}} &= \frac{y_{j_l} - (2^q - 1)x_{t_l}}{(-1)^{x_{t_l}}}, \quad l = 1, \ldots, n. && (4.16) \end{aligned}$$

Therefore,

$$
\begin{aligned}
&\tilde{z}_{si} - \tilde{z}_{mi} \\
&= \sum_{l=1}^{k} (-1)^{x_{sc_l}} \left(2y_{ic_l} - (2^q - 1)\right) 2^{q(c_l-1)} \\
&\overset{(a)}{=} \sum_{l=1}^{k} (-1)^{x_{sc_l}} \left(2\left((2^q - 1)x_{sc_l} + (-1)^{x_{sc_l}} z_{sic_l}\right) - (2^q - 1)\right) 2^{q(c_l-1)} \\
&\overset{(b)}{=} \sum_{l=1}^{k} (-1)^{1-x_{tc_l}} \left(2\left((2^q - 1)(1 - x_{tc_l}) + (-1)^{1-x_{tc_l}} z_{sic_l}\right) - (2^q - 1)\right) 2^{q(c_l-1)} \\
&\overset{(c)}{=} \sum_{l=1}^{k} (-1)^{1-x_{tc_l}} \left(2\left((2^q - 1)(1 - x_{tc_l}) + (-1)^{1-x_{tc_l}} z_{tjc_l}\right) - (2^q - 1)\right) 2^{q(c_l-1)} \\
&\overset{(d)}{=} \sum_{l=1}^{k} (-1)^{x_{tc_l}} \left((2^q - 1) - 2\left((2^q - 1)(1 - x_{tc_l}) - (-1)^{x_{tc_l}} z_{tjc_l}\right)\right) 2^{q(c_l-1)} \\
&= \sum_{l=1}^{k} (-1)^{x_{tc_l}} \left((2^q - 1) - 2(2^q - 1) + 2(2^q - 1)x_{tc_l} + 2(-1)^{x_{tc_l}} z_{tjc_l}\right) 2^{q(c_l-1)} \\
&= \sum_{l=1}^{k} (-1)^{x_{tc_l}} \left(2\left((2^q - 1)x_{tc_l} + (-1)^{x_{tc_l}} z_{tjc_l}\right) - (2^q - 1)\right) 2^{q(c_l-1)} \\
&= \sum_{l=1}^{k} (-1)^{x_{tc_l}} \left(2y_{jc_l} - (2^q - 1)\right) 2^{q(c_l-1)} \qquad (4.17)
\end{aligned}
$$

where (a) is by equation (4.6), (b) is due to (4.16), property (c) and x_{sc_l} being binary, (c) is due to (4.15) and (d) is valid since $(-1)^{1-x} = -(-1)^x$. The proof is complete since equation (4.17) shows that $\tilde{z}_{si} - \tilde{z}_{mi}$ is achieved by toggling the same coordinates of x_t^n which indicates that $\tilde{z}_{mi}$ also appears in the column y_j^n. This shows that $\tilde{\mathbf{Q}}^n$ is quasi-symmetric and by Remark (4.3), $\mathbf{Q}^n$ is also quasi-symmetric.

□

Since the channel transition matrix for the channel given by (4.6) satisfies the quasi-symmetric condition, by Theorem 2.2.2 the input distribution that maximizes $\frac{1}{n}I(X^n;Y^n)$ is the uniform distribution. With the next proposition, the value of

$[H(Y^n)]$ under uniform distribution is formulated.

