THESIS REPORT

Ph.D.

New Results on the Analysis of Discrete Communication Channels with Memory

by F. Alajaji
Advisor: T.E. Fuja

Ph.D. 94-8

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

NEW RESULTS ON THE ANALYSIS OF DISCRETE COMMUNICATION CHANNELS WITH MEMORY

by

Fady Alajaji

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1994

Advisory Committee:

Associate Professor Thomas E. Fuja, Chairman/Advisor
Professor Prakash Narayan
Professor Nariman Farvardin
Associate Professor Steven Tretter
Professor Eric Slud

ABSTRACT

Title of Dissertation: NEW RESULTS ON THE ANALYSIS OF DISCRETE COMMUNICATION CHANNELS WITH MEMORY

Fady Alajaji, Doctor of Philosophy, 1994

Dissertation directed by: Associate Professor Thomas E. Fuja
Department of Electrical Engineering

The reliable transmission of information bearing signals over a communication channel constitutes a fundamental problem in communication theory. An important objective in analyzing this problem is to understand and investigate its "information theoretic" aspects – i.e., to determine the fundamental limits to how efficiently one can encode information and still be able to recover it with negligible loss. In this work, we address this problem for the case where the communication channel is assumed to have memory – i.e., the effect of noise lingers over many transmitted symbols. Our motivation is founded on the fact that most real-world communication channels have memory.

We begin by proposing and analyzing a contagion communication channel. A contagion channel is a system in which noise propagates in a way similar to the spread of an infectious disease through a population; each "unfavorable" event (i.e., an error) increases the probability of future unfavorable events. A contagion-based model offers an interesting and less complex alternative to other models of channels with memory like the Gilbert-Elliott burst channel. We call the model

set forth the Polya-contagion channel – a discrete binary communication channel with additive errors modeled according to the famous urn scheme of George Polya for the spread of contagion.

We next consider discrete channels with arbitrary (not necessarily stationary ergodic) additive noise. Note that such channels need not be memoryless; in general, they have memory. We show that output feedback does not increase the capacity of such channels. The same result is also shown for a larger class of channels to which additive channels belong, the class of discrete symmetric channels with memory. These channels have the property that their inf-information rate is maximized for equally likely iid input processes.

Finally, we impose average cost constraints on the input of the additive channels, rendering them non-symmetric. We demonstrate that in the case where the additive noise is a binary stationary mixing Markov process, output feedback can increase the capacity-cost function of these channels.

Dedication

To Naman Alajaji

Acknowledgements

I would like to thank my advisor, Professor Thomas Fuja, for his guidance, continuous encouragement and support. My time with him has been a fruitful learning experience. I would like to also thank Professor Imre Csiszàr from the Mathematical Institute of the Hungarian Academy of Sciences, Professor Prakash Narayan, Professor Nariman Farvardin and Professor Steven Tretter. I have profited greatly from many discussions with them. I am grateful to my friends in the Communications and Signal Processing Laboratory for their assistance and numerous interesting discussions. I also acknowledge the generous Fellowships offered to me during the course of my study by the Institute of Systems Research and the Graduate School at the University of Maryland. Finally, I would like to thank all my family and friends for their constant love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Transmission of information from one point to another is at the heart of what we call communication. As an area of concern, communication theory is so vast as to touch the preoccupations of philosophers as well as scientists, and to give rise to a thriving technology.

A communication system consists mainly of three parts: (1) The source, which generates messages at the transmitting end of the system, (2) The destination, which tries to reproduce the messages as accurately as possible, and (3) The channel, consisting of a noisy (in general) transmission medium or device for conveying signals from the source to the destination. The source and destination ends can be separated via the channel link in two ways: separation in space (as in satellite transmission) or separation in time (as in storage). In this dissertation, we will focus our attention on the study of communication channels.

Associated with most communication channels is a non-negative number called the *capacity* which, loosely interpreted, is the maximum amount of information (in bits/channel-use) that can be *reliably* transmitted across this channel. In other words, if C is the capacity of a given channel, then there exist methods of encoding

data for transmission over the channel at rates arbitrarily close to C bits/channel-use – and these methods provide arbitrarily good performance, provided a sufficiently complex encoder/decoder is permitted. Conversely, it is impossible to reliably convey data at rates above the capacity.

The simplest kind of channel studied is the *discrete* (discrete time, discrete alphabet) *memoryless channel* (DMC). A DMC is a channel for which the output letter at a given time depends statistically only on the corresponding input letter. An example of a DMC is a channel for which the output sequence is obtained by adding the channel input sequence to a noise sequence that consists of a series of *iid* random variables.

A discrete channel is said to have *memory*, if each letter in the output sequence depends statistically on the corresponding input letter, as well as on the past inputs, past outputs and future inputs. If the output letter does not depend on the past input letters, then the channel is said to be *historyless* (or with no input memory); thus the channel has anticipation and output memory. The channel is *causal* (or has no anticipation) if for a given input and a given input-output history, the current output letter is statistically independent of future inputs. If the channel is causal and historyless, then it possibly has output memory only. An example of a channel with output memory, would be the additive noise channel described above, with the exception that the noise sequence is a series of *dependent* random variables. In this dissertation, new results are obtained on the analysis of discrete channels with output memory.

1.1 A Communication Channel Modeled on Contagion

In the first part of the dissertation, we propose and analyze a *contagion* communication channel. A contagion channel is a system in which noise propagates in a way similar to the spread of an infectious disease through a population. The noise process of the channel is generated according to the contagion model of George Polya; our motivation is the empirical observation of Stapper *et. al.* that defects in semiconductor memories are well described by distributions derived from Polya's urn scheme. The resulting channel is stationary but not ergodic, and it has many interesting properties.

We first derive a maximum likelihood (ML) decoding algorithm for the channel; it turns out that ML decoding is equivalent to decoding a received vector onto *either* the closest codeword *or* the codeword that is farthest away, depending on whether an "apparent epidemic" has occurred. We next show that the Polya-contagion channel is an "averaged" channel in the sense of Ahlswede (and others) and that its capacity is zero. Finally, we consider a finite-memory version of the Polya-contagion model; this channel is (unlike the original) ergodic with a non-zero capacity that increases with increasing memory.

1.2 Feedback Capacity of Discrete Additive Channels

In the second part of the dissertation, we consider discrete channels with additive random noise. Note that such channels need not be memoryless; in general,

they have memory. The Polya-contagion channel discussed in the previous section belongs to the class of such channels.

We investigate the effect of output feedback on the capacity of additive noise channels with memory. By output feedback, we mean the existence of a "return channel" from the receiver to the transmitter; we assume this return channel is noiseless, delayless, and has large capacity. The receiver uses the return channel to inform the transmitter what letters were actually received; these letters are received at the transmitter before the next letter is transmitted, and therefore can be used in choosing the next transmitted letter.

Intuitively, it is plausible that if we use feedback on channels with memory, then we can use some encoding techniques at the transmitter end in order to combat the noise of the channel and hence increase the capacity of the channel. However we reach the seemingly surprising result that the capacity of additive noise channels with feedback does not exceed their respective capacity without feedback. This is shown for both ergodic and non-ergodic additive stationary noise. In light of recent results on channel capacity by Verdú and Han [35], we then generalize our result for discrete channels with *arbitrary* additive noise. This result extends Shannon's work of more than 35 years ago [30], which showed the same result for *memoryless* channels.

Finally, we introduce the notion of *symmetric* channels with memory to which additive noise channels belong. These channels are obtained by combining an input process with an arbitrary noise process that is independent of the input. They have the property that equiprobable input vectors maximize their block mutual information. We show that feedback does not increase the capacity of *symmetric* channels.

1.3 Feedback Capacity-Cost Function of Discrete Additive Channels

In the last part of the dissertation, we analyze the capacity-cost function of additive noise channels. We consider a binary additive noise channel, where the noise process is a stationary mixing Markov process of order M. This channel is *symmetric* as observed in the previous section. We incorporate average cost constraints on the possible input sequences of the additive channel, thus "destroying" its symmetry. We establish a tight upper bound to the capacity-cost function of the channel with no feedback ($C_{NFB}(\beta)$) and a lower bound to the capacity-cost function with feedback ($C_{FB}(\beta)$). Numerical results indicate that output feedback *can increase* the capacity-cost function of this channel. An upper bound to $C_{FB}(\beta)$ is also obtained, by proving the converse to the coding theorem.

1.4 Outline of the Dissertation

The rest of this dissertation is organized as follows. In Chapter 2, we present the communication channel modeled by the spread of disease. We then look at the effect of output feedback on the capacity of discrete channels with additive noise in Chapter 3. The analysis of the capacity-cost function of discrete additive Markov channels with feedback is addressed in Chapter 4. Finally, conclusions are stated in Chapter 5.

Chapter 2

A Communication Channel Modeled by the Spread of Disease

2.1 Introduction and Motivation: Communication via Contagion

We consider a discrete communication channel with memory in which errors spread in a fashion similar to the spread of a contagious disease through a population. The errors propagate through the channel in such a way that the occurrence of each "unfavorable" event (i.e., an error) increases the probability of future unfavorable events.

One motivation for the study of such channels is the "clustering" of defects in silicon; Stapper *et. al.* [32] have shown that the distribution of defects in semiconductor memories fits the Polya-Eggenberger (PE) distribution much better than the commonly used Poisson distribution. The PE distribution is one of the "contagious" distributions that can be generated by George Polya's urn model for the spread of contagion [26, 27]. More generally, real-world communication channels –

in particular the digital cellular channel – often have memory; a contagion-based model offers an interesting alternative to the Gilbert model and others [20].

We begin by introducing a communication channel with additive noise modeled by the Polya contagion urn scheme; the channel is stationary but not ergodic. We then present a maximum likelihood (ML) decoding algorithm for the channel; ML decoding for the Polya-contagion channel is carried out by mapping the received vector onto either the codeword that is closest to the received vector *or* the codeword that is farthest away – depending on which possibility is more *extreme.* We then show that the Polya-contagion channel is in fact an "averaged" channel [1, 19] – i.e., its block transition probability is the average of those of a class of binary symmetric channels, where the expectation is taken with respect to the beta distribution. Using De Finetti's results on exchangeability, we note that binary channels with additive exchangeable noise processes are averaged channels with binary symmetric channels as components.

We show that the capacity of the Polya channel is zero, and we also obtain the ϵ-capacity of the channel. The zero capacity result provides a counter-example to the adage "memory can only increase capacity". We note that this adage applies only to causal historyless *information stable* channels [10, 25], and that for more general channels, memory may *increase or decrease* capacity.

Finally, we consider a finite-memory version of the Polya-contagion model. The resulting channel is a stationary ergodic Markov channel with memory M; its capacity is positive and increases with M. As M grows, the n-fold conditional distribution of the finite-memory channel converges to the n-fold conditional distribution of the original Polya channel; however the capacity of the finite-memory channel *does not converge* to the capacity of the Polya channel.

2.2 Polya-Contagion Communication Channel

Consider a discrete binary additive communication channel – i.e., a channel for which the i^{th} output $Y_i \in \{0,1\}$ is the modulo-two sum of the i^{th} input $X_i \in \{0,1\}$ and the i^{th} noise symbol $Z_i \in \{0,1\}$; more succinctly, $Y_i = X_i \oplus Z_i$, for $i = 1, 2, 3, \ldots$

We assume that the input and noise sequences are independent of each other. The noise sequence $\{Z_i\}_{i=1}^{\infty}$ is drawn according to the Polya contagion urn scheme [28], as follows: An urn originally contains T balls, of which R are red and S are black ($T = R + S$); let $\rho = R/T$ and $\sigma = 1 - \rho = S/T$. We make successive draws from the urn; after each draw, we return to the urn $1 + \Delta$ balls of the same color as was just drawn. Note that if $\Delta = 0$, we get the classic case of independent drawings with replacement. In our problem we will assume that $\Delta > 0$ (contagion case) and that $\rho < \sigma$ – i.e. $\rho < 1/2$. Furthermore, we denote $\delta = \Delta/T$. Our sequence $\{Z_i\}$ corresponds to the outcomes of the draws from our Polya urn with parameters ρ and δ, where:

$$Z_i = \begin{cases} 1, & \text{if the } i^{th} \text{ ball drawn is red;} \\ 0, & \text{if the } i^{th} \text{ ball drawn is black.} \end{cases}$$

In Polya's model, a red ball in the urn represents a sick person in the population and a black ball in the urn represents a healthy person.

2.2.1 Block Transition Probability of the Channel

Definition 1 (Channel state) We define the state of the channel after the n^{th} transmission to be the total number of red balls drawn after n trials:

$$S_n \triangleq Z_1 + Z_2 + \cdots + Z_n = S_{n-1} + Z_n, \qquad S_0 = 0.$$

The possible values of S_n are the elements of the set $\{0, 1, \ldots, n\}$. Furthermore, the sequence of states $\{S_n\}_{n=1}^{\infty}$ form a Markov chain, i.e.

$$P(S_n = s_n \mid S_{n-1} = s_{n-1}, S_{n-2} = s_{n-2}, \ldots, S_1 = s_1) = P(S_n = s_n \mid S_{n-1} = s_{n-1}).$$

For a given input block $\underline{X} = [X_1, X_2, \ldots, X_n]$ and a given output block $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]$, the block (or n-fold) transition probability of the channel is given by

$$P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) = \prod_{i=1}^{n} P(Y_i = y_i \mid X_i = x_i, S_{i-1} = s_{i-1}),$$

where

$$P(Y_i = y_i \mid X_i = x_i, S_{i-1} = s_{i-1}) = \begin{cases} \left[\frac{\rho + s_{i-1}\delta}{1+(i-1)\delta}\right], & \text{if } y_i \oplus x_i = 1; \\ \left[\frac{\sigma + (i-1-s_{i-1})\delta}{1+(i-1)\delta}\right], & \text{if } y_i \oplus x_i = 0. \end{cases}$$

We thus obtain:

$$P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) = \frac{\rho(\rho+\delta)\cdots(\rho+(d-1)\delta)\sigma(\sigma+\delta)\cdots(\sigma+(\tilde{d}-1)\delta)}{(1+\delta)(1+2\delta)\cdots(1+(n-1)\delta)}, \quad (2.1)$$

or

$$P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) = \frac{\Gamma(\frac{1}{\delta})\ \Gamma(\frac{\rho}{\delta}+d)\ \Gamma(\frac{\sigma}{\delta}+n-d)}{\Gamma(\frac{\rho}{\delta})\ \Gamma(\frac{\sigma}{\delta})\ \Gamma(\frac{1}{\delta}+n)}, \quad (2.2)$$

where $\tilde{d} \triangleq n - d$, $d = d(\underline{y}, \underline{x}) = weight(\underline{z} = \underline{y} \oplus \underline{x}) = s_n$ and $\Gamma(\cdot)$ is the gamma function, $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$ for $x > 0$. To obtain equation (2.2) from equation (2.1), we used the fact that $\Gamma(x+1) = x\,\Gamma(x)$ which leads to the following identity:

$$\prod_{j=0}^{n-1}(\alpha + j\beta) = \beta^n \frac{\Gamma(\frac{\alpha}{\beta}+n)}{\Gamma(\frac{\alpha}{\beta})}.$$

2.2.2 Properties of the Channel

We first define a discrete channel to be stationary if for every *stationary* input process $\{X_i\}_{i=1}^{\infty}$, the joint input-output process $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is stationary. Furthermore, a discrete channel is ergodic if for every *ergodic* input process $\{X_i\}_{i=1}^{\infty}$, the joint input-output process $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is ergodic [21, 16].

Before analyzing the characteristics of the channel, we state from [33] the following definitions and lemma.

Definition 2 A finite sequence of random variables $\{Z_1, Z_2, \cdots, Z_n\}$ is said to be *exchangeable* if the joint distribution of $\{Z_1, Z_2, \cdots, Z_n\}$ is invariant with respect to permutations of the indices $1, 2, \cdots, n$.

Definition 3 An infinite sequence of random variables $\{Z_i\}_{i=1}^{\infty}$ is said to be *exchangeable* if for every finite n, the collection $\{Z_{i_1}, Z_{i_2}, \cdots, Z_{i_n}\}$ is exchangeable.

Lemma 1 Exchangeable random processes are *strictly stationary.*

Exchangeability was investigated by De Finetti (1931) who recognized its fundamental role for Bayesian statistics and modern probability. The main interest in adopting this concept is to use exchangeable random variables as an alternative to independent identically distributed (*iid*) random variables. Note that *iid* random variables are exchangeable. However, exchangeable random variables are *dependent* in general but symmetric in their dependence. We now can study the properties of the channel:

1. **Symmetry:** The channel is *symmetric.* By this we mean that $P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$ depends only on $\underline{x} \oplus \underline{y}$ since $P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) = P(\underline{Z} = \underline{y} \oplus \underline{x})$. Due to the symmetry, if we want to maximize the mutual information $I(\underline{X}; \underline{Y})$ over all input distributions on $\underline{X}$, the result is maximized for equiprobable input n-tuples (uniform *iid* input process). The resulting output process is also uniform iid.

2. **Stationarity:** From equation (2.1) and the above definitions, we can conclude that the noise process $\{Z_i\}_{i=1}^{\infty}$ forms an *exchangeable* random process.

The noise process is thus *strictly stationary* (by Lemma 1) and thus *identically distributed.* We get:

$$P(Z_i = 1) = \rho = 1 - P(Z_i = 0) \quad \forall\, i = 1, 2, 3, \ldots$$

and the correlation coefficient

$$Cor(Z_i, Z_j) = \frac{Cov(Z_i, Z_j)}{\sqrt{Var(Z_i)\ Var(Z_j)}} = \frac{\delta}{1+\delta} > 0 \quad \forall\, i \neq j$$

indicates the positive correlation among the random variables of the noise process.

3. **Non-Ergodicity :** It is shown in [28, 11] that $Z \triangleq \lim_{n\to\infty} S_n/n$ exists almost surely, where Z has the beta distribution with parameters ρ/δ and σ/δ. Thus the noise process $\{Z_i\}_{i=1}^{\infty}$ is *not ergodic* since its sample average does not converge to a constant.

2.3 Maximum Likelihood (ML) Decoding

Suppose M codewords are possible inputs to the channel with transition probability $P(\underline{Y} = \underline{y}|\underline{X} = \underline{x})$; the codebook is given by $\mathcal{C} = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_M\}$, with each $\underline{x}_k \in \{0,1\}^n$. For a given received vector $\underline{y} \in \{0,1\}^n$ the maximum likelihood estimate of the transmitted codeword is

$$\underline{x} = \arg\max\{P(\underline{Y} = \underline{y}|\underline{X} = \underline{x}_k) : \underline{x}_k \in \mathcal{C}\}.$$

From equation (2.2), we can rewrite the transition probability of the channel as:

$$P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) = g(d(\underline{x}, \underline{y})),$$

where $g : [0, n] \to [0, 1]$ is defined by

$$g(d) = A \cdot \Gamma\left(\frac{\rho}{\delta} + d\right) \cdot \Gamma\left(\frac{\sigma}{\delta} + n - d\right),$$

and A is a constant depending on n, ρ, and δ.

In order to analyze the behavior of $P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$, we refer to the following definition and lemmas without stating their proofs from [7]:

Definition 4 Let f be a real-valued function defined and strictly positive on an interval $\mathbf{I} \subset \mathbf{R}$. If

$$f[ux + (1-u)y] \leq [f(x)]^u [f(y)]^{1-u}$$

for all x, y $\in \mathbf{I}$ and all $u \in [0, 1]$, then f is said to be *logarithmically convex* or briefly *log-convex* on $\mathbf{I}$. If the inequality is strict, then f is said to be strictly log-convex.

Another definition of log-convexity is the following: If $f(x) > 0$ for all x $\in \mathbf{I}$ and if *log* f is convex on $\mathbf{I}$, then f is said to be log-convex on $\mathbf{I}$. Furthermore, by the inequality of the arithmetic and geometric means, we have:

$$[f(x)]^u [f(y)]^{1-u} \leq u\, f(x) + (1-u)\, f(y).$$

Thus log-convexity implies convexity, but not conversely.

Lemma 2 The set of all log-convex functions is closed under both addition and multiplication- i.e., if f_1 and f_2 are log-convex, then so are $f_1 + f_2$ and $f_1.f_2$.

Lemma 3 The gamma function is strictly log-convex on $\mathbf{R}_+$.

We thus obtain that $g(\cdot)$ defined above is strictly log-convex on the interval $[0, n]$. This observation leads to the following result.

Proposition 1 The transition probability function $P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$ of the Polya-contagion channel is strictly log-convex in $d(\underline{x}, \underline{y})$ and has a unique minimum at

$$d_0 = \frac{n}{2} + \frac{1 - 2\rho}{2\delta}.$$

Furthermore, $P(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$ is symmetric in $d(\underline{x}, \underline{y})$ about d_0.

Proof 1 As above, define $g(d) = P(\underline{Y} = \underline{y} | \underline{X} = \underline{x})$ for any $\underline{x}, \underline{y}$ such that $d(\underline{x}, \underline{y}) = d$; then $g(\cdot)$ is strictly log-convex. For $d_0 = (n/2) + ((1 - 2\rho)/2\delta)$, we obtain

$$g(d_0 + \epsilon) = g(d_0 - \epsilon) = A\, \Gamma\left(\frac{n}{2} + \frac{1}{2\delta} + \epsilon\right)\, \Gamma\left(\frac{n}{2} + \frac{1}{2\delta} - \epsilon\right)$$

for any ϵ; therefore $g(\cdot)$ is symmetric about d_0 and the strict convexity of $g(\cdot)$ means that a unique minimum occurs there. ■

Decoding Algorithm: From the results above, the ML decoding algorithm for the channel is as follows:

1. For a given n-tuple $\underline{y}$ received at the channel output, compute $d_i \stackrel{\triangle}{=} d(\underline{y}, \underline{x}_i)$, for $i = 1, \ldots, M$. Compute also $d_{max} \stackrel{\triangle}{=} \max_{1 \le i \le M}\{d_i\}$ and $d_{min} \stackrel{\triangle}{=} \min_{1 \le i \le M}\{d_i\}$.