Proposition 4.3.2. *The value of* $[H(Y^n)]$ *under a uniform distribution over* $\mathcal{X}^n = \{0,1\}^n$ *is given by*

$$\max_{p(x^n)} H(Y^n) = n + H(W^n). \tag{4.18}$$

Proof. We need to calculate

$$H(Y^n) = -\sum_{y^n \in \ \mathcal{Y}^n} Pr(Y^n = y^n) \log_2 Pr(Y^n = y^n) \tag{4.19}$$

when x^n has a uniform distribution. But,

$$Pr(Y^n = y^n) = \frac{1}{2^n} \sum_{x^n \in \ \mathcal{X}^n} Pr\left(Z^n = \frac{y^n - (2^q - 1)x^n}{(-1)^{x^n}})\right). \tag{4.20}$$

Since $\mathbf{Q}^n$ is quasi-symmetric, the probability in (4.20) is the same for all the 2^n distinct values of y^n. Substituting (4.20) into (4.19) and using Definition 4.3.1, we get

$$\max_{p(x^n)} H(Y^n) = -\sum_{w^n \in \ \mathcal{W}^n} Pr(W^n = w^n) \log_2 \left(\frac{Pr(W^n = w^n)}{2^n}\right) \tag{4.21}$$

and the result follows. □

To find the channel capacity, we just need to substitute (4.18) into (4.10). This gives us

$$C^{(n)} = 1 + \frac{1}{n}[H(W^n) - H(Z^n)] \tag{4.22}$$

and the channel capacity is thus given by

$$\begin{aligned} C_{NFB} &= \lim_{n\to\infty} C^{(n)} \\ &= 1 + \lim_{n\to\infty} \frac{1}{n}[H(W^n) - H(Z^n)] \qquad (4.23) \\ &= 1 + H(\mathcal{W}) - H(\mathcal{Z}) \qquad (4.24) \end{aligned}$$

in bits/channel use, where $H(\mathcal{W}) = \lim_{n\to\infty}(1/n)H(W^n)$ and $H(\mathcal{Z}) = \lim_{n\to\infty}(1/n)H(Z^n)$ denote the entropy rates of $\{W_n\}$ and $\{Z_n\}$, respectively.

Chapter 5

Feedback Capacity of NBNDC

In this chapter, we will show that feedback does not increase the capacity of the NBNDC. Without loss of generality, we assume that $q \geq 2$, since for $q = 1$, the NBNDC reduces to the modulo-2 additive noise channel and hence the result trivially holds from [1].

5.1 Capacity with Feedback

In the derivation of feedback capacity we frequently use the properties of NBNDC that we defined in Chapter 4. Let us recall these properties:

(a) For any fixed input $x \in \mathcal{X}$, $f(x, \cdot) : \mathcal{Z} \to \mathcal{Y}$ is invertible.

(b) Every output symbol is the image of exactly two distinct input-noise pairs; i.e., for any $y \in \mathcal{Y}$, there are exactly two pairs (x_1, z_1) and (x_2, z_2) in $\mathcal{X} \times \mathcal{Z}$ such that $x_1 \neq x_2$, $z_1 \neq z_2$ and $y = f(x_1, z_1) = f(x_2, z_2)$.

In a feedback communication system, by feedback we mean that there exists a channel from the receiver to the transmitter which is noiseless, delayless and has large capacity. Thus at any given time, all previously received outputs are unambiguously known by the transmitter and can be used for encoding the message into the next code symbol. Therefore,a feedback code with blocklength n and rate R consists of a sequence of mappings

$$\psi_i : \{1, 2, ..., 2^{nR}\} \times \mathcal{Y}^{i-1} \rightarrow \mathcal{X}$$

for $i = 1, 2, ...n$ and an associated decoding function

$$\phi : \mathcal{Y}^n \rightarrow \{1, 2, ..., 2^{nR}\}.$$

Thus when the transmitter wants to send a message, say $V \in \{1, 2, ..., 2^{nR}\}$, it sends the codeword X^n, where $X_1 = \psi_1(V)$ and $X_i = \psi_i(V, Y_1, \cdots, Y_{i-1})$, for $i = 2, \cdots, n$. For a received Y^n at the channel output, the receiver uses the decoding function to estimate the transmitted message as $\hat{V} = \phi(Y^n)$. A decoding error is made when $\hat{V} \neq V$.

We assume that the message V is uniformly distributed over $\{1, 2, ..., 2^{nR}\}$. Therefore, the probability of error is given by

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} P\left\{\phi(Y^n) \neq V | V = k\right\} = P\left\{\phi(Y^n) \neq V\right\}.$$

The capacity with feedback, C_{FB}, is the supremum of all admissible feedback code rates (i.e., all rates for which there exists sequences of feedback codes with asymptotically vanishing probability of error). From Fano's inequality, we have

$$\begin{aligned} H(V|Y_n) &\leq h_b({P_e}^{(n)}) + {P_e}^{(n)} \log_2(2^{nR} - 1) \\ &\leq 1 + {P_e}^{(n)} nR \end{aligned}$$

where the second inequality holds since $h_b(P_e^{(n)}) \leq 1$, where $h_b(\cdot)$ is the binary entropy function. We also know that

$$\begin{aligned} nR &= H(V) \\ &= H(V|Y^n) + I(V;Y^n) \\ &\leq 1 + {P_e}^{(n)} nR + I(V;Y^n) \end{aligned}$$

where R is any admissible rate. Dividing both sides above by n and taking the limit yields

$$C_{FB} \leq \lim_{n\to\infty} \sup \frac{1}{n} I(V;Y^n)$$

where the supremum is taken over all feedback policies $\{P(x_i|x^{i-1}, y^{i-1})\}_{i=1}^n$. We can write $I(V;Y^n)$ as follows

$$\begin{aligned} I(V;Y^n) = & \sum_{i=1}^n I(V;Y_i|Y^{i-1}) \\ = & \sum_{i=1}^n \big(H(Y_i|Y^{i-1}) - H(Y_i|V, Y^{i-1})\big) \\ = & \sum_{i=1}^n \big(H(Y_i|Y^{i-1}) - H(Y_i|V, Y^{i-1}, X_i, X^{i-1})\big) \end{aligned}$$

where the last equality follows from the fact that $X_k = \psi_k(V, Y_1, Y_2, \ldots, Y_{k-1})$ for $k = 1, \cdots, i$. We also can write

$$\begin{aligned} & H(Y_i|V, Y^{i-1}, X_i, X^{i-1}) \\ & \quad = H(f(X_i, Z_i)|V, Y^{i-1}, X_i, X^{i-1}) \\ & \quad = H(Z_i|V, Y^{i-1}, X_i, X^{i-1}) \\ & \quad = H(Z_i|V, Y^{i-1}, X_i, X^{i-1}, Z^{i-1}) \\ & \quad = H(Z_i|Z^{i-1}) \end{aligned}$$

where the second and third equalities follow from channel property (a) and the last equality holds since Z_i and (V, X_i, Y^{i-1}) are conditionally independent given Z^{i-1}. Therefore, we get that

$$\begin{aligned} I(V;Y^n) &= \sum_{i=1}^{n} I(V;Y_i|Y^{i-1}) \\ &= \sum_{i=1}^{n} \left[H(Y_i|Y^{i-1}) - H(Z_i|Z^{i-1})\right]. \end{aligned} \tag{5.1}$$

We next prove that all of the output conditional entropies $H(Y^i|Y^{i-1})$ in (5.1) are maximized by uniform conditional input distributions $P(X_i|X^{i-1}, Y^{i-1})$ (feedback policies). With this result in hand, we can then directly deduce that feedback does not increase the capacity of the NBNDC as the right hand side of (5.1) will equal C_{NFB} after normalizing by n and taking the limit.

Lemma 5.1.1. *For a general noise process $\{Z_k\}$, each conditional output entropy $H(Y_i|Y^{i-1})$, $i = 1, \cdots, n$ in (5.1) is maximized by a uniform feedback policy:*

$$P(X_i = a|X^{i-1} = x^{i-1}, Y^{i-1} = y^{i-1}) = \frac{1}{2}$$

for all $a \in \{0,1\}$, $x^{i-1} \in \{0,1\}^{i-1}$ and $y^{i-1} \in \mathcal{Y}^{i-1}$.