2. If $|d_{max} - d_0| \le |d_{min} - d_0|$, map $\underline{y}$ onto a codeword $\underline{x}_j$ for which $d_j = d_{min}$. In this case ML decoding $\Longleftrightarrow$ minimum distance decoding.

3. If $|d_{max} - d_0| > |d_{min} - d_0|$, map $\underline{y}$ onto a codeword $\underline{x}_j$ for which $d_j = d_{max}$. In this case ML decoding $\Longleftrightarrow$ maximum distance decoding.

In Figure (2.1), we have that

$$a = g(0) = \frac{\Gamma(\frac{1}{\delta})\, \Gamma(\frac{\sigma}{\delta} + n)}{\Gamma(\frac{\sigma}{\delta})\, \Gamma(\frac{1}{\delta} + n)}, \qquad \text{and} \qquad b = g(n) = \frac{\Gamma(\frac{1}{\delta})\, \Gamma(\frac{\rho}{\delta} + n)}{\Gamma(\frac{\rho}{\delta})\, \Gamma(\frac{1}{\delta} + n)}.$$

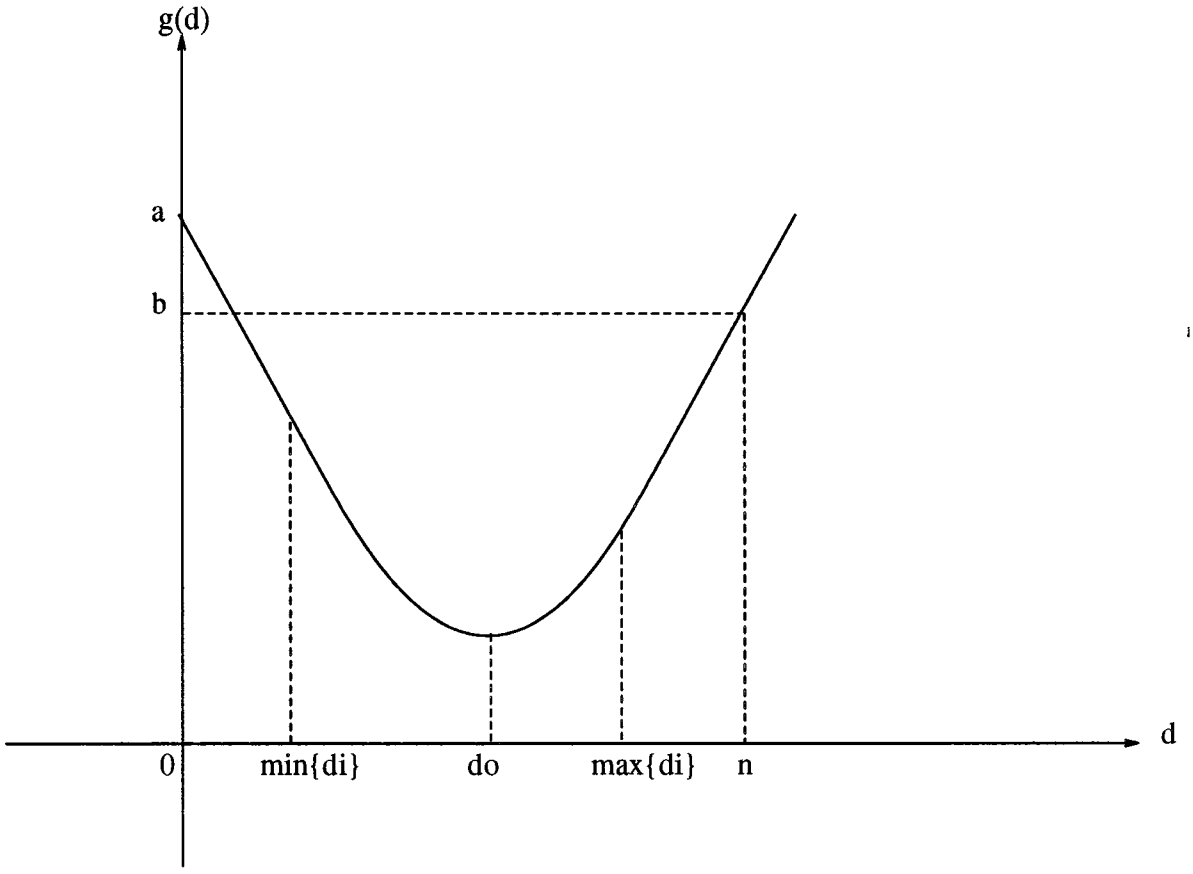

Figure 2.1: Transition probability function vs Hamming distance.

Observations:

- Insight into the decoding rule:
  - We can rewrite d_0 as:

$$d_0 = \frac{n}{2} + \frac{1}{\Delta}\left(\frac{T}{2} - R\right).$$

Note that $n/2$ is (of course) the distance the received n-tuple would be from the transmitted codeword if *half* of the bits get flipped; note also that $(T/2 - R)$ is the initial offset from having an equal number of red and black balls in the urn. Thus d_0 may be thought of as an equilibrium

point.

- The best estimate is then specified by the value of d_i that is furthest away from the equilibrium point d_0. In other words, the best decision is based on the following reasoning: either *many* errors occurred during transmission – an *apparent epidemic,* to use the contagion interpretation – or very *few* errors occurred – an apparently healthy population.

- We note that, if $d_0 > n - 0.5$, then condition (2) in the above algorithm is always satisfied – meaning minimum distance decoding is optimal. The requirement $d_0 > n - 0.5$ is equivalent to the condition

$$\delta < \frac{1 - 2\rho}{n - 1},$$

so if the parameter $\delta = \Delta/T$ is sufficiently small – i.e., there is sufficiently little memory in the system – minimum distance decoding is optimal. In particular, if $\delta = 0$, the draws from the urn are independent and the channel reduces to a binary symmetric channel with crossover probability ρ. Thus this observation is consistent with the fact that, for a BSC with crossover probability less than one-half, minimum-distance decoding is maximum likelihood decoding.

2.4 Averaged Communication Channels

Averaged channels with discrete memoryless components were first introduced by Jacobs [19] and then were analyzed by Ahlswede [1] and Kieffer [21] who investigated their operational capacity. We will show that the Polya-contagion channel is an averaged channel with components that are binary symmetric channels (BSC's).

Consider a family of stationary channels parameterized by θ:

$$\left\{W_\theta^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}), \theta \in \Theta\right\}_{n=1}^{\infty},$$

where $\underline{Y}$ and $\underline{X}$ are respectively the input and output blocks of the channel, each of length n. $W_\theta^{(n)}(\cdot)$ is the n-fold transition probability of the channel specified by $\theta \in \Theta$.

Definition 5 We say a channel is an "averaged" channel with stationary ergodic components if its block transition probability is the expected value of the transition probabilities of a class of stationary channels parameterized by θ – i.e., if it's of the form:

$$\begin{aligned} W_{\text{avg}}^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) &= \int_\Theta W_\theta^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})\, dG(\theta) \\ &= E_\theta[W_\theta^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})], \end{aligned} \tag{2.3}$$

for some distribution $G(\cdot)$ on θ.

Note that if a channel is averaged with stationary components then it is stationary, and it may have memory. One way an averaged channel may be realized is as follows: From among the components, nature selects one according to some probability distribution G. This component is then used for the entire transmission. However the selection is unknown to both the encoder and the decoder.

We will show that the Polya channel – and indeed *any* additive channel – belongs to this class of channels. But first we need to recall some results from [13, 14, 15]:

Notation: Consider a discrete time random process with alphabet D; let $\sigma(D^\infty)$ denote a σ-field consisting of subsets of D^∞, and let μ be a probability measure such that $(D^\infty, \sigma(D^\infty), \mu)$ forms a probability space. Finally, let $\mathbf{U}_n$:

$D^\infty \to D$ denote a sampling function defined by $\mathbf{U}_n(u) = u_n$. Then the sequence of random variables $\{\mathbf{U}_n; n = 1, 2, \ldots\}$ is a discrete time random process, to be denoted $[D, \mu, \mathbf{U}]$.

Lemma 4 (Ergodic Decomposition) Let $[D, \mu, \mathbf{U}]$ be a stationary, discrete time random process. There exists a class of stationary ergodic measures $\{\mu_\theta; \theta \in \Theta\}$ and a probability measure G on an event space of Θ such that for every event $F \subset \sigma(D^\infty)$ we can write:

$$\mu(F) = \int_\Theta \mu_\theta(F)\, dG(\theta).$$

Remark: The ergodic decomposition theorem states that, in an appropriate sense, all stationary non-ergodic random processes are a mixture of stationary ergodic processes; that is if we are viewing a stationary non-ergodic process, we are in reality viewing a stationary ergodic process selected by nature according to some probability measure G. By directly applying the ergodic decomposition theorem we get the following result.

Proposition 2 Any discrete channel with stationary (non-ergodic) additive noise is an averaged channel whose components are channels with additive stationary ergodic noise.

Proof 2 Let $\{Z_i\}$ be the (non-ergodic) noise sequence. Then the ergodic decomposition theorem states that $P(\underline{Z} = \underline{z}) = P(Z_1 = z_1, \ldots, Z_n = z_n)$ may be written as the expected value of the distribution of a class of stationary ergodic processes; since the noise and input sequences are independent, we have $W^{(n)}(\underline{Y} = \underline{y}|\underline{X} = \underline{x}) = P(\underline{Z} = \underline{y} - \underline{x})$ and so $W^{(n)}(\underline{Y} = \underline{y}|\underline{X} = \underline{x})$ may likewise be expressed as the expected value of the transition probabilities of a class of channels with stationary ergodic additive noise. ■

Proposition 3 The binary Polya-contagion channel is an averaged channel; its components are BSC's with crossover probability θ, where θ is a beta-distributed random variable with parameters ρ/δ and σ/δ.

Proof 3 We showed in Proposition 2 that the Polya channel is an averaged channel whose components are channels with additive stationary ergodic noise. To prove the rest of the proposition we just note that, if we let $f_\Theta(\theta)$ be the pdf of a beta-distributed(ρ/δ,σ/δ) random variable – i.e.,

$$f_\Theta(\theta) = \begin{cases} \dfrac{\Gamma(1/\delta)}{\Gamma(\rho/\delta)\Gamma(\sigma/\delta)}\theta^{\rho/\delta-1}(1-\theta)^{\sigma/\delta-1}, & \text{if } 0<\theta<1; \\ 0, & \text{otherwise,} \end{cases}$$

then

$$\int_0^1 \theta^{d(\underline{x},\underline{y})}(1-\theta)^{n-d(\underline{x},\underline{y})} f_\Theta(\theta)d\theta = P(\underline{Y}=\underline{y}|\underline{X}=\underline{x}),$$

where $P(\underline{Y}=\underline{y}|\underline{X}=\underline{x})$ describes the Polya-contagion channel as in (2.2). ■

Observation: We could have proved part of Proposition 3 by using De Finetti's results on exchangeability, since the additive noise process of the Polya channel is a binary exchangeable random process. De Finetti's results are summarized below [11, 34].

Theorem 1 (De Finetti) For an infinite sequence of random variables, the concept of exchangeability is equivalent to that of conditional independence with a common marginal distribution; i.e. if $Z_1, Z_2, \ldots$ is an infinite sequence of exchangeable random variables, then there exists a σ-field $\mathcal{F}$ and a distribution G such that, given $\mathcal{F}$, the random variables $Z_1, Z_2, \ldots$ are conditionally independent with distribution function G.

Corollary 1 For every infinite sequence of exchangeable random variables $\{Z_i\}$ such that $Z_i \in \{0,1\}$, there corresponds a probability distribution G concentrated on the interval $(0,1)$ such that:

$$P(Z_1 = e_1, Z_2 = e_2, \ldots, Z_n = e_n) = \int_0^1 \theta^k \ (1-\theta)^{n-k} \ dG(\theta),$$

where $k = e_1 + e_2 + \cdots + e_n$ and $e_i \in \{0,1\}$ for $i = 1, 2, \ldots, n$.

This brings us to the following more general result:

Proposition 4 Any binary channel with an exchangeable additive noise process is an averaged channel with binary symmetric channels (BSC's) as its components.

2.5 Capacity of the Polya Channel

Consider a discrete (not necessarily memoryless) channel with input alphabet A and output alphabet B; let $W^{(n)}(\underline{Y} = \underline{y} | \underline{X} = \underline{x})$ be the n-fold transition probability describing the channel.

Definition 6 An (M, n, ϵ) code has M codewords, each with blocklength n, and average error probability not larger than ϵ. $R \geq 0$ is an *ϵ-achievable rate* if for every $\gamma > 0$ there exists, for sufficiently large n, (M, n, ϵ) codes with rate

$$\frac{log_2(M)}{n} > R - \gamma.$$

The maximum ϵ-achievable rate is called the *ϵ-capacity*, C_ϵ. The *channel capacity*, C, is the maximum rate that is ϵ-achievable for all $0 < \epsilon < 1$. It follows immediately from the definition that

$$C = \lim_{\epsilon \to 0} C_\epsilon.$$

In [35], Verdú and Han derived a formula for the capacity of arbitrary single-users channels (not necessarily stationary, ergodic, information stable, etc.):

Lemma 5 The channel capacity C is given by

$$C = \sup_{\underline{X}} \underline{I}\,(\underline{X};\underline{Y}), \tag{2.4}$$

where the symbol $\underline{I}\,(\underline{X};\underline{Y})$ is the *inf-information rate* between $\underline{X}$ and $\underline{Y}$ and is defined as the *liminf in probability*[1] of the sequence of normalized information densities $\frac{1}{n}\, i_{\underline{X};\underline{Y}}(\underline{X};\underline{Y})$, where

$$i_{\underline{X};\underline{Y}}(\underline{a};\underline{b}) = \log_2 \frac{P^{(n)}_{\underline{Y}|\underline{X}}(\underline{b}|\underline{a})}{P^{(n)}_{\underline{Y}}(\underline{b})}. \tag{2.5}$$

Using the above lemma as well as the properties of the inf-information rate derived in [35], we obtain that that the inf–information rate in (2.4) is maximized when the input process is equally likely Bernoulli (symmetry property), yielding the following expression for the capacity of the Polya channel

$$C_{\text{Polya}} = 1 - \overline{H}(\underline{Z}),$$

where $\overline{H}(\underline{Z})$ is the sup-entropy rate of the additive Polya noise process $\{Z_n\}$, defined as the limsup in probability of $\frac{1}{n}\, \log_2 \frac{1}{P^{(n)}_{\underline{Z}}(\underline{Z})}$. Since the noise process is stationary, we obtain that the sup-entropy rate is equal to the supremum over the the entropies of almost every ergodic component of the noise process [21, 35]:

$$C_{\text{Polya}} = 1 - \text{ess}_{\Theta} \sup\ h(W_\theta), \tag{2.6}$$

[1] If A_n is a sequence of random variables, then its *liminf in probability* is the supremum of all reals α for which $P(A_n \leq \alpha) \longrightarrow 0$ as $n \longrightarrow \infty$. Similarly, its *limsup in probability* is the infimum of all reals β for which $P(A_n \geq \beta) \longrightarrow 0$ as $n \longrightarrow \infty$ [35].

where

- the noise entropy rate $h(W_\theta)$ is given by

$$h(W_\theta) = -\lim_{n\to\infty} \frac{1}{n} \sum_{\underline{x},\underline{y}\in A^n} W_\theta^{(n)}(\underline{y} \mid \underline{x})\, Q^{(n)}(\underline{x})\, \log_2\, W_\theta^{(n)}(\underline{y} \mid \underline{x}),$$

- and the essential supremum is defined by

$$\text{ess}_\Theta \sup\ f(\theta) \stackrel{\triangle}{=} \inf\ [r : dG(f(\theta) \leq r) = 1].$$

We know that the stationary ergodic components of the Polya channel are BSC's with crossover probability θ; therefore the noise entropy rate is given by $H(W_\theta) = h_b(\theta)$, where $h_b(x) = -x\log_2(x) - (1-x)\log_2(1-x)$. Equation (2.6) then yields the capacity of the channel:

$$C_{\text{Polya}} = 1 - \text{ess}_\Theta \sup\ h_b(\theta).$$

Since θ has the beta distribution on [0,1], we obtain $\text{ess}_\Theta \sup\ h_b(\theta) = 1$ which and so $C_{\text{Polya}} = 0$.

ϵ-Capacity of the Polya Channel: Since the stationary Polya noise process $\{Z_n\}$ is a mixture of Bernoulli(θ) processes where the parameter θ is beta distributed with parameters ρ/δ and σ/δ, it can be shown using the ergodic decomposition theorem [14, 16] that $\frac{1}{n}\ \log_2\ \frac{1}{P_{\underline{Z}}^{(n)}(\underline{Z})}$ converges in L^1 and almost surely (hence in distribution) to the random variable $V \stackrel{\triangle}{=} h_b(U)$ where U is beta-distributed(ρ/δ,σ/δ). The cumulative distribution function (cdf) of V is given by

$$F_V(a) \ \stackrel{\triangle}{=}\ P(V \leq a) = F_U(h_b^{-1}(a)) + 1 - F_U(1 - h_b^{-1}(a)), \qquad (2.7)$$

where $h_b^{-1}(a) \in [0, 1/2]$ is the smallest root of the equation $a = h_b(u)$, and $F_U(\cdot)$ is the cdf of U. Note that since U is beta distributed, $F_V(\cdot)$ is strictly increasing in the interval $[0, 1]$; it therefore admits an inverse $F_V^{-1}(\cdot)$. Now, applying the formula for ϵ-capacity in Theorem 6 in [35], we obtain

$$C_\epsilon = 1 - F_V^{-1}(1 - \epsilon). \tag{2.8}$$

Note that $\lim_{\epsilon \to 0} C_\epsilon = 1 - F_V^{-1}(1) = 1 - 1 = 0$, as expected.

Observations:

- The zero capacity of the Polya channel is due to the fact that θ can occur in any neighborhood of the point $1/2$ with positive probability. This channel behaves like a compound channel with BSC's as components and the capacity of such a compound channel is equal to the infimum of the capacities of the BSC's.

- The zero capacity result suggests that the Polya channel might not be a good model for a realistic channel. However in Section 2.7 we will consider a finite-memory channel that approximates the Polya channel as memory increases, but with a capacity that does *not* approach zero. Before we do so, however, we first point out that the Polya channel provides a counterexample to the adage "memory increases capacity"; this is the subject of the next section.

2.6 Effect of Memory on the Capacity of the Polya Channel

In [10], Pinsker and Dobrushin showed that "for a wide class" of channels, the capacity of a channel with memory is *not less* than the capacity of the "equivalent" memoryless channel. They considered a channel with input alphabet A, output alphabet B, and n-fold transition probability $W^{(n)}(y_1, \ldots, y_n \mid x_1, \ldots, x_n)$, $x_i \in A$, $y_i \in B$, such that:

$$W^{(n-1)}(y_1, \ldots, y_{n-1} \mid x_1, \ldots, x_{n-1}) = \sum_{y_n \in B} W^{(n)}(y_1, \ldots, y_n \mid x_1, \ldots, x_n). \quad (2.9)$$

Specifically, they considered such channels with an operational capacity given by

$$C = \lim_{n \to \infty} \frac{1}{n} C_n \quad (2.10)$$

where

$$C_n = \sup_{(X_1, \ldots, X_n)} I\left((X_1, \ldots, X_n); (Y_1, \ldots, Y_n)\right).$$

Here, $I\left((X_1, \ldots, X_n); (Y_1, \ldots, Y_n)\right)$ is the block mutual information between the input and the output. They then defined the memoryless channel associated with this channel to have n-fold transition probability

$$\tilde{W}^{(n)}(y_1, \ldots, y_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} \tilde{W}_i(y_i | x_i),$$

where

$$\tilde{W}_i(y_i \mid x_i) = \sum_{y_1, \ldots, y_{i-1} \in B} W^{(i)}(y_1, \ldots, y_i \mid x_1, \ldots, x_i) \quad (2.11)$$

and the left-hand side of (2.11) is assumed to be independent of $(x_1, \ldots, x_{n-1})$. Thus the one-step transition probabilities of the memoryless channel are equal to

the per-letter marginals of the channel with memory. The capacity of the associated memoryless channel is denoted by $\tilde{C}_n$, and they showed that:

$$C_n \geq \tilde{C}_1 + \tilde{C}_2 + \cdots + \tilde{C}_n.$$

By "a wide class of channels", they made some implicit assumptions:

- The channels are non-anticipatory (causal) and historyless (with no input memory). The non-anticipatory property can be seen from equation (2.9), where it is implicitely assumed that

  $$W^{(n-1)}(y_1, \ldots, y_{n-1}|x_1, \ldots, x_{n-1}) = W^{(n-1)}(y_1, \ldots, y_{n-1}|x_1, \ldots, x_n).$$

  Furthermore, equation (2.11) assumes the distribution on the n^{th} channel output given the first n channel inputs depends only on the n^{th} input – i.e., the original channel has no input memory; it is historyless.

- The channels are *information stable.* Information stable channels have the property that the input that maximizes mutual information and its corresponding output behave ergodically [35, 25]. The class of information stable channels is the most general class of channels for which equation (2.10) represents the operational capacity.