Proof. Let us first write the output conditional entropy $H(Y_i|Y^{i-1})$ as

$$H(Y_i|Y^{i-1}) = \sum_{y^{i-1}} P(y^{i-1}) H(Y_i|Y^{i-1} = y^{i-1}) \tag{5.2}$$

where

$$H(Y_i|Y^{i-1} = y^{i-1}) = -\sum_{y_i} P(y_i|y^{i-1}) \log P(y_i|y^{i-1}). \tag{5.3}$$

To show that $H(Y_i|Y^{i-1})$ in (5.2) is maximized by a uniform feedback policy, it is enough to show that such uniform policy maximizes each of the $H(Y_i|Y^{i-1} = y^{i-1})$ terms.

We now expand $P(y_i|y^{i-1})$ as follows

$$\begin{aligned}
&\sum_{x_i}\sum_{x^{i-1}}\sum_{z_i}\sum_{z^{i-1}} P(y_i, x_i, z_i, x^{i-1}, z^{i-1}|y^{i-1}) \\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i, x^{i-1}, z^{i-1}, y^{i-1})P(x_i, z_i, x^{i-1}, z^{i-1}|y^{i-1}) && (5.4)\\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(x_i, z_i, x^{i-1}, z^{i-1}|y^{i-1}) && (5.5)\\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(z_i, x^{i-1}, z^{i-1}|y^{i-1})P(x_i|z_i, x^{i-1}, z^{i-1}y^{i-1}) && (5.6)\\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(x_i|x^{i-1}, y^{i-1})P(z_i, x^{i-1}, z^{i-1}|y^{i-1}) && (5.7)\\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(x_i, x^{i-1}, z^{i-1}|y^{i-1})P(z_i|x_i, x^{i-1}, z^{i-1}, y^{i-1}) && (5.8)\\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(z_i|z^{i-1})P(x_i|x^{i-1}, z^{i-1}, y^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}) && (5.9)\\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(x_i|x^{i-1}, y^{i-1})P(z_i|x^{i-1}, z^{i-1}, y^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(x_i|x^{i-1}, y^{i-1})P(z_i|z^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}). && (5.10)
\end{aligned}$$

Thus

$$\begin{aligned}
P(y_i|y^{i-1}) &= \sum_{x_i}\cdots\sum_{z^{i-1}} P(y_i|x_i, z_i)P(z_i|z^{i-1}) \\
&\quad P(x_i|x^{i-1}, y^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}). && (5.11)
\end{aligned}$$

The equation (5.11) encompasses the properties of channel such as the symmetry.

This can be seen when going through the sum over y_i in (5.3) as follow:

$$\begin{aligned}
P(y_i = 0|y^{i-1}) &= \sum_{\substack{x_i,z_i\\x^{i-1},z^{i-1}}} P(y_i = 0|x_i, z_i)P(z_i|z^{i-1})P(x_i|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \qquad (5.12) \\
&= \sum_{\substack{z_i,x^{i-1}\\z^{i-1}}} P(y_i = 0|x_i = 0, z_i)P(z_i|z^{i-1})P(x_i = 0|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&+ \sum_{\substack{z_i,x^{i-1}\\z^{i-1}}} P(y_i = 0|x_i = 1, z_i)P(z_i|z^{i-1})P(x_i = 1|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \qquad (5.13) \\
&= \sum_{x^{i-1},z^{i-1}} P(y_i = 0|x_i = 0, z_i = 0)P(z_i = 0|z^{i-1})P(x_i = 0|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&+ \sum_{x^{i-1},z^{i-1}} P(y_i = 0|x_i = 1, z_i = 2^q - 1)P(z_i = 2^q - 1|z^{i-1}) \\
&\quad P(x_i = 1|x^{i-1}, y^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}) \qquad (5.14)
\end{aligned}$$

where in (5.14) we used the fact that, $P(y_i|x_i, z_i)$ is deterministic given x_i and z_i and moreover, it is only non-zero for exactly two values of input-noise pairs (channel's properties (a) and (b)).

Similar to the derivation above, we can write $P(y_i = 2^q - 1|y^{i-1})$ as follows;

$$\begin{aligned}
P(y_i = 2^q - 1|y^{i-1}) &= \sum_{\substack{x_i,z_i\\ x^{i-1},z^{i-1}}} P(y_i = 2^q - 1|x_i, z_i)P(z_i|z^{i-1})P(x_i|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \qquad (5.15)\\
&= \sum_{\substack{z_i,x^{i-1}\\ z^{i-1}}} P(y_i = 2^q - 1|x_i = 0, z_i)P(z_i|z^{i-1})P(x_i = 0|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&+ \sum_{\substack{z_i,x^{i-1}\\ z^{i-1}}} P(y_i = 2^q - 1|x_i = 1, z_i)P(z_i|z^{i-1})P(x_i = 1|x^{i-1}, y^{i-1}) \\
&\quad P(x^{i-1}, z^{i-1}|y^{i-1}) \qquad (5.16)\\
&= \sum_{x^{i-1},z^{i-1}} P(y_i = 2^q - 1|x_i = 0, z_i = 2^q - 1)P(z_i = 2^q - 1|z^{i-1}) \\
&\quad P(x_i = 0|x^{i-1}, y^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&+ \sum_{x^{i-1},z^{i-1}} P(y_i = 2^q - 1|x_i = 1, z_i = 0)P(z_i = 0|z^{i-1}) \\
&\quad P(x_i = 1|x^{i-1}, y^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}). \qquad (5.17)
\end{aligned}$$

Equations (5.14) and (5.17) are quite similar. Let us define $P(x_i = 0|x^{i-1}, y^{i-1}) :=$ p and $P(x_i = 1|x^{i-1}, y^{i-1}) := 1 - p$ and look at their sum. Then

$$\begin{aligned}
&P(y_i = 0|y^{i-1}) + P(y_i = 2^q - 1|y^{i-1}) = \\
&\sum_{x^{i-1},z^{i-1}} P(z_i = 0|z^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1})\,(p + (1 - p)) \qquad (5.18)\\
&+ \sum_{x^{i-1},z^{i-1}} P(z_i = 2^q - 1|z^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1})\,(p + (1 - p)) \qquad (5.19)
\end{aligned}$$

where we can observe that, the sum is independent of the feedback policy, $P(x_i = 0|x^{i-1}, y^{i-1}) = p$ over which that we are trying to maximize (5.2).

Considering the channel properties (a) and (b), it can be seen that this argument holds for any $j = 0, 1, \cdots, 2^{q-1} - 1$,

$$\begin{aligned}
&P(Y_i = j|y^{i-1}) + P(Y_i = 2^q - 1 - j|y^{i-1}) \\
&= \sum_{x^{i-1},z^{i-1}} P(Z_i = j|z^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1})\,(p + (1-p)) \\
&\quad + \sum_{x^{i-1},z^{i-1}} P(Z_i = 2^q - 1 - j|z^{i-1})P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&\qquad\qquad \times (p + (1-p)) \\
&= \sum_{x^{i-1},z^{i-1}} \left[P(Z_i = j|z^{i-1}) + P(Z_i = 2^q - 1 - j|z^{i-1})\right] \\
&\qquad\qquad \times P(x^{i-1}, z^{i-1}|y^{i-1}) \\
&= \sum_{z^{i-1}} \left[P(Z_i = j|z^{i-1}) + P(Z_i = 2^q - 1 - j|z^{i-1})\right] \\
&\qquad\qquad \times P(z^{i-1}|y^{i-1}) \\
&:= k_j. \qquad (5.20)
\end{aligned}$$