If we restrict ourselves to stationary channels, then $\tilde{W}_n(y_n \mid x_n) = \tilde{W}(y_n \mid x_n)$ for all n and so $C_n \geq n\tilde{C}$, where $\tilde{C}_1 = \tilde{C}_2 = \ldots = \tilde{C}_n = \tilde{C}$. Thus we get $C \geq \tilde{C}$.

In [2], Ahlswede showed that there are averaged channels for which the introduction of memory *decreases* capacity. We briefly show that the Polya-contagion channel is such a channel.

We showed in the previous section that the capacity of the Polya channel is zero. Now, let us compute $\tilde{C}$, the capacity of the associated memoryless channel. The transition probability of the associated memoryless channel $\tilde{W}(\cdot)$ is

$$\begin{aligned}
\tilde{W}(Y_n = y_n | X_n = x_n) &= \sum_{y_1,\ldots,y_{n-1} \in \{0,1\}} \int_\Theta W_\theta^{(n)}(\underline{Y} = \underline{y} | \underline{X} = \underline{x}) \; dG(\theta) \\
&= \int_\Theta \sum_{y_1,\ldots,y_{n-1} \in \{0,1\}} W_\theta^{(n)}(\underline{Y} = \underline{y} | \underline{X} = \underline{x}) \; dG(\theta) \\
&= \int_\Theta W_\theta^{(1)}(Y_n = y_n | X_n = x_n) \; dG(\theta) \\
&= \int_0^1 \theta^{y_n \oplus x_n} \; (1-\theta)^{1-(y_n \oplus x_n)} \; f_\Theta(\theta) \; d\theta \\
&= \rho^{y_n \oplus x_n} \; (1-\rho)^{1-(y_n \oplus x_n)},
\end{aligned}$$

where $\Theta = [0,1]$, and $dG(\theta) = f_\Theta(\theta) \; d\theta$ is the beta distribution with parameters ρ/δ and σ/δ given in Proposition 3.

Thus we observe that the memoryless channel equivalent to the Polya channel is a BSC with crossover probability ρ; this leads us to the conclusion that, for $\rho \neq 1/2$, the memory in the Polya channel *decreases* capacity.

One can obtain an even simpler example of an averaged channel which has a smaller capacity than the associated memoryless channel. Consider an averaged channel consisting of two BSC's - one with crossover probability 0 and the other with crossover probability $1/2$ - and the two BSC's are equally likely. Then simple calculations reveal that the averaged channel has zero capacity, while the equivalent memoryless channel has capacity $1 - h_b(1/4)$, where $h_b(\cdot)$ is the binary entropy function.

On the other hand, one can derive examples of stationary non-ergodic channels for which memory increases capacity. Consider an averaged channel consisting of two BSC's - one with crossover probability 0 and the other with crossover probabil-

ity 1 - and the two BSC's are equally likely. The capacity of the averaged channel is now 1, while the equivalent memoryless channel has zero capacity. This leads us to conclude that for arbitrary discrete channels, memory may *increase* or *decrease* capacity.

2.7 Finite-Memory Contagion Channel

An unrealistic aspect of the Polya channel is its infinite memory. Consider, for instance, the millionth ball drawn from Polya's urn; the very *first* ball drawn from the urn and the 999,999'th ball drawn from the urn have an identical effect on the outcome of the millionth draw. In the context of a communication channel, this is not reasonable; we would assume that the effects of the "disease" fade in time. We now consider a more realistic model for a contagion channel with finite memory, where the noise in the additive channel is generated according to a modified version of the Polya urn scheme.

Assume once again that the channel output Y_i is the modulo-two sum of the input X_i and the noise Z_i; as for the Polya channel, assume that the input and noise sequences are independent. Then $\{Z_i\}_{i=1}^{\infty}$ is drawn according to the following urn scheme: An urn initially contains T balls – R red and S black ($T = R + S$). At the j'th draw, $j = 1, 2, \ldots$, we select a ball from the urn and replace it with $1 + \Delta$ balls of the same color ($\Delta > 0$); then, M draws later – after the $(j + M)$'th draw – we retrieve from the urn Δ balls of the color picked at time j. Once again let $\rho = R/T < 1/2$, $\sigma = 1 - \rho = S/T$ and $\delta = \Delta/T$. Then the noise process $\{Z_i\}$ corresponds to the outcomes of the draws from the urn, where:

$$Z_i = \begin{cases} 1, & \text{if the } i^{th} \text{ ball drawn is red;} \\ 0, & \text{if the } i^{th} \text{ ball drawn is black.} \end{cases}$$

Observation: With this modification of the original Polya urn scheme, the number of balls in the urn is constant ($T + M\Delta$ balls) after an initialization period of M draws. It also limits the effect of any draw to M draws in the future.

2.7.1 The Distribution of the Noise

During the initialization period ($n \leq M$), the process $\{Z_i\}$ of the finite-state channel is identical to the Polya noise process discussed earlier. We now study the noise process for $n \geq M + 1$.

Let R_n be the number of red balls in the urn after n draws, T_n be the total number of balls in the urn after n draws, and $r_n = R_n/T_n$. Then $T_n = T + M\Delta$ for $n \geq M + 1$, and so

$$\begin{aligned} r_n &= \frac{R + (Z_n + Z_{n-1} + \cdots + Z_{n-M+1})\Delta}{T + M\Delta} \\ &= \frac{\rho + (Z_n + Z_{n-1} + \cdots + Z_{n-M+1})\delta}{1 + M\delta}. \end{aligned}$$

We now have that:

$$\begin{aligned} P(Z_n = 1 | Z_1 = e_1, \ldots, Z_{n-1} = e_{n-1}) &= \frac{\rho + (e_{n-1} + e_{n-2} + \cdots + e_{n-M})\delta}{1 + M\delta} \\ &= r_{n-1} \\ &= P(Z_n = 1 | Z_{n-M} = e_{n-M}, \ldots, Z_{n-1} = e_{n-1}), \end{aligned}$$

where $e_i \in \{0, 1\}$. Thus the noise process $\{Z_i\}_{i=M+1}^{\infty}$ is a Markov process of order M. We shall refer to the resulting channel as the finite-memory contagion channel. For an input block $\underline{X} = [X_1, X_2, \ldots, X_n]$ and an output block $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]$, the block transition probability of the resulting binary channel is as follows:

- For blocklength $n \leq M$, the block transition probability of this channel is identical to that of the Polya-contagion channel given by equations (2.1) and (2.2).

- For $n \geq M+1$, we obtain:

$$\begin{aligned} P_M(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) &= P(\underline{Z} = \underline{e}) \\ &= \prod_{i=1}^{n} P(Z_i = e_i \mid Z_{i-1} = e_{i-1}, \ldots, Z_{i-M} = e_{i-M}) \\ &= L \prod_{i=M+1}^{n} \left[\frac{\rho + s_{i-1}\delta}{1 + M\delta}\right]^{e_i} \left[\frac{\sigma + (M - s_{i-1})\delta}{1 + M\delta}\right]^{1-e_i} \end{aligned} \tag{2.12}$$

where

$$L = \frac{\prod_{i=0}^{k-1}(\rho + i\delta)\prod_{j=0}^{M-1-k}(\sigma + j\delta)}{\prod_{\ell=1}^{M-1}(1 + \ell\delta)},$$

$$e_i = x_i \oplus y_i,$$

$$k = e_1 + \cdots + e_M,$$

and

$$s_{i-1} = e_{i-1} + \cdots + e_{i-M}.$$

By examining the above equation we can show that the noise process (and so the channel) is *stationary* since the block distribution of the noise is invariant under all finite shifts, i.e.

$$P(Z_1 = e_1, Z_2 = e_2, \ldots, Z_n = e_n) = P(Z_{1+\tau} = e_1, Z_{2+\tau} = e_2, \ldots, Z_{n+\tau} = e_n)$$

for all finite $\tau > 0$.

Remark: Obviously, as M grows, the n-fold transition distribution of the finite-memory contagion channel converges to the n-fold transition distribution of the Polya-contagion channel, i.e. $P_M^{(n)}(\cdot) \to P_{Polya}^{(n)}(\cdot)$.

2.7.2 Properties of the Noise Process

We now consider the properties of the noise process $\{Z_i\}$. Define $\{W_n\}$ to be the process obtained by M-step blocking $\{Z_n\}$ – i.e. $W_n = (Z_n, Z_{n+1}, \ldots, Z_{n+M-1})$. Then $\{W_n\}$ is a one-step Markov process with 2^M states; we denote each state by its decimal representation; i.e. state 0 corresponds to state $(0\cdots00)$, state 1 corresponds to state $(0\cdots01)$, ..., and state (2^M-1) corresponds to state $(1\cdots11)$.

Tedious calculations reveal the following properties about the process $\{W_n\}$.

- $\{W_n\}$ is a homogeneous stationary Markov process with stationary distribution $\Pi = [\pi_0, \pi_1, \ldots, \pi_{2^M-1}]$, where π_i is computed as follows. Let $w(i)$ denote the number of 1's in the binary representation of the decimal integer i, where $i \in \{0, 1, \ldots, 2^M - 1\}$. Then

$$\pi_i = \frac{\prod_{j=0}^{w(i)-1} (\rho + j\delta) \prod_{k=0}^{M-1-w(i)} (\sigma + k\delta)}{\prod_{\ell=1}^{M-1} (1 + \ell\delta)}.$$

- If we let $\{p_{ij}\}$ be the one-step transition probabilities – i.e., $p_{ij} \triangleq Pr(W_n = j | W_{n-1} = i)$, $i, j \in \{0, 1, \ldots, 2^M - 1\}$ – then

$$p_{ij} = \begin{cases} \frac{\sigma + (M - w(i))\delta}{1 + M\delta}, & \text{if } j = 2i \ (modulo\ 2^M); \\ \frac{\rho + w(i)\delta}{1 + M\delta}, & \text{if } j = (2i+1) \ (modulo\ 2^M); \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

We've thus shown that the Markov process $\{W_n\}$ is irreducible and aperiodic; therefore it is *strongly mixing* [24] and hence ergodic. Since the additive noise process is stationary and mixing, the resulting additive noise channel is *ergodic* [16, 25].

Observation: For $M = 1$ the one-step transition probability matrix of $\{Z_n\}$ is

$$Q \triangleq [p_{ij}] = \frac{1}{1+\delta} \begin{bmatrix} \sigma+\delta & 1-\sigma \\ \sigma & 1-\sigma+\delta \end{bmatrix} \tag{2.14}$$

Clearly one can choose δ and σ to "match" the transition probabilities of an arbitrary irreducible two-state Markov chain.

2.7.3 Capacity of the Finite-Memory Contagion Channel

Using the results in the previous subsection, we arrive at the following proposition.

Proposition 5 The capacity C_M of M-memory contagion channel is non-decreasing in M. It is given by:

$$C_M = 1 - \sum_{k=0}^{M} \binom{M}{k} L_k \; h_b\left(\frac{\rho + k\delta}{1+M\delta}\right), \tag{2.15}$$

where

$$L_k = \frac{\prod_{j=0}^{k-1} (\rho + j\delta) \; \prod_{\ell=0}^{M-k-1} (\sigma + \ell\delta)}{\prod_{m=1}^{M-1} (1+m\delta)},$$

and $h_b(\cdot)$ is the binary entropy function.

Proof 5 Since the channel is stationary ergodic (and hence information stable), its capacity is given by equation (2.10) which yields

$$\begin{aligned} C_M &= 1 - H(Z_{M+1} \mid Z_M, Z_{M-1}, \ldots, Z_1) \\ &= 1 + \sum_{i,j=0}^{2^M-1} \pi_i \; p_{ij} \; \log_2 p_{ij} \\ &= 1 - \sum_{k=0}^{M} \binom{M}{k} L_k \; h_b\left(\frac{\rho + k\delta}{1+M\delta}\right). \end{aligned} \tag{2.16}$$

The monotonicity of C_M in M follows from (2.16) because the Markov noise process is stationary and conditioning can only decrease entropy. ■

Proposition 6 The following equality holds:

$$\lim_{M\to\infty} C_M = 1 - \int_0^1 h_b(z)\ f_Z(z)dz, \tag{2.17}$$

where $f_Z(z)$ is the beta(ρ/δ, σ/δ) pdf and $h_b(\cdot)$ is the binary entropy function.

Proof 6 If we examine the quantity $\begin{pmatrix} M \\ k \end{pmatrix} L_k$ in the formula of C_M, we note that it is equal to the probability that $S_M = k$, where S_M is the state of the original Polya-contagion channel after the M'th draw, as defined in Section 2.2.1. We thus have:

$$\begin{aligned} C_M &= 1 - \sum_{k=0}^{M} h_b\left(\frac{\rho + k\delta}{1 + M\delta}\right)\ P(S_M = k) \\ &= 1 - \sum_{\tau\in\{k/M:k=0,1,\ldots,M\}} h_b\left(\frac{\frac{\rho}{M} + \tau\delta}{\frac{1}{M} + \delta}\right) P\left(\frac{S_M}{M} = \tau\right) \\ &= 1 - E_{T_M}\left[h_b\left(\frac{\frac{\rho}{M} + T_M\delta}{\frac{1}{M} + \delta}\right)\right], \end{aligned}$$

where $T_M = S_M/M$. We know by Property 3 in Section 2.2.2, that $T_M = S_M/M$ converges almost surely to a beta-distributed random variable Z with parameters ρ/δ and σ/δ. This almost surely convergence implies convergence in distribution. We now state the "weak equivalence theorem" [4]:

Lemma 6 Let the random variables X_n and X have respectively distributions μ_n and μ. Then the following two conditions are equivalent:

1. $\mu_n \Longrightarrow \mu$ as $n \longrightarrow \infty$; i.e. X_n converges to X in distribution.

2. $\int f d\mu_n \longrightarrow \int f d\mu$ as $n \longrightarrow \infty$, or equivalently $E_{X_n}[f(X_n)] \longrightarrow E_X[f(X)]$ as $n \longrightarrow \infty$ for every bounded, continuous real function $f(\cdot)$.

Therefore using the above lemma with the fact that $h_b(\cdot)$ is bounded and continuous, and that $\lim_{M\to\infty} h_b\left(\frac{\frac{\rho}{M}+\frac{k}{M}\delta}{\frac{1}{M}+\delta}\right) = h_b(z)$, we get:

$$\begin{aligned}\lim_{M\to\infty} E_{T_M}\left[h_b\left(\frac{\frac{\rho}{M}+\frac{S_M}{M}\delta}{\frac{1}{M}+\delta}\right)\right] &= E_Z[h_b(Z)] \\ &= \int_0^1 h_b(z)\ f_Z(z)dz.\end{aligned}$$

■

Proposition 7 If we let $I(\underline{X};\underline{Y})$ denote the mutual information between the input vector $\underline{X}$ and output vector $\underline{Y}$ connected over the original (non-ergodic) Polya channel, then

$$\lim_{n\to\infty}\frac{1}{n}\sup_{\underline{X}} I(\underline{X};\underline{Y}) = 1 - \int_0^1 h_b(z)\ f_Z(z)dz.$$

where $f_Z(z)$ is the beta(ρ/δ, σ/δ) pdf.

Proof 7 We have

$$\sup_{\underline{X}}\ \frac{1}{n}\ I(\underline{X};\underline{Y}) = 1 - \ \frac{1}{n}\ H(\underline{Y}\mid \underline{X}) = 1 - \ \frac{1}{n}\ \sum_{j=1}^{n}\ H(Y_j \mid Y^{j-1}, X^j),$$

where $\underline{X} = X^n = [X_1,\ldots,X_n]$ and $X^j = [X_1,\ldots,X_j]$. From Section 2.2.1, we know that

$$\begin{aligned}P(Y_j = y_j | Y^{j-1} = y^{j-1}, X^j = x^j) &= P(Y_j = y_j \mid X_j = x_j, S_{j-1} = s_{j-1}) \\ &= \begin{cases}\left[\frac{\rho+s_{j-1}\delta}{1+(j-1)\delta}\right], & \text{if } y_j \oplus x_j = 1; \\ \left[\frac{\sigma+(j-1-s_{j-1})\delta}{1+(j-1)\delta}\right], & \text{if } y_j \oplus x_j = 0,\end{cases}\end{aligned}$$

where S_{j-1} is the state of the Polya channel after $j-1$ draws.

We can then compute $H(Y_j \mid Y^{j-1}, X^j)$ to obtain:

$$\begin{aligned} H(Y_j \mid Y^{j-1}, X^j) &= H(Y_j \mid X_j, S_{j-1}) \\ &= \sum_{k=0}^{j-1} h_b\left(\frac{\rho + k\delta}{1 + (j-1)\delta}\right) P(S_{j-1} = k). \end{aligned}$$

As in the previous proof, we use the "weak equivalence theorem" to get

$$\lim_{j\to\infty} H(Y_j \mid Y^{j-1}, X^j) = \int_0^1 h_b(z)\ f_Z(z)dz,$$

where $f_Z(\cdot)$ is given as above. Finally using the Cesaro-mean average, we obtain

$$\lim_{n\to\infty} \frac{1}{n} H(\underline{Y} \mid \underline{X}) = \lim_{j\to\infty} H(Y_j \mid Y^{j-1}, X^j) = \int_0^1 h_b(z)\ f_Z(z)dz.$$

■

Observation: As the memory M grows, $P_M^{(n)}(\cdot) \to P_{Polya}^{(n)}(\cdot)$, but the capacity C_M of the (ergodic) finite-memory channel *does not converge* to the capacity of the (non-ergodic) Polya-contagion channel (which is zero). On the contrary, C_M increases in M and converges to $1 - \int_0^1 h_b(z)\ f_Z(z)dz$. In addition, if we let $I(\underline{X}; \underline{Y})$ denote the mutual information between the input vector $\underline{X}$ and output vector $\underline{Y}$ connected over the original (non-ergodic) Polya channel, then

$$\lim_{n\to\infty} \frac{1}{n} \sup_{\underline{X}} I(\underline{X}; \underline{Y}) = 1 - \int_0^1 h_b(z)\ f_Z(z)dz. \tag{2.18}$$

The left side of equation (2.18) is called the *information rate capacity* of the Polya channel; we have thus demonstrated that, as we let the memory in the finite-memory contagion channel increase, not only does the channel block transition distribution converge to that of the Polya channel, but the information rate capacities also converge to that of the Polya channel. However, there is no convergence

in the *operational* capacity. It seems reasonable to assume that this is due to the non-ergodic nature of the Polya channel. In the following proposition we examine this question.

Proposition 8 Consider a sequence of historyless non-anticipatory stationary ergodic channels; let the n-fold transition probability of the M^{th} channel be denoted $W_M^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$. Let C_M denote the capacity of the M^{th} channel. Finally, suppose this sequence of channels satisfies the following conditions.

1. As M grows, their n-fold transition distributions converge to the n-fold transition distribution of a historyless non-anticipatory stationary channel – i.e., if we let $W_*^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$ denote the n-fold transition probability of the limiting channel, then for any real n-tuples $\underline{x}$ and $\underline{y}$,

$$\lim_{M\to\infty} W_M^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) = W_*^{(n)}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}).$$

2. The "information rate capacities" of the channels converge to that of the limiting channel – i.e., if $I_M(\underline{X};\underline{Y})$ denotes the n-fold mutual information between the inputs and outputs of the M^{th} channel, and $I_*(\underline{X};\underline{Y})$ denotes the same for the limiting channel, then

$$\lim_{M\to\infty} \lim_{n\to\infty} \frac{1}{n} \sup_{\underline{X}} I_M(\underline{X};\underline{Y}) = \lim_{n\to\infty} \frac{1}{n} \sup_{\underline{X}} I_*(\underline{X};\underline{Y}).$$

Let C_* denote the capacity of the limiting channel; then a *sufficient but not necessary* condition that

$$\lim_{M\to\infty} C_M = C_*$$

is that the limiting channel be *ergodic.*

Proof 8 The proof that ergodicity is sufficient is trivial. All the channels are ergodic, so the information rate capacities are equal to the corresponding operational capacities; condition (2.) says the information rate capacities converge, and so the operational capacities must too.

To see that ergodicity is not necessary, we briefly sketch a counter-example. Let $\{U_i^M\}_{i=0}^{\infty}$ be a stationary, mixing binary Markov process indexed by the parameter M; assume that $P(U_0^M = 0) = P(U_0^M = 1) = 1/2$ and that U_i^M has one-step transition matrix

$$Q_{U^M} = \begin{pmatrix} 2^{-M} & 1-2^{-M} \\ 1-2^{-M} & 2^{-M} \end{pmatrix}.$$

We create a noise process $\{Z_i^M\}_{i=0}^{\infty}$ by two-blocking the process $\{U_i^M\}$ – i.e., $Z_i^M = (Z_{i1}^M, Z_{i2}^M) = (U_{2i}^M, U_{2i+1}^M)$ for $i = 0, 1, 2, \ldots$. Then $\{Z_i^M\}$ is a one-step Markov chain with four states and transition matrix

$$Q_{Z^M} = \begin{pmatrix} 2^{-2M} & 2^{-M}(1-2^{-M}) & (1-2^{-M})^2 & 2^{-M}(1-2^{-M}) \\ 2^{-M}(1-2^{-M}) & (1-2^{-M})^2 & 2^{-M}(1-2^{-M}) & 2^{-2M} \\ 2^{-2M} & 2^{-M}(1-2^{-M}) & (1-2^{-M})^2 & 2^{-M}(1-2^{-M}) \\ 2^{-M}(1-2^{-M}) & (1-2^{-M})^2 & 2^{-M}(1-2^{-M}) & 2^{-2M} \end{pmatrix}.$$

Now consider the channel with input/output alphabet $\{00, 01, 10, 11\}$, where the i^{th} input $X_i = (X_{i1}, X_{i2})$ is related to the i^{th} output $Y_i = (Y_{i1}, Y_{i2})$ by $Y_i = (X_{i1} \oplus Z_{i1}^M, X_{i2} \oplus Z_{i2}^M)$. $\{Z_i^M\}$ is a stationary mixing process; thus the channel is stationary ergodic [25]. For finite M, the capacity – both operational and information rate – is give n by $C_M = 2 - H(Z_2^M | Z_1^M)$ bits/channel use. From Q_{Z^M} we observe that $\lim_{M\to\infty} H(Z_2^M | Z_1^M) = 0$; thus, $\lim_{M\to\infty} C_M = 2$ bits/channel use.

As M increases, the process $\{Z_i^M\}$ converges in distribution to a stationary non-ergodic process $\{Z_i^*\}$ with two equiprobable components – $\{01, 01, 01, 01, \ldots\}$ and $\{10, 10, 10, 10, \ldots\}$. The information rate capacity of this – that is,

$\lim_{n\to\infty} \max_{\underline{X}} (1/n) I_*(\underline{X};\underline{Y})$ – is two bits/channel use. Thus both of the conditions above are met. However, this limiting channel is a mixture of two deterministic channels, and its operational capacity is also two bits per channel use. Thus the ergodicity of the limiting channel is not a necessary condition.

■

2.8 Summary

In this chapter we considered a discrete channel with memory in which errors "spread" like the spread of a contagious disease through a population; the channel is based on Polya's model for contagion. The channel is stationary and non-ergodic. We first presented a maximum likelihood (ML) decoding algorithm for the channel, and then showed that this channel is in fact an "averaged" channel, and its capacity is zero. Using De Finetti's results on exchangeability, we noted that binary channels with additive exchangeable noise processes are averaged channels with binary symmetric channels as components. The zero capacity result illustrates a counter-example to the adage "memory can only increase capacity". The ϵ-capacity of the Polya channel was also derived.

Finally, we considered a finite-memory version of the Polya-contagion model. The resulting channel is a stationary ergodic Markov channel with memory M; its capacity is positive and increases with M. As M increases, the n-fold transition distribution of the finite-memory contagion channel converges to the n-fold transition distribution of the original Polya-contagion channel, but its capacity does not converge to the capacity of the Polya channel.

Chapter 3

Feedback Capacity of Discrete Channels with Memory

3.1 Introduction

We consider discrete channels with additive random noise. Note that such channels need not be memoryless; in general, they have memory. The Gilbert burst-noise channel [12], as well as the Polya-contagion channel analyzed in the previous chapter, belong to the class of such channels. We assume that these channels are each accompanied by a noiseless, delayless feedback channel with large capacity. We show that the capacity of the channels with feedback does not exceed their respective capacity without feedback. This is shown for both ergodic and non-ergodic additive stationary noise processes. In light of recent results on a general channel capacity formula by Verdú and Han [35] we then generalize our result for discrete channels with *arbitrary* (non-stationary, non-ergodic in general) additive noise processes.

We remark that for these channels, the capacities with and without feedback

are equal because additive noise channels are *symmetric* channels. By this we mean that the block mutual information (respectively the *inf-information* rate for the case of arbitrary additive noise) between input and output processes is maximized by equally likely *iid* input process. Finally, we introduce the notion of *symmetric* channels with memory. These channels are obtained by combining an input process with an arbitrary noise process that is independent of the input. We also show that feedback does not increase the capacity of discrete symmetric channels. Additive noise channels belong to the class of symmetric channels.

In earlier related work, Shannon [30] showed that feedback does not increase the capacity of discrete memoryless channels. The same result was proven to be true for continuous channels with additive white Gaussian noise. Later, Cover and Pombra [8] and others considered continuous channels with additive non-white Gaussian noise and showed that feedback increases their capacity by at most half a bit; similarly, it has been shown [8] that feedback can at most double the capacity of a non-white Gaussian channel.

3.2 Discrete Channels with Stationary Ergodic Additive Noise

3.2.1 Capacity with no Feedback

Consider a discrete channel with common input, noise and output q-ary alphabet A where $A = \{0, 1, \ldots, q-1\}$, described by the following equation: $Y_n = X_n \oplus Z_n$, for $n = 1, 2, 3, \ldots$ where:

- $\oplus$ represents the addition operation modulo q.

- The random variables X_n, Z_n and Y_n are respectively the input, noise and output of the channel.

- $\{X_n\} \perp \{Z_n\}$, i.e. the input and noise sequences are independent from each other.

- The noise process $\{Z_n\}_{n=1}^{n=\infty}$ is stationary and ergodic.

Note that additive channels defined above, are "non-anticipatory" channels; where by "non-anticipatory" we mean channels with no input memory (i.e., historyless) and no anticipation (i.e., causal) [18]. Recall that a channel is said to have no anticipation if for a given input and a given input-output history, its current output is independent of future inputs. Furthermore, a channel is said to have no input memory if its current output is independent of previous inputs. Refer to [18] for more rigorous definitions of causal and historyless channels.

We furthermore note that discrete additive noise channels are *symmetric* channels. Symmetric channels are channels for which the block mutual information (respectively the inf-information rate for general channels) is maximized by equally likely *iid* input process. This class of channels will be considered in Section 3.5. This is due to the facts that the input and noise processes of the channel are independent from each other, the addition operation (modulo q) is invertible and the input and output alphabets are *finite* and have the *same cardinality.*

A channel code with blocklength n and rate R consists of an encoder

$$f : \{1, 2, \ldots, 2^{nR}\} \rightarrow A^n$$

and a decoder

$$g : A^n \rightarrow \{1, 2, \ldots, 2^{nR}\}.$$

The encoder represents the message $V \in \{1, 2, \ldots, 2^{nR}\}$ with the codeword $f(V) = X^n = [X_1, X_2, \ldots, X_n]$ which is then transmitted over the channel; at the receiver, the decoder observes the channel output $Y^n = [Y_1, Y_2, \ldots, Y_n]$, and chooses as its estimate of the message $\hat{V} = g(Y^n)$. A decoding error occurs if $\hat{V} \neq V$.

For additive channels, $Y_i = X_i \oplus Z_i$ for all i. We assume that V is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$. The probability of decoding error is thus given by:

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} Pr\{g(Y^n) \neq V | V = k\} = Pr\{g(Y^n) \neq V\}$$

We say that a rate R is *achievable (admissible)* if there exists a sequence of codes with blocklength n and rate R such that

$$\lim_{n \to \infty} P_e^{(n)} = 0.$$

The objective is to find an admissible sequence of codes with as high a rate as possible. The capacity of the channel is defined as the supremum of the rate over all admissible sequences of codes. We denote it by C_{NFB}, to stand for capacity with no feedback.

Because the channel is a discrete channel with additive stationary ergodic noise, the nonfeedback capacity C_{NFB} of this channel is known and is equal to ([35], [22]):

$$\begin{aligned} C_{NFB} &= \lim_{n \to \infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n) \\ &= \log_2(q) - \lim_{n \to \infty} \frac{1}{n} H(Z^n), \end{aligned} \tag{3.1}$$

where

$$X^n = (X_1, X_2, \ldots, X_n),$$

$$Y^n = (Y_1, Y_2, \ldots, Y_n),$$

$$Z^n = (Z_1, Z_2, \ldots, Z_n),$$

$I(X^n; Y^n)$ is the mutual information between the input vector X^n and the output vector Y^n, and the supremum is taken over the input distributions of X^n. $H(Z^n)$ is the entropy of the noise vector Z^n. ' The expression in (3.1) can be shown to be the capacity of the channel using the Shannon-McMillan (AEP) theorem [22], [35].

3.2.2 Capacity with Feedback

We now consider the corresponding problem for the discrete additive channel with complete output feedback. By this we mean that there exists a "return channel" from the receiver to the transmitter; we assume this return channel is noiseless, delayless, and has large capacity. The receiver uses the return channel to inform the transmitter what letters were actually received; these letters are received at the transmitter before the next letter is transmitted, and therefore can be used in choosing the next transmitted letter.

A feedback code with blocklength n and rate R consists of sequence of encoders

$$f_i : \{1, 2, \ldots, 2^{nR}\} \times A^{i-1} \to A$$

for $i = 1, 2, \ldots, n$, along with a decoding function

$$g : A^n \to \{1, 2, \ldots, 2^{nR}\}.$$

The interpretation is simple: If the user wishes to convey message $V \in \{1, \ldots, 2^{nR}\}$ then the first code symbol transmitted is $X_1 = f_1(V)$; the second code symbol transmitted is $X_2 = f_2(V, Y_1)$, where Y_1 is the channel's output due to X_1. The third code symbol transmitted is $X_3 = f_3(V, Y_1, Y_2)$, where Y_2 is the channel's

output due to X_2. This process is continued until the encoder transmits $X_n = f_n(V, Y_1, Y_2, \ldots, Y_{n-1})$. At this point the decoder estimates the message to be $g(Y^n)$, where $Y^n = [Y_1, Y_2, \ldots, Y_n]$.

Assuming our additive channel, $Y_i = X_i \oplus Z_i$ where $\{Z_i\}$ is a stationary ergodic noise process. Again, we assume that V is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$, and we define the probability of error and achievability as in Section 3.2.1.

Note, however, that because of the feedback, X^n and Z^n are no longer independent; X_i may depend on Z^{i-1}.

We will denote the capacity of the channel with feedback by C_{FB}. As before, C_{FB} is the supremum of all admissible feedback code rates.

Proposition 9 Feedback does not increase the capacity of channels with additive stationary ergodic noise:

$$C_{FB} = C_{NFB} = \log_2(q) - \lim_{n\to\infty} \frac{1}{n} H(Z^n). \tag{3.2}$$

Proof 9 Since V is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$, we have that $H(V) = nR$. Furthermore, $H(V) = H(V|Y^n) + I(V; Y^n)$. Now by Fano's inequality,

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

since $h_b(P_e^{(n)}) \leq 1$, where $h_b(\cdot)$ is the binary entropy function.

We then have:

$$nR = H(V)$$

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

where R is any admissible rate.

Dividing both sides by n and taking n to infinity, we get:

$$C_{FB} \leq \lim_{n\to\infty} \frac{1}{n} I(V;Y^n). \tag{3.3}$$

Let us thus study $I(V;Y^n)$:

$$I(V;Y^n) = \sum_{i=1}^{n} I(V;Y_i|Y^{i-1}), \tag{3.4}$$

but

$$
\begin{aligned}
I(V;Y_i|Y^{i-1}) &= H(Y_i|Y^{i-1}) - H(Y_i|V,Y^{i-1}) && (3.5)\\
&= H(Y_i|Y^{i-1}) - H(X_i \oplus Z_i|V,Y^{i-1}). && (3.6)
\end{aligned}
$$

Now the fact that $X_i = f_i(V, Y_1, \ldots, Y_{i-1})$ implies that

$$
\begin{aligned}
H(X_i \oplus Z_i|V,Y^{i-1}) &= H(Z_i|V,Y^{i-1},X_i) && (3.7)\\
&= H(Z_i|V,Y^{i-1},X_i,X^{i-1},Z^{i-1}) && (3.8)\\
&= H(Z_i|V,Y^{i-1},X^i,Z^{i-1}) && (3.9)\\
&= H(Z_i|Z^{i-1}). && (3.10)
\end{aligned}
$$

Here,

- Equation (3.7) follows from the fact that given V and Y^{i-1}, X_i is known deterministically and $H(Z+X|X) = H(Z|X)$.
- Equations (3.8) and (3.9) follow from the fact that given V and Y^{i-1}, we know all the previous transmitted letters $X_1, X_2, \ldots, X_{i-1}$ and thus we can recover all the previous noise letters $Z_j = Y_j - X_j \pmod q$ for $j = 1, 2, \ldots, i-1$.

- Equation (3.10) follows from the fact that Z_i and (V, Y^{i-1}, X^i) are conditionally independent given Z^{i-1}.

Therefore

$$I(V; Y_i|Y^{i-1}) = H(Y_i|Y^{i-1}) \; - \; H(Z_i|Z^{i-1}), \tag{3.11}$$

and

$$\begin{aligned} I(V; Y^n) &= \sum_{i=1}^{n} \left[H(Y_i|Y^{i-1}) \; - \; H(Z_i|Z^{i-1})\right] && (3.12) \\ &= H(Y^n) \; - \; H(Z^n). && (3.13) \end{aligned}$$

But $H(Y^n) \leq \log_2 q^n$ because the channel is discrete. Therefore, if we divide both sides of (3.13) by n, and take n to infinity, we obtain that

$$C_{FB} \leq C_{NFB}.$$

But by definition of a feedback code, $C_{FB} \geq C_{NFB}$ since a non-feedback code is a special case of a feedback code. Thus we get:

$$C_{FB} = C_{NFB} = \log_2(q) \; - \lim_{n\to\infty} \frac{1}{n} H(Z^n) \tag{3.14}$$

. ■

Observations:

1. It is important to note that for additive channels, the conditional noise entropy (given in equations (3.7)-(3.10)) remains the same *with or without feedback.* This is because addition is invertible; in general $H(X) \geq H(f(X))$ with equality holding for invertible functions $f(\cdot)$. This is true for both discrete and continuous alphabet additive channels.

2. The reason why output feedback potentially increases the capacity of additive non-white Gaussian channels [8] is because for continuous channels we have power constraints on the input, which upon optimization may increase $\lim_{n\to\infty} \frac{1}{n} H(Y^n)$ when feedback is used; while for discrete channels this quantity is upperbounded by $\log_2(q)$ and cannot be increased with feedback. In particular for discrete additive channels, the output entropy rate is equal to $\log_2(q)$ without feedback (symmetry property). It is therefore suspected that feedback might increase the capacity of discrete additive channels if we impose power constraints on the input.

3. The result given in Proposition 9 can be easily extended to discrete non-anticipatory channels with additive asymptotically mean stationary (AMS) ergodic noise process. Such class of noise processes include time-homogeneous ergodic Markov chains with arbitrary initial distributions. The proof is identical to that of Proposition 9, since the non-feedback capacity for the channel with AMS ergodic additive noise is still given by equation (3.2) [16], [35]. A random process has the AMS property (or is an AMS process) if its sample averages converge for a sufficiently large class of measurements (e.g., the indicator functions of all events); furthermore, there exists a stationary measure, called the "stationary mean" of the process, that has the same sample averages. A necessary and sufficient condition for a random process to possess ergodic properties with respect to the class of all bounded measurements is that it is AMS [15].

Finally, with the result of Proposition 9 in mind, it would be interesting to investigate discrete *non-additive* channels with known non-feedback capacities, and see whether output feedback would increase their capacities.

3.3 Discrete Channels with Stationary Non-Ergodic Additive Noise

3.3.1 Capacity with no Feedback

Consider a discrete channel similar to the one considered in Section 3.2 with the exception that the additive noise process $\{Z_n\}$ to the channel is stationary but *non-ergodic*. We know from Proposition 2 in the previous chapter that the resulting channel is an averaged channel whose components are discrete channels with *additive* stationary *ergodic* noise.

The non-feedback capacity of this channel with additive non-ergodic noise is [21], [23]:

$$C_{NFB} = \log_2(q) - \text{ess}_\Theta \sup \; h(W_\theta), \tag{3.15}$$

where

- the noise entropy rate $h(W_\theta)$ is given by

$$h(W_\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} H_n(W_\theta^{(n)}), \tag{3.16}$$

  with

$$H_n(W_\theta^{(n)}) \triangleq - \sum_{x^n, y^n \in A^n} W_\theta^{(n)}(y^n|x^n)\; Q^{(n)}(x^n)\; \log_2\; W_\theta^{(n)}(y^n|x^n), \tag{3.17}$$

  where the input block distribution $Q^{(n)}(x^n) = \frac{1}{q^n}$.

- and the essential supremum is defined by

$$\text{ess}_\Theta \sup \; f(\theta) \triangleq \inf \; [r : dG(f(\theta) \leq r) = 1]\,. \tag{3.18}$$

3.3.2 Capacity with Feedback

As in the previous section, we consider the corresponding problem for the discrete additive channel with complete output feedback. Similarly, we define a feedback code with blocklength n and rate R, as a sequence of encoders

$$f_i : \{1, 2, \ldots, 2^{nR}\} \times A^{i-1} \to A$$

for $i = 1, 2, \ldots, n$, along with a decoding function

$$g : A^n \to \{1, 2, \ldots, 2^{nR}\}.$$

The interpretation of the functions is identical to those in Section 3.2.2.

Assuming our additive channel, $Y_i = X_i \oplus Z_i$ where $\{Z_i\}$ is a stationary non-ergodic noise process.

Here again, we assume that V is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$ and we use the same definitions of achievable rates, probability of decoding error and capacity as in Section 3.2.2.

Because of the feedback, X^n and Z^n are no longer independent; X_i depends causally on Z^{i-1}. We will denote the capacity of the channel *with feedback* by C_{FB}. We now get the following result:

Proposition 10 Feedback does not increase the capacity of channels with additive stationary non-ergodic noise:

$$C_{FB} = C_{NFB} = \log_2(q) - \text{ess}_\Theta \sup \; h(W_\theta).$$

Proof 10 The main idea of the proof is the following. The channel is an averaged channel whose components are stationary channels with additive ergodic noise. Since feedback does not increase the capacity of *each* of these components (as

shown in Section 3.2), it therefore does not increase the capacity of the averaged channel.

To formalize this reasoning, we will show that the (weak) converse to the channel coding theorem still holds with feedback. The coding theorem itself obviously holds since a non-feedback code is a special case of a feedback code, and thus any rate that can be achieved without feedback, can also be achieved with feedback.

The additive channel is a mixture of channels with *additive* stationary ergodic noise, thus by Proposition 9, we obtain that for *each of these components*:

$$C_{FB}^{(\theta)} = C_{NFB}^{(\theta)}.$$

Now, examining equation (3.15), we have: $h(W_\theta) \leq \text{ess}_\Theta \sup\ h(W_\theta)$ a.e. Then for some small $\epsilon > 0$, there exists components $\theta \in \Theta$ such that:

$$h(W_\theta) > \text{ess}_\Theta \sup\ h(W_\theta) - \epsilon,$$

or

$$\log_2(q) - h(W_\theta) < \log_2(q) - \text{ess}_\Theta \sup\ h(W_\theta) + \epsilon,$$

or

$$C_{NFB}^{(\theta)} < C_{NFB} + \epsilon.$$

And the probability of such components is $\delta > 0$.

By this we mean, that we can find among the stationary components, with probability $\delta > 0$, components with capacity $C_{NFB}^{(\theta)} < C_{NFB} + \epsilon$ for some small $\epsilon > 0$; i.e.

$$\delta = Pr\{\theta \in \Theta :\ C_{NFB}^{(\theta)} < C_{NFB} + \epsilon\} > 0.$$

With feedback encoder f_i and message $V = k$, we define

$$A(x_k^n) = \left\{ y^n \in A^n : f_i(k, y^{i-1}) = x_k^{(i)}\ (i = 1, 2, \ldots, n) \right\},$$

where $x_k^n = \left(x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(n)}\right)$. The probability that the feedback codeword for the message k is x_k^n is given by $P_k(x_k^n) = \sum_{y^n \in A(x_k^n)} W^{(n)}(Y^n = y^n | X^n = x_k^n)$. Letting D_k be the decoding set for message k, the probability that the feedback codeword for the message k is x_k^n *and* a decoding error takes place is given by $Pe_k(x_k^n) = \sum_{y^n \in A(x_k^n) \cap D_k^c} W^{(n)}(Y^n = y^n | X^n = x_k^n)$, where D_k^c is the complement of D_k.

Hence, the probability of decoding error when message k was sent is written as

$$\begin{aligned} Pe_k^{(n)} &= Pr\{g(Y^n) \neq V | V = k\} \\ &= \sum_{x_k^n \in A^n} \sum_{y^n \in A(x_k^n) \cap D_k^c} W^{(n)}(Y^n = y^n | X^n = x_k^n). \end{aligned} \tag{3.19}$$

It should be noted here that $A(x_k^n)$ and D_k^c in the above summation do not depend on the channel $W^{(n)}(\cdot)$. Therefore, the overall average probability of decoding error is

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} \sum_{x_k^n \in A^n} \sum_{y^n \in A(x_k^n) \cap D_k^c} W^{(n)}(Y^n = y^n | X^n = x_k^n). \tag{3.20}$$

It is evident using Proposition 2, that this probability of error can be expressed as

$$P_e^{(n)} = \int_\Theta P_e^{(n)}(\theta)\, dG(\theta). \tag{3.21}$$

where

$$P_e^{(n)}(\theta) = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} \sum_{x_k^n \in A^n} \sum_{y^n \in A(x_k^n) \cap D_k^c} W_\theta^{(n)}(Y^n = y^n | X^n = x_k^n), \tag{3.22}$$

and $P_e^{(n)}(\theta)$ is the average probability of decoding error for the channel component $W_\theta^{(n)}(\cdot)$.

Now, suppose there exists a sequence of feedback codes with blocklength n and rate R, such that $R > C_{NFB} + 2\epsilon$. Thus we have:

$$P_e^{(n)} = \int_\Theta P_e^{(n)}(\theta)\, dG(\theta)$$

$$\geq \int_{\{\theta\in\Theta: C_{NFB}^{(\theta)} < C_{NFB}+\epsilon\}} P_e^{(n)}(\theta)\, dG(\theta). \tag{3.23}$$

We now recall the weak converse to the nonfeedback channel coding theorem for stationary channels with additive ergodic noise: if $R > C_{NFB} + \epsilon^{'}$, for some small $\epsilon^{'} > 0$, then there exists $\gamma > 0$, such that $P_e^{(n)} > \gamma$ for sufficiently large n. To show this, using Fano's inequality along with the fact that $H(V) = nR$ we have

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

where the second inequality follows from the data processing theorem with $V \rightarrow X^n \rightarrow Y^n$ forming a Markov chain. Thus

$$\begin{aligned} P_e^{(n)} &\geq \frac{1}{R}\left(R - \frac{1}{n} I(X^n;Y^n) - \frac{1}{n}\right) \\ &\geq \frac{1}{R}\left(R - C_{NFB} - \frac{1}{n}\right) \\ &> \frac{1}{R}\left(\epsilon^{'} - \frac{1}{n}\right), \end{aligned}$$

and the result is shown. Note that $\gamma \triangleq \frac{1}{R}\left(\epsilon^{'} - \frac{1}{n}\right)$ is independent of the characteristics of the channel.