This fact reduces the problem to the maximization of the following expression

$$H(Y_i|Y^{i-1} = y^{i-1}) = -\sum_{j=0}^{2^{q-1}-1} \left[a_j \log a_j + (k_j - a_j)\log(k_j - a_j)\right] \qquad (5.21)$$

where

$$a_j = P(Y_i = j|Y^{i-1} = y^{i-1})$$

and

$$k_j - a_j = P(Y_i = 2^q - 1 - j|Y^{i-1} = y^{i-1}),$$

applying the log-sum inequality on each summand (within brackets) in (5.21) yields that

$$H(Y_i|Y^{i-1} = y^{i-1}) \le -\sum_{j=0}^{2^{q-1}-1} k_j \log(k_j/2) \qquad (5.22)$$

with equality iff $a_j = k_j - a_j$ for $j = 0, 1, ..., 2^{q-1} - 1$. In other words, $H(Y_i|Y^{i-1} = y^{i-1})$

is maximized iff

$$P(Y_i = j|Y^{i-1}) = P(Y_i = 2^q - 1 - j|Y^{i-1}). \tag{5.23}$$

By examining (5.11) and using the channel's properties, it can be directly shown that (5.23) is satisfied when

$$P(X_i = 0|x^{i-1}, y^{i-1}) = P(X_i = 1|x^{i-1}, y^{i-1}) = \frac{1}{2}. \tag{5.24}$$

Hence a uniform feedback policy maximizes the conditional entropy $H(Y_i|Y^{i-1} = y^{i-1})$ for each y^{i-1}; this completes the proof. □

Lemma 5.1.1 directly implies that a uniform feedback policy yields a uniformly distributed input X^n and maximizes the channel's output block entropy $H(Y^n)$, resulting in $H(Y^n) = n + H(W^n)$ as in (4). Substituting the later in (5.1), normalizing by n and taking the limit yield that

$$C_{FB} \;\; \leq \;\; 1 + H(\mathcal{W}) - H(\mathcal{Z}) = C_{NFB} \tag{5.25}$$

for a stationary ergodic noise. But by definition of the feedback capacity, we know that $C_{NFB} \leq C_{FB}$. Thus, we have shown the following.

Theorem 5.1.1. *Feedback does not increase the capacity of the NBNDC with stationary ergodic noise:*

$$C_{FB} = C_{NFB} = 1 + H(\mathcal{W}) - H(\mathcal{Z}).$$

Observation: We should remark that, since Lemma 5.1.1 holds for arbitrary noise processes, Theorem 5.1.1 can be extended for such noise sources (i.e., without requiring them to be stationary ergodic) by using Verdú and Han's non-feedback capacity formula for general channels with memory [21] as discussed in Chapter (3) [17].

Chapter 6

Summary and Future Work

6.1 Summary

In this project, we first introduced a discrete binary-input 2^q-ary output discrete channel (denoted by NBNDC) to properly represent both the statistical memory and the soft-decision information of BPSK-modulated time-correlated Rayleigh fading channels when they are coherently demodulated via a q-bit output quantizer. We next observed that the NBNDCs output is explicitly described in terms of its binary input and a 2^q-ary noise.

To compute the capacity of this channel, we first observed that the transition probability matrix of the channel is quasi-symmetric and therefore its capacity is achieved by a uniform input. Using this fact, Pimentel and Alajaji computed the channel capacity and showed that it is equal to 1 plus the difference between the entropy of a process with a reduced alphabet and the noise entropy.

In the last chapter we showed that feedback does not increase the capacity of NBNDC. In a sense, it is an unexpected result since one might expect that with feedback there exists some encoding mechanism which makes the output more uniform and increases the capacity.