Therefore, applying the weak converse of the coding theorem for the stationary channel components with additive ergodic noise, we get that for $R > C_{NFB} + 2\epsilon > C_{NFB}^{(\theta)} + \epsilon$, there exists some small $\gamma > 0$, such that $P_e^{(n)}(\theta) > \gamma$, as $n \rightarrow \infty$. As mentioned above, γ is independent of θ and depends only on ϵ and R.

Then

$$\lim_{n\rightarrow\infty} P_e^{(n)} > Pr\{\theta \in \Theta :\ C_{NFB}^{(\theta)} < C_{NFB} + \epsilon\}\gamma = \delta\ \gamma > 0. \tag{3.24}$$

Therefore the weak converse is proved and $C_{FB} = C_{NFB}$. ■

Observation: It should be noted that for general averaged channels, i.e. *non-additive* averaged channels, feedback might *increase* capacity. For example, if we consider an averaged channel with a *finite* number of *non-additive* discrete memoryless channels (DMC's), then the non-feedback capacity of the averaged channel is equal to the capacity of the corresponding compound memoryless channel [1]:

$$C_{NFB}^{(ac)} = \max_{Q^{(1)}} \inf_{\theta \in \Theta} I(Q^{(1)}; W_\theta^{(1)}). \tag{3.25}$$

Note that:

$$\begin{aligned} C_{NFB}^{(ac)} &\leq \inf_{\theta \in \Theta} \max_{Q^{(1)}} I(Q^{(1)}; W_\theta^{(1)}) \\ &= \inf_{\theta \in \Theta} C^{(\theta)}, \end{aligned} \tag{3.26}$$

where $C^{(\theta)} = \max_{Q^{(1)}} I(Q^{(1)}; W_\theta^{(1)})$ is the non-feedback capacity of each of the DMC components.

Now, if we use output feedback, the encoder knows the previous received outputs, and thus can determine by some statistical means, which one of the DMC components is being used. In the most pessimistic case, the capacity of this DMC component may be equal to $\inf_{\theta \in \Theta} C^{(\theta)}$. Thus the capacity with feedback of the averaged channel will be:

$$C_{FB}^{(ac)} = \inf_{\theta \in \Theta} C^{(\theta)}. \tag{3.27}$$

Therefore $C_{FB}^{(ac)} \geq C_{NFB}^{(ac)}$. This result (equation (3.27)) is equivalent to the result already derived by Ahlswede for the discrete averaged channel with sender informed [2].

Finally, in the case for which the inequality in (3.26) holds with the *strict* inequality, we obtain that feedback *increases* capacity: $C_{FB}^{(ac)} > C_{NFB}^{(ac)}$. Refer to Section 2 in [5] for an example of a finite collection of DMC's for which (3.26) holds with the strict inequality.

3.4 Discrete Channels with Arbitrary Additive Noise

3.4.1 Capacity with no Feedback

Consider a discrete channel similar to the one considered in Section 3.2 with the exception that the additive noise process $\{Z_n\}$ to the channel is an *arbitrary* random process (non-stationary, non-ergodic in general). We again use the same definitions as stated in Section 3.2.1 for channel block code, probability of error, achievable (or admissible) code rates and operational capacity (the supremum of all achievable rates). We denote the nonfeedback capacity by C_{NFB}.

In [35], Verdú and Han derived a formula for the operational capacity of arbitrary single-users channels (not necessarily stationary, ergodic, information stable, etc.). The rbitrary single-users channels (not necessarily stationary, ergodic, information stable, etc.). The (nonfeedback) capacity was shown to equal the supremum, over all input processes, of the input-output *inf-information rate* defined as the liminf in probability of the normalized information density:

$$C_{NFB} = \sup_{X^n} \underline{I}\,(X^n; Y^n), \tag{3.28}$$

where $X^n = (X_1, \ldots, X_n)$, for $n = 1, 2, \ldots$, is the block input vector and $Y^n = (Y_1, \ldots, Y_n)$ is the corresponding output sequence induced by X^n via the channel $W^{(n)} = P_{Y^n|X^n} : A^n \to B^n$; $n = 1, 2, \ldots$, which is an arbitrary sequence of n-dimensional conditional output distributions from A^n to B^n, where A and B are the input and output alphabets respectively.

The symbol $\underline{I}\,(X^n; Y^n)$ appearing in (3.28) is the *inf-information rate* between

X^n and Y^n and is defined as the *liminf in probability* of the sequence of normalized information densities $\frac{1}{n}\, i_{X^nY^n}(X^n;Y^n)$, where

$$i_{X^nY^n}(a^n;b^n) = \log_2 \frac{P_{Y^n|X^n}(b^n|a^n)}{P_{Y^n}(b^n)}. \tag{3.29}$$

The *liminf in probability* of a sequence of random variables is defined as follows: if A_n is a sequence of random variables, then its *liminf in probability* is the supremum of all reals α for which $P(A_n \leq \alpha) \longrightarrow 0$ as $n \longrightarrow \infty$. Similarly, its *limsup in probability* is the infimum of all reals β for which $P(A_n \geq \beta) \longrightarrow 0$ as $n \longrightarrow \infty$. Note that these two quantities are always defined; if they are equal, then the sequence of random variables converges in probability to a constant (which is α).

Using equation (3.28) as well as the properties of the inf-information rate derived in [35], we obtain that the inf–information rate in (3.28) is maximized for equiprobable iid X^n (symmetry property), yielding the following expression for the nonfeedback capacity of our discrete channel with arbitrary additive noise:

$$C_{NFB} = \log_2(q) - \overline{H}(Z^n), \tag{3.30}$$

where $Z^n = (Z_1, Z_2, \ldots, Z_n)$ and $\overline{H}(Z^n)$ is the sup-entropy rate of the additive noise process $\{Z_n\}$, which is defined as the limsup in probability of the normalized noise entropy density

$$\frac{1}{n} \log_2 \frac{1}{P_{Z^n}(Z^n)}.$$

3.4.2 Capacity with Feedback

As in the previous section, we consider the corresponding problem for the discrete additive channel with complete output feedback. Similarly, we use the same definitions as stated in Section 3.2.2 for feedback channel block code, probability of error, achievability and operational capacity with feedback (supremum of all achievable feedback code rates). We denote the capacity of the channel with feedback by C_{FB}.

Note again, that because of the feedback, X^n and Z^n are no longer independent; X_i may depend on Z^{i-1}.

We now state the key result (Theorem 4) of [35] which is a new converse approach based on a simple new lower bound on the error probability of an arbitrary channel code as a function of its size.

Lemma 7 Let (n, M, ϵ) represent a channel block code with blocklength n, M codewords and error probability ϵ. Then every (n, M, ϵ) code satisfies

$$\epsilon \geq P\left[\frac{1}{n}\, i_{X^nY^n}(X^n;Y^n) \leq \frac{1}{n}\log_2 M - \gamma\right] - \exp\left(-\gamma n\right) \tag{3.31}$$

for every $\gamma > 0$, where X^n places probability mass $1/M$ on each codeword.

We now obtain the following result:

Proposition 11 Feedback does not increase the capacity of discrete channels with *arbitrary* additive noise:

$$C_{FB} = C_{NFB} = \log_2(q) - \overline{H}(Z^n). \tag{3.32}$$

Proof 11 We start by noting that the result given in Lemma 7 still holds if we replace the input vector X^n by the message random variable V where V is uniform over the set of messages $\{1, 2, \ldots, M\}$. That is, every (n, M, ϵ) feedback code satisfies

$$\epsilon \geq P\left[\frac{1}{n} i_{VY^n}(V; Y^n) \leq \frac{1}{n}\log_2 M - \gamma\right] - \exp(-\gamma n) \tag{3.33}$$

for every $\gamma > 0$, where V is uniform over $\{1, 2, \ldots, M\}$.

We refer to the sequence (n, M, ϵ_n) of feedback codes with vanishingly small error probability (i.e., $\epsilon_n \to 0$ as $n \to \infty$) as a *reliable feedback code sequence.*

Using equation (3.33), we first show that

$$C_{FB} \leq \sup_{X^n} \underline{I}(V; Y^n), \tag{3.34}$$

where the supremum is taken over all possible feedback encoding schemes.[1]

We prove (3.34) by contradiction. Assume that for some $\rho > 0$,

$$C_{FB} = \sup_{X^n} \underline{I}(V; Y^n) + 3\rho. \tag{3.35}$$

By definition of capacity, there exists a reliable feedback code sequence with rate

$$R = \frac{1}{n}\log_2 M > C_{FB} - \rho. \tag{3.36}$$

Now using (3.33) (with $\gamma = \rho$) along with (3.35) and (3.36), we obtain that the error probability of the sequence (n, M, ϵ_n) of feedback codes must be lower bounded by

---

[1] $$\sup_{X^n} \underline{I}(V; Y^n) = \sup_{X^n = (f_1(V), f_2(V, Y_1), \ldots, f_n(V, Y^{n-1}))} \underline{I}(V; Y^n) = \sup_{(f_1, f_2, \ldots, f_n)} \underline{I}(V; Y^n).$$

$$\epsilon_n \geq P\left[\frac{1}{n}\, i_{VY^n}(V;Y^n) \leq \sup_{X^n} \underline{I}(V;Y^n) + \rho\right] - \exp(-\rho n). \tag{3.37}$$

However by definition of $\underline{I}(V;Y^n)$ the probability in the right-hand side of (3.37) cannot vanish asymptotically; therefore contradicting the fact that $\epsilon_n \to 0$ as $n \to \infty$. Thus (3.34) is proved.

Now using the properties of the inf-information rate in [35], we can write

$$\begin{aligned} \underline{I}(V;Y^n) &\leq \overline{H}(Y^n) - \overline{H}(Y^n|V) \\ &\leq \log_2(q) - \overline{H}(Y^n|V). \end{aligned} \tag{3.38}$$

The conditional sup-entropy rate $\overline{H}(Y^n|V)$ is the limsup in probability (according to P_{VY^n}) of $\frac{1}{n}\log_2 \frac{1}{P_{Y^n|V}(Y^n|V)}$. That is $\overline{H}(Y^n|V)$ is the infimum of all reals β such that $Pr\left\{\frac{1}{n}\log_2 \frac{1}{P_{Y^n|V}(Y^n|V)} \geq \beta\right\} \longrightarrow 0$, as $n \longrightarrow \infty$. But we can write

$$Pr\left\{\frac{1}{n}\log_2 \frac{1}{P_{Y^n|V}(Y^n|V)} \geq \beta\right\} = \sum_v P(V=v) \sum_{y^n : P(Y^n=y^n|V=v) \leq 2^{-n\beta}} P(Y^n = y^n|V=v).$$

Now, letting

$$f_i \stackrel{\triangle}{=} f_i(v, y^{i-1}),$$

and

$$f^i \stackrel{\triangle}{=} [f_1(v), f_2(v,y_1), \ldots, f_i(v,y^{i-1})] = [f_1, f_2, \ldots, f_i],$$

we have

$$\begin{aligned} P(Y^n = y^n|V=v) &= \prod_{i=1}^{n} P(Y_i = y_i|Y^{i-1} = y^{i-1}, V=v) \\ &= \prod_{i=1}^{n} P(X_i \oplus Z_i = y_i|Y^{i-1} = y^{i-1}, V=v, X_i = f_i) \end{aligned} \tag{3.39}$$

$$= \prod_{i=1}^{n} P(Z_i = y_i \oplus f_i | Y^{i-1} = y^{i-1}, V = v, X_i = f_i) \tag{3.40}$$

$$= \prod_{i=1}^{n} P(Z_i = y_i \oplus f_i | Y^{i-1} = y^{i-1}, V = v, X^i = f^i, Z^{i-1} = y^{i-1} \oplus f^{i-1}) \tag{3.41}$$

$$= \prod_{i=1}^{n} P(Z_i = y_i \oplus f_i | Z^{i-1} = y^{i-1} \oplus f^{i-1}) \tag{3.42}$$

$$= P(Z^n = y^n \oplus f^n). \tag{3.43}$$

Here,

- Equation (3.39) follows from the fact that $X_i = f_i(V, Y_1, \ldots, Y_{i-1})$ due to feedback.
- Equation (3.40) holds since $P(Z + X = y | X = x) = P(Z = y - x | X = x)$.
- Equation (3.41) follows from the fact that given V and Y^{i-1}, we know all the previous transmitted letters $X_1, X_2, \ldots, X_{i-1}$ and thus we can recover all the previous noise letters $Z_j = Y_j - X_j \pmod q$ for $j = 1, 2, \ldots, i-1$.
- Equation (3.42) follows from the fact that Z_i and (V, Y^{i-1}, X^i) are conditionally independent given Z^{i-1}.

Hence,

$$\begin{aligned} Pr\left\{\frac{1}{n}\log_2 \frac{1}{P_{Y^n|V}(Y^n|V)} \geq \beta\right\} &= \sum_v P(V = v) \sum_{y^n : P(Z^n = y^n \oplus f^n) \leq 2^{-n\beta}} P(Z^n = y^n \oplus f^n) \\ &= \sum_v P(V = v) \sum_{z^n : P(Z^n = z^n) \leq 2^{-n\beta}} P(Z^n = z^n) \\ &= \sum_{z^n : P(Z^n = z^n) \leq 2^{-n\beta}} P(Z^n = z^n). \end{aligned}$$

Therefore we obtain that

$$\overline{H}(Y^n|V) = \overline{H}(Z^n). \tag{3.44}$$

Thus from (3.34), (3.38) and (3.44) we conclude that

$$C_{FB} \leq \log_2(q) - \overline{H}(Z^n) = C_{NFB}. \tag{3.45}$$

But by definition of a feedback code, $C_{FB} \geq C_{NFB}$ since a non-feedback code is a special case of a feedback code. Thus we get:

$$C_{FB} = C_{NFB} = \log_2(q) - \overline{H}(Z^n). \tag{3.46}$$

■

Observation: Note that if the noise process is stationary, then its sup-entropy rate is equal to the supremum over the entropies of almost every ergodic component of the stationary noise. If the noise process is stationary ergodic, then its sup-entropy rate is equal to the entropy rate of the noise [35].

3.5 Discrete Symmetric Channels with Memory

Most physical channels produce outputs by combining the input process with a seperate noise process that is independent of the input signal. In light of this observation, we consider a physically motivated model for a discrete channel given by $Y_n = f(X_n, Z_n)$, for $n = 1, 2, 3, \ldots$ where:

- The random variables X_n, Z_n and Y_n are respectively the input, noise and output of the channel.
- The function $f : \mathcal{X} * \mathcal{Z} \to \mathcal{Y}$, is a mapping from the cartesian product of the input alphabet $\mathcal{X}$ and the noise alphabet $\mathcal{Z}$ into the output alphabet $\mathcal{Y}$. We assume that the sets $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{Y}$ are all finite.

- $\{X_n\} \perp \{Z_n\}$, i.e. the input and noise sequences are independent from each other.

- The noise process $\{Z_n\}_{n=1}^{n=\infty}$ is an *arbitrary* process (non-stationary non-ergodic in general).

Note that the channel described above is non-anticipatory (causal and historyless and thus it only has output memory (which is due to the memory in the noise process).

Definition 7 Consider the above channel given by $Y_n = f(X_n, Z_n)$, for $n = 1, 2, 3, \ldots$ We say that such a channel is *symmetric* if the following conditions hold:

1. $|\mathcal{X}| = |\mathcal{Z}| = q$; i.e., the input and noise alphabets have same cardinality.

2. Given the input x, the function $f(x, .)$ is one-to-one; i.e. for all $x \in \mathcal{X}$, $f(x, z) = f(x, \tilde{z}) \Longrightarrow z = \tilde{z}$.

3. $f^{-1}(\cdot)$ exists such that $z = f^{-1}(x, y)$ and given the output y, $f^{-1}(\cdot, y)$ is one-to-one; i.e. for all $y \in \mathcal{Y}$, $f^{-1}(x, y) = f^{-1}(\tilde{x}, y) \Longrightarrow x = \tilde{x}$.

Realize that condition (2) implies the following

$$P_{Y_n|X_n}(y_n|x_n) = P_{Z_n}(z_n), \tag{3.47}$$

where $z_n = f^{-1}(x_n, y_n)$.

Furthermore, conditions (1) and (3) guarantee that iid uniform input vectors yield iid uniform output vectors:

$$P(Y^n = y^n) = \sum_{x^n} P(Y^n = y^n | X^n = x^n) \left(\frac{1}{q}\right)^n$$

$$= \left(\frac{1}{q}\right)^n \sum_{x^n} P(Z^n = f^{-1}(x^n, y^n))$$
$$= \left(\frac{1}{q}\right)^n \sum_{z^n} P(Z^n = z^n) = \left(\frac{1}{q}\right)^n.$$

Note also that our definition of symmetric channels is consistent with the definition of *weakly symmetric* discrete memoryless channels (DMC's) as defined in [9]. That is if we take our noise process $\{Z_n\}$ to be iid, then by (3.47), all the rows of the square transition probability matrix of the channel will be permutations of each others. Furthermore, by conditions 1-3, all the columns will sum to 1 since

$$\sum_{x \in \mathcal{X}} P(Y = y|X = x) = \sum_{z \in \mathcal{Z}} P(Z = z) = 1.$$

Thus for symmetric channels (as defined above), $\underline{I}(X^n; Y^n)$ is maximized when the input process is equally likely Bernoulli (which yields an equally likely Bernoulli output process).

Refering to Section 3.2.2 for the definition of feedback, we obtain the following result:

Proposition 12 Feedback does not increase the capacity of discrete symmetric channels:

$$C_{FB} = C_{NFB} = \log_2(q) - \overline{H}(Z^n). \tag{3.48}$$

Proof 12 From (3.28) we have

$$C_{NFB} = \sup_{X^n} \underline{I}(X^n; Y^n)$$

But

$$\begin{aligned} \underline{I}(X^n; Y^n) &\leq \overline{H}(Y^n) - \overline{H}(Y^n|X^n) \\ &\leq \log_2(q) - \overline{H}(Y^n|X^n) \\ &= \log_2(q) - \overline{H}(Z^n), \end{aligned} \tag{3.49}$$

where the equality in (3.49) is due to (3.47).

Finally, (3.49) holds with equality for equiprobable iid X^n, and thus

$$C_{NFB} = \log_2(q) \ - \overline{H}(Z^n).$$

Using Definition 7 and equation (3.47), the proof of (3.48) follows directly from the proof of Proposition 11. ■

Example: Multiplicative Channel:

Consider the multiplicative channel with $\mathcal{X} = \mathcal{Z} = \mathcal{Y} = \{-1, 1\}$, given by $Y_n = X_n * Z_n$, for $n = 1, 2, \ldots$ where "$*$" represents the multiplication operation and $\{Z_n\}$ is an arbitrary noise process that is independent of the input process.

It can easily be verified that this channel is *symmetric* and that

$$C_{FB} = C_{NFB} = 1 \ - \overline{H}(Z^n). \tag{3.50}$$

Observation: It is pertinent to remark that channels with additive noise satisfy our definition of symmetry, and therefore belong to the class of *symmetric* channels.

3.6 Conclusion

In this chapter, we considered a discrete additive noise channel with output feedback. We showed that the capacity of the channel without feedback equals its capacity with feedback. This was first shown for a stationary ergodic and non-ergodic additive noise process. We then generalized the result for discrete channels with arbitrary additive noise.

We introduced the notion of *symmetric* channels with memory. These channels are obtained by combining an input process with an arbitrary noise process

that is independent of the input. These channels have the property that their inf-information rate is maximized when the input process is an equally likely iid process. We showed that feedback does not also increase the capacity of these channels. Additive noise channels belong to the class of symmetric channels.

Chapter 4

Capacity-Cost Function of Discrete Additive Markov Channels with and without Feedback

4.1 Introduction

We consider a binary communication channel with additive noise. The noise process is a stationary mixing Markov process of order M. We assume that the channel is accompanied by a noiseless, delayless feedback channel with large capacity.

Additive noise channels are *symmetric* channels; by this we mean that the block mutual information between input and output vectors of an additive channel is maximized for equiprobable input blocks (uniform iid input process). Furthermore, the uniform iid input process yields a uniform iid outpur process. In the previous chapter, we showed that the capacity of discrete additive channels with feedback, does not exceed their respective capacity without feedback. Indeed, this result holds for the class of discrete symmetric channels.

In achieving our result in Chapter 3, we did not introduce any restriction on

the possible input sequences of the channels. However, in many communication channels, not every combination of the elements of the input alphabet can be used for transmission; this is due to physical restrictions that may arise in the channels. We therefore incorporate average cost constraints on the input sequences of the Markov additive channel and analyze its capacity-cost function with and without feedback.

Since this additive channel with memory becomes *non-symmetric* under input constraints, closed form expressions for its nonfeedback and feedback capacity-cost functions do not exist. This lead us to establish bounds to the nonfeedback and feedback capacity-cost functions – denoted by $C_{NFB}(\beta)$ and $C_{FB}(\beta)$ respectively – in order to determine whether the introduction of output feedback brings about any potential increase in the capacity-cost function of the channel.

We first derive a *tight* upper bound to $C_{NFB}(\beta)$ which holds for all discrete channels with stationary mixing additive noise. The bound is the counterpart of the Wyner-Ziv lower bound to the rate-distortion function of stationary ergodic sources; this illustrates the striking duality that exists between the rate-distortion function and the capacity-cost function. A lower bound to $C_{NFB}(\beta)$ is also given.

We then study the capacity-cost function of the channel with feedback. An encoding feedback strategy is developed for the case where the M'th order Markov noise process is assumed to be generated by the finite memory conation urn scheme of Polya. It consists of adding at each instant of time, the maximum a posteriori (MAP) estimate of the current noise letter to the unencoded input letters representing the message to be sent. This results in a lower bound to $C_{FB}(\beta)$. An upper bound to $C_{FB}(\beta)$ is also derived, by proving the converse to the coding theorem.

The lower bound to $C_{FB}(\beta)$ and the upper bound to $C_{NFB}(\beta)$ are compared

numerically, using Blahut's algorithm for the computation of the capacity-cost function. The results, computed assuming power cost constraints on the input, demonstrate that the lower bound to $C_{FB}(\beta)$ *exceeds* the upper bound to $C_{NFB}(\beta)$ for certain parameters of the channel.

The rest of this chapter is organized as follows. In Section 4.2, we define the capacity-cost functions of stationary ergodic channels and present its properties. The analysis of the nonfeedback and feedback capacity-cost functions of the Markov channel is given in Section 4.3. In Section 4.4, numerical results indicating that $C_{FB}(\beta) > C_{NFB}(\beta)$ are provided. Finally, conclusions are stated in Section 4.5.

4.2 Preliminaries: The Capacity-Cost Function

Consider a discrete channel with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$ and n-fold transition probability $W^{(n)}(y_1, y_2, \ldots, y_n|x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i = 1, 2, \ldots, n$. In general, the use of the channel is not free; we associate with each input letter x a nonnegative number $b(x)$, that we call the "cost" of x. The function $b(\cdot)$ is called the cost function. We assume that the cost function is "bounded"; i.e., there exists $b_{\max}$ such that $b(x) \leq b_{\max}$ for all $x \in \mathcal{X}$. If we use the channel n consecutive times, i.e., we send an input vector $x^n = (x_1, x_2, \ldots, x_n)$, the cost associated with this input vector is "additive"; i.e.,

$$b(x^n) = \sum_{i=1}^{n} b(x_i).$$

For an input process $\{X_i\}_{i=1}^{\infty}$ with block input distribution $P^{(n)}(X^n = x^n) = P^{(n)}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, the *average cost* for sending X^n is defined by

$$E\left[b(X^n)\right] = \sum_{x^n} P^{(n)}(x^n)\, b(x^n) = \sum_{i=1}^{n} E\left[b(X_i)\right].$$

Definition 8 An n-dimensional input random vector $X^n = (X_1, X_2, \ldots, X_n)$ that satisfies

$$\frac{1}{n} E\left[b(X^n)\right] \leq \beta,$$

is called a *β-admissible* input vector. We denote the set of n-dimensional *β-admissible* input distributions by $\tau_n(\beta)$:

$$\tau_n(\beta) = \left\{ P^{(n)}(X^n) : \frac{1}{n} E\left[b(X^n)\right] \leq \beta \right\}.$$

We recall that a discrete channel is stationary if for every *stationary* input process $\{X_i\}_{i=1}^{\infty}$, the induced input-output process $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is stationary. Furthermore, a discrete channel is ergodic if for every *ergodic* input process $\{X_i\}_{i=1}^{\infty}$, the induced input-output process $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is ergodic [18].

A discrete channel with stationary mixing (or weakly mixing) additive noise is stationary and ergodic [16, 25]. Furthermore, it has no anticipation and no input memory. Recall that a channel is said to have no anticipation (i.e., causal) if for a given input and a given input-output history, its current output is independent of future inputs. Furthermore, a channel is said to have no input memory (i.e., historyless) if its current output is independent of previous inputs. Refer to [16, 18] for more rigorous definitions of causal and historyless channels.

Definition 9 A channel block code of length n over $\mathcal{X}$ is a subset

$$\mathcal{C} = \{c_{(1)}, c_{(2)}, \ldots, c_{(|\mathcal{C}|)}\}$$

of $\mathcal{X}^n$ where each $c_{(i)}$ is an n-tuple. The rate of the code is $R = \frac{1}{n}\log_2 |\mathcal{C}|$. The code is β-admissible if $b(c_{(i)}^n) \leq n\beta$ for $i = 1, 2, \ldots, |\mathcal{C}|$. If the encoder wants to transmit message W where W is uniform over $\{1, 2, \ldots, |\mathcal{C}|\}$, he sends the codeword $c_{(W)}$. At the channel output, the decoder receives Y^n and chooses as estimate of the

message $\hat{W} = g(Y^n)$, where $g(\cdot)$ is a decoding rule. The (average) probability of decoding error is then $P_e^{(n)} = Pr\{g(Y^n) \neq W\}$.

The *capacity-cost function* $C(\beta)$ is the supremum of all rates R for which there exist sequences of β-admissible block codes with vanishing probability of error as n grows to infinity (achievable codes). In other words, $C(\beta)$ is the maximum amount of information that can be transmitted reliably over the channel, if the channel must be used in such a way that the average cost is $\leq \beta$. If $b(x) = 0$ for every letter $x \in \mathcal{X}$, then $C(\beta)$ is just the channel *capacity* C as we know it.

It is known that under regularity conditions [16] (e.g., for a discrete stationary ergodic channel with no input memory and no anticipation and with a bounded and additive cost function [1]), the formulas of $C(\beta)$ and C are respectively given by:

$$C(\beta) = \sup_n \; C_n(\beta) = \lim_{n\to\infty} \; C_n(\beta), \tag{4.1}$$

where $C_n(\beta)$ is the *n'th capacity-cost function* given by

$$C_n(\beta) \stackrel{\triangle}{=} \max_{P^{(n)}(X^n)\in\tau_n(\beta)} \; \frac{1}{n} \, I(X^n;Y^n), \tag{4.2}$$

and

$$C = \sup_n \; C_n = \lim_{n\to\infty} \; C_n, \tag{4.3}$$

where

$$C_n \stackrel{\triangle}{=} \max_{P^{(n)}(X^n)} \; \frac{1}{n} \, I(X^n;Y^n), \tag{4.4}$$

[1] This follows from the dual result on the distortion rate function $D(R)$ of stationary ergodic sources (cf. next remark or p. 61 in [17] and Theorem 10.6 in [16]).

where $I(X^n;Y^n)$ is the block mutual information between the input vector X^n and the output vector Y^n. Indeed, the above formulas for capacity hold for a larger class of channels, the class of *information stable* channels [25, 35]. A general formula for the capacity of *arbitrary* channels (not necessarily information stable) was recently derived in [35].

Remark: Note that if the channel is stationary ergodic, and the cost function is additive and bounded, then there exists a stationary ergodic input process that achieves $C(\beta)$. This follows from the dual result on the distortion rate function $D(R)$ of stationary ergodic sources, which states that for a stationary ergodic source with additive and bounded distortion measure, there exists a stationary ergodic input-output process $P_{X^nY^n}$ that achieves $D(R)$ such that the induced marginal P_{X^n} is the source distribution [16, 17]. Therefore, the maximization in (4.2) can be taken with respect to n-dimensional β-admissible vector distributions $p^{(n)}(x^n)$ of stationary ergodic input processes.

We now state the properties of $C(\beta)$ [22]. We first define respectively β_{min}, $\beta_{max}^{(n)}$ and β_{max} by

$$\beta_{min} = \min_{x\in\mathcal{X}} b(x),$$

$$\beta_{max}^{(n)} = \min\left\{\frac{1}{n}E\left[b(X^n)\right] : \frac{1}{n}I(X^n;Y^n) = C_n\right\},$$

and

$$\beta_{max} \triangleq \beta_{max}^{(\infty)} = \min\left\{\lim_{n\to\infty}\frac{1}{n}E\left[b(X^n)\right] : \lim_{n\to\infty}\frac{1}{n}I(X^n;Y^n) = C\right\}.$$

From the definition of β_{min} above, we can see that $\frac{1}{n}E\left[b(X^n)\right] \geq \beta_{min}$ and so $C(\beta)$ is defined only for $\beta \geq \beta_{min}$.

Remark: For a discrete binary channel with additive noise $\{Z_n\}$ and power cost constraints on the input – i.e. $b(x) = x^2$ – we get that $\beta_{min} = 0$, $\beta_{max} = 1/2$, $C(\beta_{min}) = 0$ and $C(\beta_{max}) = C = 1 - H(Z_\infty)$ where $H(Z_\infty)$ is entropy rate of the noise.

Lemma 8 The n'th capacity-cost function $C_n(\beta)$ defined in (4.2) is *concave* and *strictly increasing* in β for $\beta_{min} \leq \beta < \beta_{max}^{(n)}$ and is equal to C_n for $\beta \geq \beta_{max}^{(n)}$. Therefore, the capacity-cost function $C(\beta)$ given by (4.1) is *concave* and *strictly increasing* in β for $\beta_{min} \leq \beta < \beta_{max}$, and is equal to C for $\beta \geq \beta_{max}$.

Observation: We close this section by reflecting upon the words of Shannon concerning the *duality* between a source and a channel [31]: " There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. This duality is *enhanced* if we consider channels in which there is a *cost* associated with the different input letters, and it is desired to find the capacity subject to the constraint that the expected cost not exceed a certain quantity [...] The solution of this problem leads to a capacity cost function for the channel [...], this function is concave downward [...] In a somewhat dual way, evaluating the rate distortion function $R(D)$ for a source amounts, mathematically, to minimizing a mutual information [...] with a linear inequality constraint [...] The solution leads to a function $R(D)$ which is convex downward [...] This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus, we may have knowledge of the past but cannot control it; we may control the future but not have knowledge of it."

4.3 Discrete Channels with Binary Additive Markov Noise

4.3.1 Capacity-Cost Function with no Feedback

We consider a discrete channel with memory, with common input, noise and output binary alphabet and described by the following equation: $Y_n = X_n \oplus Z_n$, for $n = 1, 2, 3, \ldots$ where:

- $\oplus$ represents the addition operation modulo 2.
- The random variables X_n, Z_n and Y_n are respectively the input, noise and output of the channel.
- $\{X_n\} \perp \{Z_n\}$, i.e. the input and noise sequences are independent from each other.
- The noise process $\{Z_n\}_{n=1}^{\infty}$ is stationary and mixing.

We now turn to the analysis of the capacity-cost function (with no feedback) of this channel. We denote it by $C_{NFB}(\beta)$. The channel is non-symmetric for $\beta < \beta_{max} = 1/2$. Thus the formula of $C_{NFB}(\beta)$ given by (4.1) will not have a closed form. We will then try to derive an upper bound to $C_{NFB}(\beta)$.

In [36], Wyner and Ziv derived a lower bound to the (operational) rate-distortion function ($R(D)$) of stationary ergodic sources:

$$R(D) \geq R_1(D) - \mu_1,$$

where

- $R_1(D)$ is the rate-distortion function of an "associated" memoryless source with distribution equal to the marginal distribution $P^{(1)}(\cdot)$ of the ergodic source.

- $\mu_1 \stackrel{\triangle}{=} H(X_1) - h(X_\infty)$, is the amount of memory in the source. $H(X_1)$ is the entropy of the associated memoryless source with distribution $P^{(1)}(\cdot)$ and $h(X_\infty)$ is the entropy rate of the original ergodic source.

This lower bound was later tightened by Berger [3]:

$$R(D) \geq R_n(D) - \mu_n \geq R_1(D) - \mu_1, \tag{4.5}$$

where $R_n(D)$ is the n'th rate-distortion function of the source, $R_1(D)$ is as defined above and $\mu_n = \frac{1}{n}H(X^n) - h(X_\infty)$.

In light of the striking duality that exists between $R(D)$ and $C(\beta)$, as remarked previously by Shannon [31], we derive an equivalent upper bound to the nonfeedback capacity-cost function of a discrete additive channel.

Proposition 13 Consider a discrete channel with additive stationary noise $\{Z_n\}$. Let $W^{(n)}(\cdot)$ denote the n-fold probability distribution function of the noise process. Then for $N = kn$, $k, n = 1, 2, \ldots$,

$$C_N(\beta) \leq C_n(\beta) + \Delta_{nN} \leq C_1(\beta) + \Delta_{1N}, \tag{4.6}$$

where

- $C_n(\beta)$ is the n-fold capacity-cost function of the channel as defined in (4.2).

- $C_1(\beta)$ is the capacity-cost function of the associated discrete memoryless channel (DMC) with iid additive noise process whose distribution is equal to the marginal distribution $W^{(1)}(\cdot)$ of the ergodic noise process.

- $\Delta_{nN} \triangleq (1/n)H(Z^n) - (1/N)H(Z^N)$ with $Z^i = (Z_1, Z_2, \ldots, Z_i)$, $i = n$ or N, and $\Delta_{1N} = H(Z_1) - (1/N)H(Z^N)$, where $H(Z_1)$ is the entropy of the iid noise process of the associated DMC.

Proof 13 The proof uses a dual generalization of Wyner and Ziv's proof of the lower bound to the rate-distortion function. We first need to use the following expression

$$I(X^N; Y^N) \leq \sum_{i=1}^{k} I(X_{(i)}^n; Y_{(i)}^n) + N\Delta_{nN}, \tag{4.7}$$

where $X^N = (X_{(1)}^n, X_{(2)}^n, \ldots, X_{(k)}^n)$ and $Y^N = (Y_{(1)}^n, Y_{(2)}^n, \ldots, Y_{(k)}^n)$ with

$$X_{(i)}^n = (X_{1,(i)}, X_{2,(i)}, \ldots, X_{n,(i)}),$$

and

$$Y_{(i)}^n = (Y_{1,(i)}, Y_{2,(i)}, \ldots, Y_{n,(i)}).$$

Proving the above inequality goes as follows:

$$\begin{aligned}
\sum_{i=1}^{k} I(X_{(i)}^n; Y_{(i)}^n) + N\Delta_{nN} - I(X^N; Y^N) = & \sum_{i=1}^{k} \left[H(Y_{(i)}^n) - H(Y_{(i)}^n|X_{(i)}^n)\right] + \frac{N}{n}H(Z^n) \\
& -H(Z^N) - H(Y^N) + H(Y^N|X^N) \\
= & \sum_{i=1}^{k} \left[H(Y_{(i)}^n) - H(Z_{(i)}^n)\right] + kH(Z_n) \\
& -H(Z^N) - H(Y^N) + H(Z^N) \\
= & \sum_{i=1}^{k} H(Y_{(i)}^n) - H(Y^N) \\
= & \sum_{i=1}^{k} H(Y_{(i)}^n) - \sum_{i=1}^{k} H(Y_{(i)}^n|Y_{(i-1)}^n, \ldots, Y_{(1)}^n) \\
\geq & \sum_{i=1}^{k} H(Y_{(i)}^n) - \sum_{i=1}^{k} H(Y_{(i)}^n) \\
= & \ 0,
\end{aligned}$$

where the third equality follows from the stationarity of the noise and the last inequality follows from the fact that conditioning decreases entropy.

We now proceed to prove equation (4.6). Let $P^{(N)}(X^N) \in \tau_N(\beta)$ where $\tau_N(\beta)$ is described in Definition 1. For this input distribution, we denote $\beta_i \stackrel{\Delta}{=} \frac{1}{n} E\left[b(X_{(i)}^n)\right]$ for $i = 1, 2, \ldots, k$; thus $(1/k)\sum_{i=1}^k \beta_i \leq \beta$. By (4.7), we obtain with this $P^{(N)}(X^N)$:

$$\frac{1}{N} I(X^N; Y^N) \leq \frac{1}{N} \sum_{i=1}^{k} I(X_{(i)}^n; Y_{(i)}^n) + \Delta_{nN};$$

but $\frac{1}{n} I(X_{(i)}^n; Y_{(i)}^n) \leq C_n(\beta_i)$ for $i = 1, 2, \ldots, k$. Thus

$$\frac{1}{N} I(X^N; Y^N) \leq \frac{1}{k} \sum_{i=1}^{k} C_n(\beta_i) + \Delta_{nN}.$$

By concavity of $C_n(\cdot)$, we have $(1/k)\sum_{i=1}^k C_n(\beta_i) \leq C_n\left((1/k)\sum_{i=1}^k \beta_i\right)$ and since $C_n(\cdot)$ is strictly increasing we have that $C_n\left((1/k)\sum_{i=1}^k \beta_i\right) \leq C_n(\beta)$. Therefore

$$\frac{1}{N} I(X^N; Y^N) \leq C_n(\beta) + \Delta_{nN},$$

or

$$\max_{P^{(N)}(X^N) \in \tau_N(\beta)} \frac{1}{N} I(X^N; Y^N) = C_N(\beta) \leq C_n(\beta) + \Delta_{nN}.$$

Thus the first inequality in (4.6) is proved. To prove the second inequality in (4.6), we need to show that $C_n(\beta) \leq C_1(\beta) + \Delta_{1n}$, or $C_k(\beta) \leq C_1(\beta) + \Delta_{1k}$. This is shown using the first inequality in (4.6) and letting $n = 1$. ■

Using (4.6) and (4.1), we obtain the following tight upper bound on $C_{NFB}(\beta)$.

Corollary Consider the channel described in Proposition 1, with the assumption that the noise process is stationary mixing. Then

$$C_{NFB}(\beta) \leq C_n(\beta) + M_n \leq C_1(\beta) + M_1, \tag{4.8}$$

where

- $C_n(\beta)$ and $C_1(\beta)$ are as defined in Proposition 13.

- $M_n \triangleq \Delta_{n\infty} = (1/n)H(Z^n) - h(Z_\infty)$, and $M_1 \triangleq \Delta_{1\infty} = H(Z_1) - h(Z_\infty)$ is the amount of memory in the noise process.

The bound given above is asymptotically tight with n.

A Lower Bound to $C_{NFB}(\beta)$:

If we take the inputs to be iid, we can apply Mrs. Gerber's Lemma in [29] to obtain a lower bound on $C_{NFB}(\beta)$. Let $P(X_i = 1) \triangleq \alpha$ be the marginal distribution of an iid input process such that $E[b(X_i)] = \beta$, then

$$C_{NFB}(\beta) \geq h_b\left(\tilde{\alpha} * h_b^{-1}(\lambda)\right) - h(Z_\infty),$$

where

$$a * b \triangleq a(1-b) + (1-a)b, \qquad \tilde{\alpha} \triangleq \min\{\alpha, 1-\alpha\},$$

$$\lambda \triangleq \min\{h(Z_\infty), 1 - h(Z_\infty)\},$$

and $h_b(\cdot)$ is binary entropy function.

4.3.2 Capacity-Cost Function with Feedback

We now consider the binary additive channel with output feedback. By this we mean that there exists a "return channel" from the receiver to the transmitter; we assume this return channel is noiseless, delayless, and has large capacity. The receiver uses the return channel to inform the transmitter what letters were actually received; these letters are received at the transmitter before the next letter is transmitted, and therefore can be used in choosing the next transmitted letter.

We look at a particular class of additive noise channels; the class of channels with additive stationary mixing homogeneous Markov noise of order M. We assume that the Markov noise process is generated by the finite-memory contagion

urn scheme derived in Chapter 2. Its transition probability is

$$P(Z_n = 1|Z_1 = e_1, \ldots, Z_{n-1} = e_{n-1}) = \frac{\rho + (e_{n-1} + e_{n-2} + \cdots + e_{n-M})\delta}{1 + M\delta}$$
$$= P(Z_n = 1|Z_{n-M} = e_{n-M}, \ldots, Z_{n-1} = e_{n-1}),$$

where $e_i = 0$ or 1, for $i = 1, 2, \ldots, n-1$ and where $n \geq M+1$. We assume that $\rho < 1/2$ and $\delta > 0$. Note that if $\delta = 0$, the noise process becomes *iid.*

For blocklength $n \leq M$, the block probability of the process is

$$\begin{aligned} Pr(Z^n = e^n) &= \frac{\rho(\rho+\delta)\cdots(\rho+(d-1)\delta)\sigma(\sigma+\delta)\cdots(\sigma+(n-d-1)\delta)}{(1+\delta)(1+2\delta)\cdots(1+(n-1)\delta)} \\ &= \frac{\Gamma(\frac{1}{\delta})\ \Gamma(\frac{\rho}{\delta}+d)\ \Gamma(\frac{\sigma}{\delta}+n-d)}{\Gamma(\frac{\rho}{\delta})\ \Gamma(\frac{\sigma}{\delta})\ \Gamma(\frac{1}{\delta}+n)}, \end{aligned}$$

where $\sigma = 1 - \rho$, $d = e_1 + e_2 + \cdots + e_n$ and $\Gamma(\cdot)$ is the gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ for $x > 0$.

For $n \geq M+1$, we have

$$\begin{aligned} Pr(Z^n = e^n) &= \prod_{i=1}^{n} P(Z_i = e_i \mid Z_{i-1} = e_{i-1}, \ldots, Z_{i-M} = e_{i-M}) \\ &= L \prod_{i=M+1}^{n} \left[\frac{\rho + s_{i-1}\delta}{1+M\delta}\right]^{e_i} \left[\frac{\sigma + (M - s_{i-1})\delta}{1+M\delta}\right]^{1-e_i}, \end{aligned}$$

where

$$L = \frac{\prod_{i=0}^{k-1}(\rho + i\delta)\prod_{j=0}^{M-1-k}(\sigma + j\delta)}{\prod_{\ell=1}^{M-1}(1+\ell\delta)},$$
$$k = e_1 + \cdots + e_M,$$

and

$$s_{i-1} = e_{i-1} + \cdots + e_{i-M}.$$

Remark: For $M = 1$, the Markov process has a marginal distribution given by

$$W^{(1)}(Z_n = 1) = \rho = 1 - W^{(1)}(Z_n = 0),$$

and a transition probability matrix given by

$$Q = [p_{ij}] = \begin{pmatrix} \frac{\sigma+\delta}{1+\delta} & \frac{\rho}{1+\delta} \\ \frac{\sigma}{1+\delta} & \frac{\rho+\delta}{1+\delta} \end{pmatrix},$$

where $p_{ij} \triangleq Pr(Z_n = j | Z_{n-1} = i)$. We note that the transition matrix of this Markov model (for $M = 1$) is *general*; it can represent all binary 1'st order Markov chains with positive[2] transition matrix.