6.2 Future Work

[1] and this work showed that via the existence of some kind of symmetry in the channel transition matrix, it is not possible to get higher capacity with feedback. The modulo additive channel in [1] is strongly symmetric and the NBNDC is quasi-symmetric. A possible direction for future work is to identify the largest class of channels with memory for which feedback does not increase the capacity. Another extension is to study the feedback capacity of finite-state channels and multiple access channels with memory.

Bibliography

[1] F. Alajaji. Feedback does not increase capacity of discrete channels with additive noise. *IEEE Transaction Information Theory*, 41(1):546–549, January 1995.

[2] F. Alajaji. *Advanced Topics in CommunicationTheory: Information Theory for Systems with Memory.* Queen's University, Lecture Notes, 2003.

[3] F. Alajaji. *Information Theory.* Queen's University, Lecture Notes, 2008.

[4] F. Alajaji and T. Fuja. Effect of feedback on the capacity of discrete additive channels with memory. In *Proceedings of the 1994 IEEE International Symposium on Information Theory*, Trondheim, Norway, 1994.

[5] F. Alajaji and N. Phamdo. Soft-decision COVQ for Rayleigh fading channels. *IEEE Communication Letters*, 2(1):162–164, June 1998.

[6] P.N. Chen and F. Alajaji. Generalized source coding theorems and hypothesis testing: Part I – information measures. *Journal of the Chinese Institute of Engineers*, 21(3):283–292, May 1998.

[7] P.N. Chen and F. Alajaji. Optimistic shannon coding theorems for arbitrary single-user systems. *IEEE Trans. Inform. Theory*, 45(7):2623–2629, November 1999.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory Second Edition.* Wiley, New Jersey, 2006.

[9] T.M Cover and S. Pombra. Gaussian feedback capacity. *IEEE Transaction Information Theory*, 35(1):37–43, January 1989.

[10] P. Ebert. The capacity of Gaussian channel with feedback. *IT Bell System Technical Journal*, pages 1705–1712, Oct. 1970.

[11] A. Feinstein. A new basic theorem of information theory. *IRE Trans. PGIT*, 4:2–22, 1954.

[12] R. G. Galleger. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[13] R.M. Gray and L.D. Davisson. Source coding theorems without the ergodic assumption. *IEEE Transaction Information Theory*, IT-20(4):502–516, 1976.

[14] T. S. Han and K. Kobayashi. *Mathematics of Information and Coding.* American Mathematical Society, Rhode Island, 2002.

[15] C. Pimentel and F. Alajaji. A discrete channel model for capturing memory and soft-decision information: A capacity study. In *Proceedings of IEEE International Conference on Commununication*, Dresden, Germany, 2009.

[16] M. Pinsker. Talk at the Soviet information theory meeting. In *No Abstracts Published*, SSCB, 1969.

[17] N. Sen, F. Alajaji, and S. Yuksel. On the feedback capacity of a discrete non-binary noise channel with memory. In *Proceedings of 11th Canadian Workshop on Information Theory*, Ottawa, Canada, May 2009.

[18] C.E. Shannon. The zero-error capacity of a noisy channel. *IRE Transaction Information Theory*, IT(2):8–19, 1956.

[19] G. Taricco. On the capacity of binary input Gaussian and Rayleigh fading channel. *Eur.Trans.Telecommun.*, 7:201–208, Mar.-Apr. 1996.

[20] S. Verdu and T.H. Han. Approximation theory of output statistics. *IEEE Transaction Information Theory*, 39:752–772, May 1993.

[21] S. Verdu and T.H. Han. A general formula for channel capacity. *IEEE Transaction Information Theory*, 40(4):11471157, July 1994.

[22] S. Verdu and V.H. Poor. A lower bound on the probability of error in multi-hypothesis testing. *IEEE Transaction on Information Theory*, 41(6):1992–1995, Nov. 1995.