As in the previous chapter, we define a feedback code with blocklength n and rate R as a sequence of encoders

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

for $i = 1, 2, \ldots, n$, along with a decoding function

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

To convey message $W \in \{1, 2, \ldots, 2^{nR}\}$, the user sends the codeword $X^n = (X_1, X_2, \ldots, X_n)$ where $X_i = f_i(W, Y_1, Y_2, \ldots, Y_{i-1})$; $i = 1, 2, \ldots, n$. The decoder receives $Y^n = (Y_1, Y_2, \ldots, Y_n)$ and estimates the message to be $g(Y^n)$. A decoder error occurs if $g(Y^n) \neq W$. We assume that W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

The probability of decoding error is thus given by:

$$\begin{aligned} P_e^{(n)} &= \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} Pr\{g(Y^n) \neq W | W = k\} \\ &= Pr\{g(Y^n) \neq W | W = k\}. \end{aligned}$$

Since we are studying the capacity-cost function, we require an average cost constraint on the input code letters. We say that a feedback rate R is *achievable* if

[2]* A positive transition matrix is a matrix whose entries are all strictly positive.

there exists a sequence of β-admissible (as defined in Definition 2) feedback codes with blocklength n and rate R such that

$$\lim_{n\to\infty} P_e^{(n)} = 0.$$

The supremum of all achievable feedback code rates is then the capacity-cost function with feedback of the channel that we denote by $C_{FB}(\beta)$.

The channel being additive, we have $Y_i = X_i \oplus Z_i$ where $\{Z_i\}$ is the binary Markov noise process described earlier. Note, however, that because of the feedback, X^n and Z^n are no longer independent; X_i may depend recursively on Z^{i-1}.

Finding a closed form for $C_{FB}(\beta)$ is a tedious if not a formidable task. We will then derive lower and upper bounds to $C_{FB}(\beta)$.

A. MAP Estimation of Current Noise Letter at the Encoder

With output feedback, the encoder is informed at time instant i about all the previously received output symbols $Y_1, Y_2, \ldots, Y_{i-1}$; and thus knows all the previous channel noise samples $Z_1, Z_2, \ldots, Z_{i-1}$, $i = 1, 2, \ldots$ (we assume that $Z_0 = 0$ almost surely). Given this information, the encoder can then estimate the current noise sample Z_i in the channel. Let $\hat{Z}_i$ represent the maximum a posteriori estimate (MAP) of Z_i given $(Z_1, Z_2, \ldots, Z_{i-1})$. Since the noise process is Markovian of order M, the encoder needs only the last M noise samples. The MAP estimate $\hat{Z}_i$ for $i > M$ is [3]:

$$\hat{Z}_i = \begin{cases} 1, & \text{if } Pr(Z_i = 1 | Z_{i-1} = e_{i-1}, Z_{i-2} = e_{i-2}, \ldots, Z_{i-M} = e_{i-M}) > 1/2; \\ 0, & \text{if } Pr(Z_i = 0 | Z_{i-1} = e_{i-1}, Z_{i-2} = e_{i-2}, \ldots, Z_{i-M} = e_{i-M}) > 1/2; \end{cases}$$

where e_j is 0 or 1 for $j = i - M, i - M + 1, \ldots, i - 1$.

[3] for $i \leq M$, we get that $\hat{Z}_i$ is the same as below, with the exception of replacing M by $i - 1$.

or

$$\hat{Z}_i = \begin{cases} 1, & \text{if } e_{i-1} + e_{i-2} + \ldots + e_{i-M} > \frac{1-2\rho+M\delta}{2\delta}; \\ 0, & \text{otherwise.} \end{cases}$$

To guarantee that the MAP estimate $\hat{Z}_i$ is not a degenerate constant (zero), we need to require that $(1 - 2\rho + M\delta)/2\delta < M$; i.e.,

$$M\delta > 1 - 2\rho. \tag{4.9}$$

Thus

$$\hat{Z}_i = f(Z_{i-1}, Z_{i-2}, \ldots, Z_{i-M}) \triangleq 1\left\{Z_{i-1} + Z_{i-2} + \ldots + Z_{i-M} > \frac{1 - 2\rho + M\delta}{2\delta}\right\};$$

where $1\{\cdot\}$ is the indicator function. For $M = 1$, we obtain that the MAP estimate of the current noise sample is nonless but the *previous* noise sample:

$$\hat{Z}_i = Z_{i-1}. \tag{4.10}$$

B. A Lower Bound to $C_{FB}(\beta)$

We obtain a lower bound to the capacity-cost function of our binary Markov channel with feedback. We first state the following definition [8, 9].

Definition 10 Let (V^n, Y^n) be jointly distributed with distribution $p(v^n, y^n)$, where $V^n = (V_1, V_2, \ldots, V_n)$ and $Y^n = (Y_1, Y_2, \ldots, Y_n)$. Let $\epsilon > 0$, then the set $A_\epsilon^{(n)}$ of *jointly ϵ-typical* (V^n, Y^n) sequences is defined by

$$\begin{aligned} A_\epsilon^{(n)} \;=\; & \left\{(v^n, y^n) \in \mathcal{V}^n \times \mathcal{Y}^n : \left|-\frac{1}{n}\log_2 p(v^n) - \frac{1}{n}H(V^n)\right| < \epsilon, \right. \\ & \left. \left|-\frac{1}{n}\log_2 p(y^n) - \frac{1}{n}H(Y^n)\right| < \epsilon, \left|-\frac{1}{n}\log_2 p(v^n, y^n) - \frac{1}{n}H(V^n, Y^n)\right| < \epsilon\right\} \end{aligned}$$

where $p(v^n)$ and $p(y^n)$ are the respective n-fold marginal distributions derived from $p(v^n, y^n)$.

We now propose the following feedback encoding strategy. Since the encoder knows the previous noise samples, he can then find the MAP estimate $\hat{Z}_i$ of the current noise sample. Thus the encoder can try to fight the noise Z_i in the channel by adding the MAP estimate $\hat{Z}_i$ to the unencoded input letters. That is if $V^n(W)$ is a n-tuple vector representing message $W \in \{1, 2, \ldots, 2^{nR}\}$, then the encoder sends the feedback codeword $X^n(W) = (X_1, X_2, \ldots, X_n)$ where $X_i = V_i \oplus \hat{Z}_i$; $i = 1, 2, \ldots, n$. The output of the channel is then $Y_i = X_i \oplus Z_i = V_i \oplus U_i$ where

$$U_i \triangleq Z_i \oplus \hat{Z}_i. \tag{4.11}$$

Note that since the noise process $\{Z_i\}$ is stationary mixing and U_i is a *time-invariant* function of $\{Z_i\}$, the process $\{U_i\}$ is also stationary and mixing [25].

Realize that the encoder transmits the feedback codeword X^n and thus the β-admissibility condition should be satisfied by the components of X^n. We define $C^{lb}(\beta)$ by

$$C^{lb}(\beta) = \sup_n \; C_n^{lb}(\beta) = \lim_{n\to\infty} \; C_n^{lb}(\beta), \tag{4.12}$$

where

$$C_n^{lb}(\beta) = \max_{P^{(n)}(V^n)\in\tilde{\tau}_n(\beta)} \; \frac{1}{n} \; I(V^n; Y^n), \tag{4.13}$$

where

$$\tilde{\tau}_n(\beta) = \left\{ P^{(n)}(V^n) : \frac{1}{n} \sum_{i=1}^{n} E\left[b(V_i \oplus \hat{Z}_i)\right] \leq \beta \right\}, \tag{4.14}$$

and $Y_i = V_i \oplus U_i$ where the process $\{U_i\}$ is defined by (11).

Remark: We can easily observe from expressions (4.12-4.14) that $C^{lb}(\beta)$ is nonless but the *nonfeedback* capacity-cost function of a binary channel with additive

stationary ergodic noise given by $\{U_i\}$, with the particularity that the average cost requirement is not imposed on the input letters V_i but on the letters obtained from $V_i \oplus \hat{Z}_i$.

This leads us to the following result:

Proposition 14 (Achievability of $C^{lb}(\beta)$) Consider the binary Markov additive-noise channel (described above) with feedback. Then for $R < C^{lb}(\beta)$ there exists a sequence of β-admissible feedback codes (i.e., satisfying the average cost constraint) with blocklength n and rate R such that $P_e^{(n)} \longrightarrow 0$ as $n \longrightarrow \infty$.

Proof 14 The proof uses random coding. Let $V^n(1), V^n(2), \ldots, V^n(2^{nR})$ be independent identically distributed n-vectors, each drawn according to a distribution $p(v^n)$ that achieves $C^{lb}(\beta)$, where $p(v^n)$ is a n-fold distribution of a stationary ergodic process.

Transmission: To send message W, where W is uniform over $\{1, 2, \ldots, 2^{nR}\}$, the transmitter sends $X^n(W) = (X_1, X_2, \ldots, X_n)$ where $X_i = V_i \oplus \hat{Z}_i$; $i = 1, 2, \ldots, n$. Note that $\{V_i\} \perp \{Z_i\}$ and thus $\{V_i\} \perp \{U_i\}$. We assume that $M\delta > 1 - 2\rho$ in order to have a non-trivial $\hat{Z}_i$. Note that the feedback codewords $X^n(W)$ are governed by a stationary ergodic distribution since $V^n(W)$ are selected according to a stationary ergodic $p(v^n)$ (that achieves $C^{lb}(\beta)$) and $\{Z_i\}$ is stationary mixing [16].

Decoding: The receiver receives $Y^n = (Y_1, Y_2, \ldots, Y_n)$ where

$$Y_i = X_i \oplus Z_i = V_i \oplus \hat{Z}_i \oplus Z_i = V_i \oplus U_i; \qquad i = 1, 2, \ldots, n.$$

The receiver declares $\hat{W} \in \{1, 2, \ldots, 2^{nR}\}$ was sent if $\left(V^n(\hat{W}), Y^n\right)$ is the only jointly ϵ-typical pair.

Error: An error is made if there is no ϵ-typical $\left(V^n(\hat{W}), Y^n\right)$ pair, more than one such pair or $\hat{W} \neq W$. Furthermore, an error is made if the β-admissibility constraint is violated.

We investigate the probability of error $P_e^{(n)}$. We can assume without loss of generality that $W = 1$ was sent. We define the events

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} b(X_i(1)) > \beta \right\},$$

and

$$E_i = \left\{ (V^n(i), Y^n) \in A_\epsilon^{(n)} \right\}; \quad i = 1, 2, \ldots, 2^{nR}.$$

Let E_i^c be the complement of E_i. Then

$$\begin{aligned} P_e^{(n)} &= Pr(\hat{W} \neq 1 | W = 1) \\ &= Pr\left(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}} | W = 1\right) \\ &\leq Pr(E_0) + Pr(E_1^c | W = 1) + \sum_{i=2}^{2^{nR}} Pr(E_i | W = 1) \\ &= Pr(E_0) + Pr(E_1^c | W = 1) + 2^{nR} Pr(E_2 | W = 1). \end{aligned} \tag{4.15}$$

Since the codewords are drawn according to a stationary ergodic distribution that achieves $C^{lb}(\beta)$, then by the law of large numbers we get that $Pr(E_0) \leq \epsilon$ for n sufficiently large. We now state the following lemma [9, 22].

Lemma 9 (Joint AEP) Let $\left|A_\epsilon^{(n)}\right|$ denote the cardinality of the set $A_\epsilon^{(n)}$.

Then

$$\left|A_\epsilon^{(n)}\right| \leq 2^{n\left[\frac{1}{n}H(V^n, Y^n) + \epsilon\right]}. \tag{4.16}$$

Furthermore, if the joint process $\{(V_n, Y_n)\}_{n=1}^{\infty}$ is stationary ergodic, then there exists n_0, such that for all $n > n_0$,

$$Pr\left((V^n, Y^n) \in A_\epsilon^{(n)}\right) \geq 1 - \epsilon. \tag{4.17}$$

Now, consider $V^n(1)$ as defined above and $Y^n = (Y_1, Y_2, \ldots, Y_n)$ such that $Y_i = V_i(1) \oplus U_i$, where $\{V_i\} \perp \{U_i\}$, $i = 1, 2, \ldots, n$, and the additive noise process $\{U_i\}$ is stationary mixing. Therefore the joint process $\{(V_n(1), Y^n)\}$ is stationary ergodic [16, 25]. By (4.17) in the above lemma, we thus get that $Pr(E_1^c|W = 1) = Pr\left((V^n(1), Y^n) \notin A_\epsilon^{(n)}\right) \leq \epsilon$ for n sufficiently large.

Now,

$$Pr(E_2|W = 1) = \sum_{(v^n, y^n) \in A_\epsilon^{(n)}} p(v^n)p(y^n),$$

since by our code generation $V^n(1)$ and $V^n(i)$ are independent for $i \neq 1$; thus Y^n and $V^n(i)$ are independent, $i \neq 1$.

$$\begin{aligned} Pr(E_2|W = 1) &\leq \sum_{(v^n, y^n) \in A_\epsilon^{(n)}} 2^{-n\left[\frac{1}{n}H(V^n) - \epsilon\right]} \, 2^{-n\left[\frac{1}{n}H(Y^n) - \epsilon\right]} \\ &= \left|A_\epsilon^{(n)}\right| \, 2^{n\left[-\frac{1}{n}H(V^n) - \frac{1}{n}H(Y^n) + 2\epsilon\right]} \\ &\leq 2^{n\left[-\frac{1}{n}H(V^n) - \frac{1}{n}H(Y^n) + \frac{1}{n}H(V^n, Y^n) + 3\epsilon\right]} \\ &= 2^{-n\left[\frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|V^n) + 3\epsilon\right]} \\ &\leq 2^{-n\left[C^{lb}(\beta) + 3\epsilon\right]}, \end{aligned}$$

where the first inequality follows from Definition 10 and the second inequality above follows from (4.16).

Therefore (4.15) yields

$$P_e^{(n)} \leq 2\epsilon + 2^{-n\left[C^{lb}(\beta) - R + 3\epsilon\right]} \leq 3\epsilon$$

if n is sufficiently large and $R < C^{lb}(\beta)$; and thus $C^{lb}(\beta)$ is achievable. ■

Observation: Realize that the main idea in proving the achievability of $C^{lb}(\beta)$ consists of transforming the problem of our original additive Markov channel with feedback into the problem of a new additive channel with *no feedback*.

C. An Upper Bound to $C_{FB}(\beta)$

We now derive an upper bound to the capacity-cost function with feedback.

Proposition 15 Any β-admissible sequence of feedback codes with blocklength n and rate R such that $P_e^{(n)} \longrightarrow 0$ as $n \longrightarrow \infty$, must have $R \leq C^{ub}(\beta)$, where

$$C^{ub}(\beta) = \lim_{n\to\infty} \max_{P^{(n)}(X^n)\in\tau_n(\beta)} \frac{1}{n} I(W, Y^n), \tag{4.18}$$

where the maximization is taken over all β-admissible feedback codewords X^n of the form $X_i = X_i(W, Y^{i-1})$ for $i = 1, 2, \ldots, n$. $W \in \{1, 2, \ldots, 2^{nR}\}$ is the message, and the output is described by $Y_i = X_i(W, Y^{i-1}) \oplus Z_i$; $i = 1, 2, \ldots, n$.

Proof 15 Let R be the rate of any β-admissible feedback code with blocklength n and vanishing probability of error (i.e., $P_e^{(n)} \longrightarrow 0$ as $n \longrightarrow \infty$).

Since W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$, we have that $H(W) = nR$. Furthermore, $H(W) = H(W|Y^n) + I(W; Y^n)$. Now by Fano's inequality,

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

since $h_b(P_e^{(n)}) \leq 1$, where $h_b(\cdot)$ is the binary entropy function.

We then obtain

$$\begin{aligned} nR &= H(W) \\ &= H(W|Y^n) + I(W; Y^n) \\ &\leq 1 + P_e^{(n)} nR + I(W; Y^n) \\ &\leq 1 + P_e^{(n)} nR + \max_{P^{(n)}(X^n)\in\tau_n(\beta)} \frac{1}{n} I(W, Y^n). \end{aligned}$$

Dividing both sides by n and taking n to infinity, we get:

$$R \leq \lim_{n\to\infty} \max_{P^{(n)}(X^n)\in\tau_n(\beta)} \frac{1}{n} I(W, Y^n) = C^{ub}(\beta).$$

■

Observation: Note that since the channel is additive, we can show that $I(W, Y^n) = H(Y^n) - H(Z^n)$ (cf. Chapter 3). By comparing Propositions 14 and 15, we can assert that $C_{FB}(\beta) = C^{lb}(\beta) = C^{ub}(\beta)$ if it can be shown that the feedback encoding scheme we propose in Proposition 14 is indeed the optimal encoding scheme; a task we do not address at the present.

4.4 Numerical Results

We have thus far shown derived an upper bound on the capacity-cost function with no feedback:

$$C_{NFB}(\beta) \leq C_n(\beta) + M_n.$$

This bound becomes tight as $n \longrightarrow \infty$. Furthermore we have derived a lower and an upper bound to the capacity-cost function with feedback

$$C^{lb}(\beta) \leq C_{FB}(\beta) \leq C^{ub}(\beta),$$

where $C^{lb}(\beta)$ and $C^{ub}(\beta)$ are given respectively by (4.12) and (4.18).

We are interested in determining whether feedback increases the capacity-cost function of our channel. Elucidating this statement would involve comparing $C^{lb}(\beta)$ with $C_n(\beta) + M_n$. But $C^{lb}(\beta) = \sup_n C_n^{lb}(\beta)$ where $C_n^{lb}(\beta)$ is given by (4.13). Thus, in order to claim that feedback helps, it suffices to show that $C_n^{lb}(\beta)$ is greater than $C_n(\beta) + M_n$. We can write

$$C_n^{lb}(\beta) = \max_{P^{(n)}(V^n)\in\tilde{\tau}_n(\beta)} \frac{1}{n} H(\tilde{Y}^n) - \frac{1}{n} H(Z^n),$$

where $\tilde{Y}_i = V_i \oplus Z_i \oplus Z_{i-1}$, $i = 1, 2, \ldots, n$. Similarly we have

$$C_n(\beta) + M_n = \max_{P^{(n)}(V^n) \in \tau_n(\beta)} \frac{1}{n} H(Y^n) - \frac{1}{n} H(Z^n),$$

where $Y_i = V_i \oplus Z_i$, $i = 1, 2, \ldots, n$. In other words we need to compare

$$\max_{P^{(n)}(V^n) \in \tilde{\tau}_n(\beta)} \frac{1}{n} H(\tilde{Y}^n) \quad \text{with} \quad \max_{P^{(n)}(V^n) \in \tau_n(\beta)} \frac{1}{n} H(Y^n).$$

The comparaison above seems difficult analytically, if not impossible. Furthermore, it is very probable that the feedback quantity above is bigger than the nonfeedback quantity only for certain values of the channel parameters δ and ρ (We already know show that for very large δ and $b(x) = x^2$, feedback does not help).

Blahut's Algorithm: We investigate numerically whether $C_n^{lb}(\beta)$ is greater than $C_n(\beta) + M_n$. To do this we use Blahut's algorithm for the computation of the capacity-cost function [6].

We will hereafter assume that the cost function $b(\cdot)$ is given by $b(x) = x^2$ – i.e. we will impose power constraints on the channel input letters. $C_n^{lb}(\beta)$ and $C_n(\beta)$ are in fact the capacity-cost functions of discrete memoryless channels whose input and output alphabets are the sets of words of length n and whose transition probabilities are given by the n-fold probability distributions of the process $\{U_i\}$ and the Markov process $\{Z_i\}$ respectively.

Using the algorithm of Theorem 10 in [6], we calculate $C_8^{lb}(\beta)$ and $C_8(\beta)$ for different values of M, δ and ρ that satisfy condition (4.9). Since the noise is a Markov process, M_n can be expressed analytically in terms of n, δ and ρ using the results in Chapter 2; thus $C_8(\beta) + M_8$ is obtained.

Tighter results can be achieved for $n > 8$; however, the tightness improves

as $1/n$ while the computation complexity increases exponentially. The results, computed to an accuracy of 10^{-4} bits are plotted in Figures 4.1 – 4.4.

The figures[4] indicate to us that output feedback *does increase* the capacity-cost function of our binary Markov additive channel. We define

$$\Delta_8 = C_8^{lb}(\beta) - (C_8(\beta) + M_8) \qquad \text{bits per channel use,}$$

and

$$G_8 = \frac{\Delta_8}{C_8(\beta) + M_8} \times 100.$$

We therefore obtain that feedback increases the capacity-cost function by *at least* Δ_8 bits per channel use, which results in a gain of *at least* G_8 %. Values of Δ_8 and G_8 are computed in Table 1.1 for some values of M, β, ρ and δ. For $M = 2$, $\delta = 0.5$, $\rho = 0.05$ and $\beta = 0.35$, we get a feedback gain of at least 0.0080 bits, yielding an improvement $\geq 1.11\%$. For $M = 2$, $\delta = 0.5$, $\rho = 0.1$ and $\beta = 0.35$, $\Delta_8 = 0.0091$ and $G_8 = 1.59\%$. Note that for our results above we used the power cost function $b(x) = x^2$; thus other numerical results, can be obtained for different cost functions.

Comment: In Chapter 3, we show that feedback does not increase the capacity of additive channels. In that problem, we do not require constraints on the input, and the additive channels are *symmetric*. The symmetry in these channels, maximizes the output entropy rate with no feedback: $\lim_{n\to\infty} \frac{1}{n} \max_{X^n} H(Y^n) = 1$; assuming binary channels. Thus this quantity cannot be increased with feedback; this leads to a nonincrease in capacity since the noise entropy rate is the same with and without feedback. However, if we impose constraints on the input, the output entropy rate with no feedback is < 1; and hence can be increased with feedback.

[4]Note that C_n^{lb} is zero for $\beta \leq P(\hat{Z}_i = 1)$, which is $\rho(\rho+\delta)/(1+\delta)$ for $M = 2$. This is because the average power in X_i cannot fall below the average power in $\hat{Z}_i$.

4.5 Conclusion

In this chapter, we analyzed a binary channel with additive stationary mixing Markov noise of order M. We introduced average cost constraints on the input sequences of the additive channel, rendering it *non-symmetric*. We examined the effect of output feedback on the capacity-cost function of the channel.

A tight upper bound to the capacity-cost function of the channel with no feedback ($C_{NFB}(\beta)$) was established. Furthermore, a lower bound to the capacity-cost function with feedback ($C_{FB}(\beta)$) was obtained. The converse to the coding theorem for the channel with feedback was proven; thus providing an upper bound to $C_{FB}(\beta)$. With the help of Blahut's algorithm for the computation of the capacity-cost function, we showed that $C_{FB}(\beta) > C_{NFB}(\beta)$ for certain channel parameters.

The results of this chapter can easily be extended to q-ary alphabet ($q > 2$) additive Markov channels. One of the questions that remain unsolved, would be to determine whether the upper and lower bounds to $C_{FB}(\beta)$ are equal. If this statement is true, it would imply that the encoding feedback method we proposed in Section 4.3.2, is indeed the *optimal* encoding method for this channel.

Further studies may include the investigation of the effect of feedback on the capacity-cost function of discrete channels with additive stationary ergodic (non-Markovian in general) noise. In that case, if we use the encoding technique of Section 4.3.2, the MAP estimate of the noise, is a time-variant function of the noise process; resulting in a *non-ergodic* new noise process $\{U_i\}$.

Other research directions, may involve the study of the capacity of non-symmetric channels with feedback, like the binary multiplicative channel (the "AND" channel) or the real adder channel. It is conjectured that feedback do cause an increase in capacity for such channels.

M	β	δ	ρ	$C_8(\beta) + M_8$	$C_8^{lb}(\beta)$	Δ_8	*Gain* G_8
2	0.38	1.00	0.05	0.7847	0.7892	0.0045	0.57 %
2	0.40	1.00	0.05	0.7951	0.7999	0.0048	0.60 %
2	0.42	1.00	0.05	0.8080	0.8040	0.0040	0.50 %
2	0.30	0.50	0.05	0.6774	0.6840	0.0066	0.97 %
2	0.35	0.50	0.05	0.7198	0.7278	0.0080	1.11 %
2	0.40	0.50	0.05	0.7495	0.7564	0.0069	0.92 %
2	0.30	0.50	0.10	0.5387	0.5446	0.0059	1.10 %
2	0.35	0.50	0.10	0.5719	0.5810	0.0091	1.59 %
2	0.40	0.50	0.10	0.5953	0.6031	0.0078	1.31%

Table 4.1: Numerical results of feedback lower bound $C_8(\beta)$ and non-feedback upper bound $C_8(\beta) + M_8$ for the M'th order Markov binary channel.

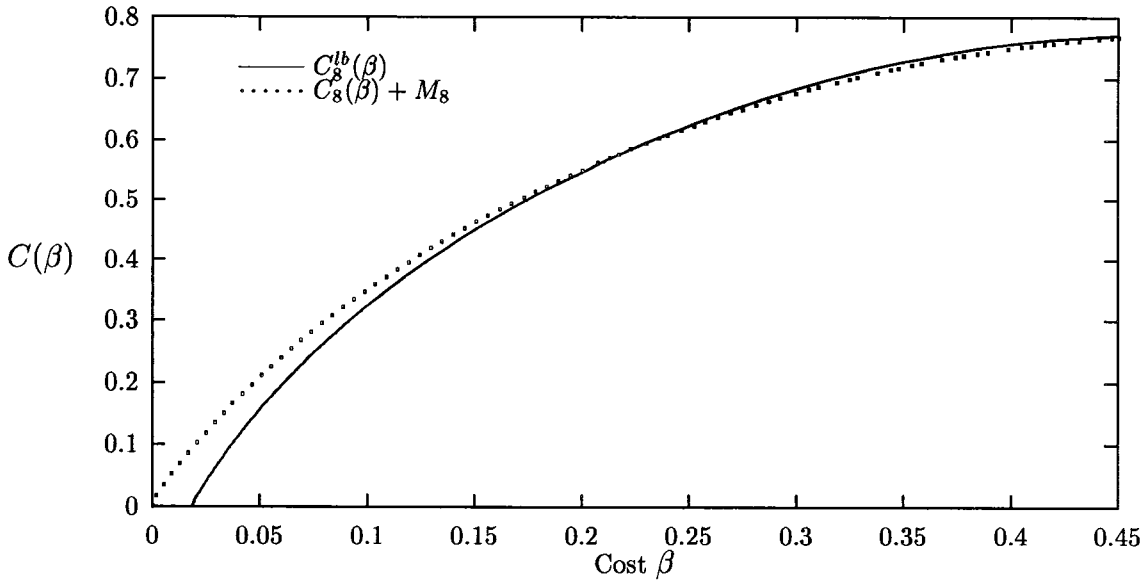

Figure 4.1: Markov binary channel of order $M = 2$ with $\delta = 0.5$ and $\rho = 0.05$.

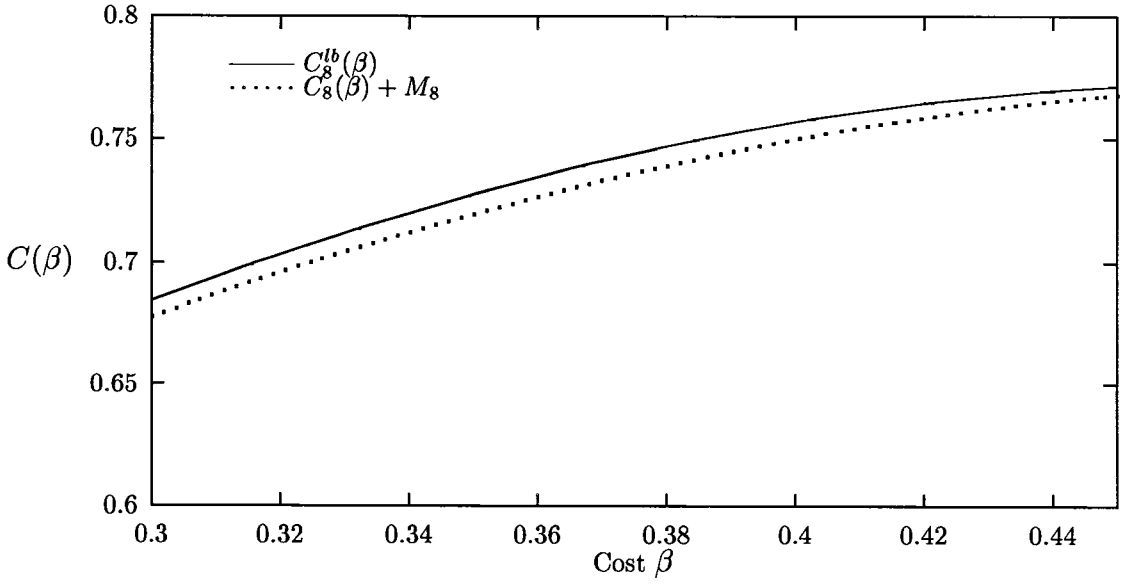

Figure 4.2: Enlargement of Figure 1 for $0.3 \leq \beta \leq 0.45$.

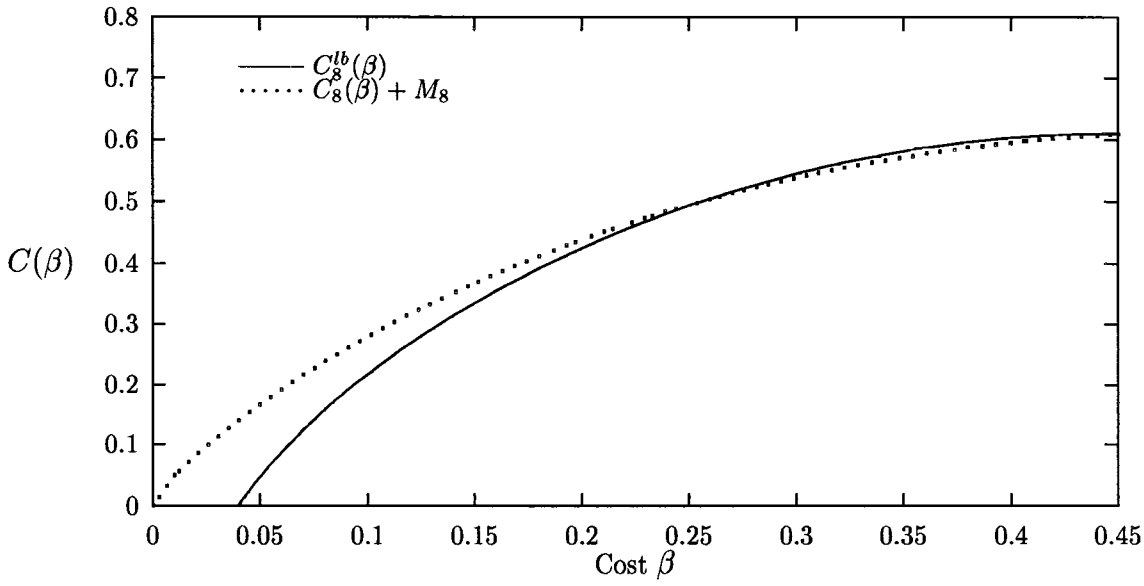

Figure 4.3: Markov binary channel of order $M = 2$ with $\delta = 0.5$ and $\rho = 0.10$.

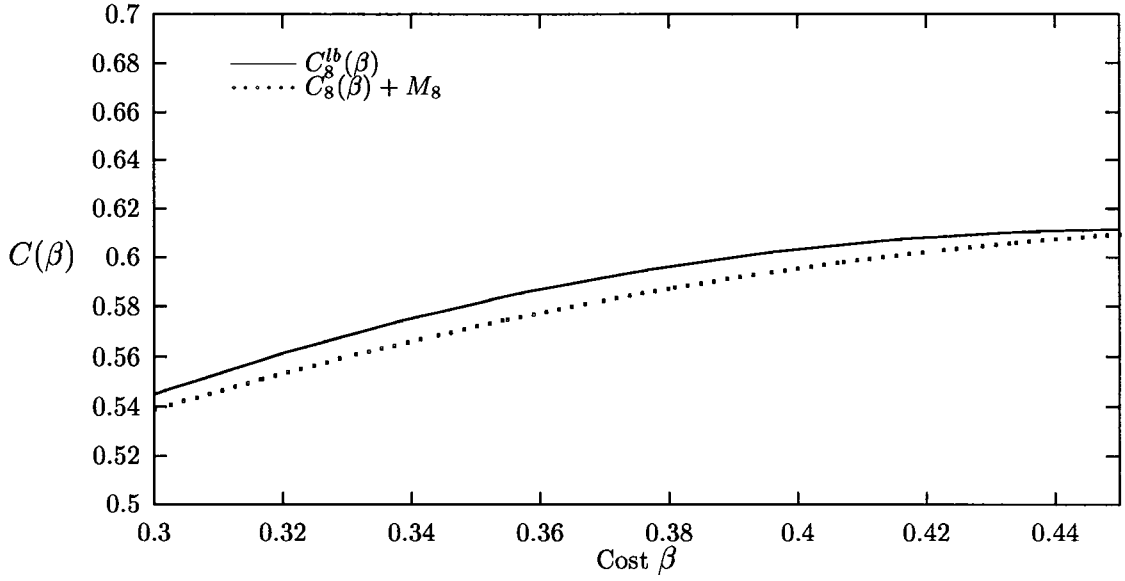

Figure 4.4: Enlargement of Figure 3 for $0.3 \leq \beta \leq 0.45$.

Chapter 5

Conclusions

In this dissertation, we obtained new results on the analysis of discrete communication channels with output memory.

First, we considered a *contagion* communication channel. A contagion channel is a system in which noise propagates in a way similar to the spread of an infectious disease through a population; each "unfavorable" event (i.e., an error) increases the probability of future unfavorable events. Our motivation was twofold: First, it has been shown that defects in silicon have distributions generated by such contagious processes. Second - and more generally - most real-world communication channels, including the digital cellular channel, have *memory* – i.e., the effect of noise lingers over many transmitted symbols. A contagion-based model offers an interesting and less complex alternative to the Gilbert-Elliott burst model and others. The model we set forth is the so-called Polya-contagion channel – a discrete binary communication channel with additive errors modeled according to the famous urn scheme of George Polya for the spread of contagion. The resulting channel is stationary but not ergodic, and it has many interesting properties. A maximum likelihood (ML) decoding algorithm for the channel was derived; it turned out

that that ML decoding is equivalent to decoding a received vector onto *either* the closest codeword *or* the codeword that is farthest away, depending on whether an "apparent epidemic" has occurred. We also showed that this channel is an "averaged" channel in the sense of Ahlswede (and others) and that its capacity is zero. A finite-memory version of this Polya-contagion model was next considered; this channel is an ergodic Markov channel of memory M with a non-zero capacity that increases with increasing memory.

In the second part of this dissertation, we investigated the effect of output feedback on the capacity of stationary additive noise channels with memory. By "output feedback", we assume there exists an ideal "return channel" from the receiver to the transmitter. Intuitively, it is plausible that, with feedback, some techniques could be used at the transmitter to control the noise, hence increasing the channel capacity. However, we arrived at the counter-intuitive result that the capacity of an additive noise channel with feedback does *not* exceed its capacity without feedback. In light of recent results by Verdú and Han on a general formula for channel capacity [35], we then extended the feedback results above to channels with *arbitrary* (not necessarily stationary, etc.) additive noise as well as discrete "symmetric" channels.

In many communication systems, there are constraints on the inputs that appear at the input of the channel; the most common is an average power constraint. We addressed this issue in the last part of the dissertation by incorporating cost constraints on the input of additive noise channels and considering the effect of output feedback on the "capacity cost function" – i.e., the maximum rate at which information can be conveyed subject to the constraint. We demonstrated that when the additive noise is a stationary mixing Markov process, output feedback

can increase the capacity-cost function.

There are several directions in which we can proceed in the future. One direction is to use the Polya contagion channel in modeling "real-world" communication channels, in particular the digital cellular channel. Preliminary results indicate that, for certain parameters, the Polya-contagion channel provides a good "fit" to the digital cellular channel. Furthermore, we intend to address the problem of combined source-channel coding over channels with memory. A combined source-channel coding system is a system in which the source and channel codes are *jointly* designed – as opposed to a tandem source-channel coding system where the source and channel codes are designed *separately.* A major consequence of the work of Claude Shannon is the source-channel separation theorem; it states that in conveying information from one point to another, the source coding and channel coding operations can be separated without sacrificing asymptotic optimality. An important drawback of the tandem coding approach is that optimality *is* asymptotic; it requires the source and channel codes to operate on arbitrarily long blocks which, in implementation, translates to large complexity and delay. An important goal here is to find source/channel coding techniques that maximize performance for a given level of complexity.

Another research direction is to pursue, from an information theory point of view, the investigation of the problem of coding of information bearing signals for transmission over communication channels – i.e., to determine the *fundamental limits* to how efficiently one can encode information and still be able to recover the information with negligible loss. One study may include the derivation of a general capacity formula for arbitrary discrete single-user channels with feedback. This would result in an extension of the work by Verdú and Han on channel capacity

[35]. Other works may involve the study of the capacity of non-symmetric channels with feedback, like the binary multiplicative channel (the "AND" channel) or the real adder channel. It is conjectured that feedback do cause an increase in capacity for such channels.

Bibliography

[1] R. Ahlswede, "The Weak Capacity of Averaged Channels", *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, Vol. 11, pp. 61-73, 1968.

[2] R. Ahlswede, "Certain Results in Coding Theory for Compound Channels I," *Proceedings Bolyai Colloquium on Information Theory*, Debrecen, Hungary, pp. 35-60, 1967.

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

[4] P. Billingsley, *Probability and Measure*, Wiley, NY, 1979.

[5] D. Blackwell, L. Breiman and A. J. Thomasian, "The Capacity of a Class of Channels", *Annals Math. Stat.*, Vol. 30, pp. 1229-1241, 1959.

[6] R. E. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions", *IEEE Transactions on Information Theory*, Vol. 18, No. 4, pp. 460-473, 1972.

[7] B. C. Carlson, *Special Functions of Applied Mathematics*, Academic Press, 1977.

[8] T. M. Cover and S. Pombra, "Gaussian Feedback Capacity", *IEEE Transactions on Information Theory*, Vol. 35, pp. 37-43, 1989.

[9] T. M. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., 1991.

[10] R. L. Dobrushin and M. S. Pinsker, "Memory Increases Transmission Capacity", *Problemy Peredachi Informatsii*, Vol. 5, No. 1, pp. 94-95, 1969.

[11] W. Feller, *An Introduction to Probability Theory and its Applications*, John Wiley & Sons Inc., Second Edition, Vol. 2, 1971.

[12] E. N. Gilbert, "Capacity of Burst-Noise Channels", *Bell Syst. Tech. Journal*, Vol. 39, pp. 1253-1265, 1960.

[13] R. Gray and L. D. Davisson, "The Ergodic Decomposition of Stationary Discrete Random Processes", *IEEE Transactions on Information Theory*, Vol. 20, No. 5, pp. 625-636, 1974.

[14] R. Gray and L. D. Davisson, *Ergodic and Information Theory*, Dowden, Hutchinson & Ross, Inc., 1977.

[15] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, Springer-Verlag New York Inc., 1988.

[16] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag New York Inc., 1990.

[17] R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, Norwell, MA, 1990.

[18] R. M. Gray and D. S. Ornstein, "Block Coding for Discrete Stationary $\overline{d}$-Continuous Noisy Channels", *IEEE Transactions on Information Theory*, Vol. 25, pp. 292-306, 1979.

[19] K. Jacobs, "Almost Periodic Channels", *Colloquium on Combinatorial Methods in Probability Theory, Aarhus"*, pp. 118-126, 1962.

[20] L. Kanal and A. Sastry, "Models for Channels with Memory and Their Applications to Error Control", *Proceedings of the IEEE*, Vol. 66, pp. 724-744, July 1978.

[21] J. C. Kieffer, "A General Formula for the Capacity of Stationary Nonanticipatory Channels", *Information and Control*, Vol. 26, pp. 381-391, 1974.

[22] R. J. McEliece, *The Theory of Information and Coding, A Mathematical Framework for Communication*, Cambridge University Press, 1984.

[23] K. R. Parthasarathy, "Effective Entropy Rate and Transmission of Information Through Channels with Additive Random Noise", *Sankhya*, Vol. A(25), pp. 75-84, 1963.

[24] K. Petersen, *Ergodic Theory*, Cambridge University Press, 1983.

[25] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Francisco, 1964.

[26] G. Polya and F. Eggenberger, "Über die Statistik Verketteter Vorgänge", *Z. Angew. Math. Mech.*, Vol. 3, pp. 279-289, 1923.

[27] G. Polya and F. Eggenberger, "Sur l'Interpretation de Certaines Courbes de Fréquences", *Comptes Rendus C. R.* , Vol. 187, pp. 870-872, 1928.

[28] G. Polya, "Sur Quelques Points de la Théorie des Probabilités", *Ann. Inst. H. Poincarré*, Vol. 1, pp. 117-161, 1931.

[29] S. Shamai (Shitz) and A. D. Wyner, "A Binary Analog to the Entropy-Power Inequality", *IEEE Transactions on Information Theory*, Vol. 36, pp. 1428-1430, 1990.

[30] C. E. Shannon, "The Zero-Error Capacity of a Noisy Channel", *IRE Transactions on Information Theory*, Vol. 2, pp. 8-19, 1956.

[31] C. E. Shannon, "Coding Theorems for a Discrete Source with a Fidelity Criterion", *IRE Nat. Conv. Rec.*, Pt. 4, pp. 142-163, 1959.

[32] C. H. Stapper, A. N. McLaren and M. Dreckmann, "Yield Model for Productivity Optimization of VLSI Memory Chips with Redundancy and Partially Good Product", *IBM J. Res. Develop.*, Vol. 24, No. 3, pp. 398-409, May 1980.

[33] R. Taylor, P. Daffer, R. Patterson, *Limit Theorems for Sums of Exchangeable Random Variables*, Rowman & Allanheld Inc., 1985.

[34] Y. C. Tong, *Probability Inequalities in Multivariate Distributions*, Academic Press, 1980.

[35] S. Verdú and T. S. Han, "A General Formula for Channel Capacity", *IEEE Transactions on Information Theory*, to appear.

[36] A. Wyner and J. Ziv, "Bounds on the Rate-Distortion Function for Stationary Sources with Memory", *IEEE Transactions on Information Theory*, Vol. 17, pp. 508-513, 1971.