THE CAPACITY-COST FUNCTION OF CHANNELS WITH ADDITIVE MARKOV NOISE

by

Nicholas J. Whalen

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements
for the degree of Master of Science (Eng.)

Queen's University
Kingston, Ontario, Canada
June 1998

Acknowledgements

Over the past two years, Dr. Fady Alajaji has been a boundless source of insight into this research topic. His supervision proved essential in moving this thesis forward and his keen abilities as an editor will surely be missed when I next attempt to write a paper. Without his help and guidance this thesis would not have been accomplished.

Also on an academic note I would like to thank Dr. Glen Takahara for his help with Chapter 2, and the other students in the communications lab – most notably Ali Nazer – for their support during my graduate studies.

My long experience with the department of Mathematics and Statistics of Queen's University has been a fulfilling one. I sincerely appreciate Dr. Bob ErDahl's invitation to join the graduate program in Mathematics and Engineering, and Dr. Ron Hirschorn and Dr. Agnes Herzberg's encouragement along the way. I would also like to personally thank Jennifer Read, Johanna Ng and Marg Hogan for dealing with my endless requests and inquiries.

Finally, I would like to thank Dr. Fady Alajaji, Dr. Lorne Campbell, the Department of Mathematics and Statistics, and the School of Graduate Studies for their financial assistance over the past two years.

Abstract

In this work, we analyze the capacity of discrete-time finite-alphabet channels with memory subject to an input cost constraint. More specifically, we consider modulo-q additive noise channels, where the noise process is a stationary Markov source of order k.

We begin by investigating the capacity-cost function ($C\left(\beta\right)$) of such additive noise channels without feedback. Since $C\left(\beta\right)$ does not admit a closed-form expression, we estimate it numerically. This is achieved by implementing two existing bounds to $C\left(\beta\right)$ (one from below and one from above). The lower bound consists of the n^{th} capacity-cost function $C_n\left(\beta\right)$ for any given block length n. The upper bound, which is due to Alajaji, is the counterpart to the Wyner-Ziv lower bound to the rate-distortion function of sources with memory. Both bounds, which are asymptotically tight with increasing block length, are calculated using the Blahut algorithm for the computation of channel capacity. In the case of channels with binary alphabet, we use Mrs. Gerber's Lemma as a means to improve the lower bound on $C\left(\beta\right)$. Numerical examples indicate that the three bounds form a tight envelope on $C\left(\beta\right)$.

We next examine the effect of output feedback on the capacity-cost function of additive Markov noise channels. We establish a lower bound to the capacity-cost function with feedback ($C_{FB}\left(\beta\right)$). We show (both numerically and analytically) that for a particular feedback encoding strategy and a class of Markov noise sources, the lower bound to $C_{FB}(\beta)$ is strictly greater than the upper bound to $C(\beta)$. This demonstrates that feedback can increase the capacity-cost function of discrete channels with memory.

Contents

List of Tables

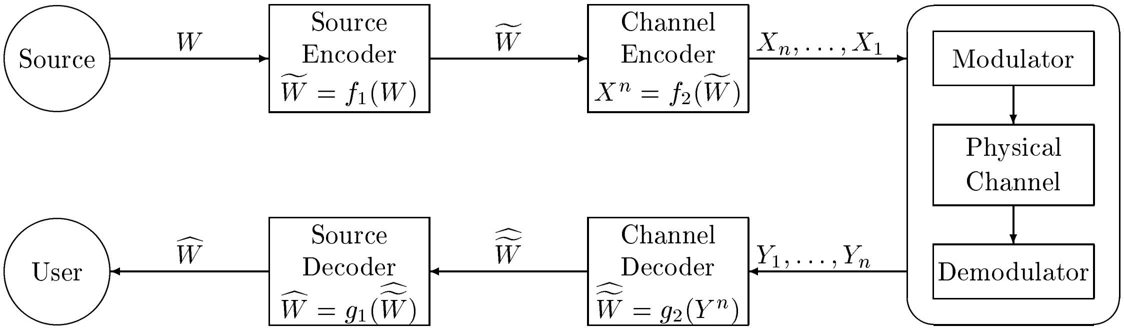
List of Figures

Chapter 1

Introduction

Within the field of information theory, numerous questions remain unanswered with regards to the capacity of constrained channels with feedback and memory. In this discourse, we will examine and implement some established techniques for bounding the constrained capacity, or capacity-cost function – $C(\beta)$ [2, 8, 20]. Then we will use these bounds to prove that feedback can increase the capacity-cost function for some q-ary addition channels for which a Markov chain generates the data corruption.

In this chapter we present the literature review of articles upon which our research is based. We then specify the main contributions of this thesis. Finally, we outline the general flow of the thesis.



Source
W
Source Encoder $\widetilde{W} = f_1(W)$
$\widetilde{W}$
Channel Encoder $X^n = f_2(\widetilde{W})$
$X_n, \ldots, X_1$
Modulator
Physical Channel
Demodulator
$Y_1, \ldots, Y_n$
Channel Decoder $\widehat{\widetilde{W}} = g_2(Y^n)$
$\widehat{\widetilde{W}}$
Source Decoder $\widehat{W} = g_1(\widehat{\widetilde{W}})$
$\widehat{W}$
User

Figure 1.1: Generalized Channel Model without Feedback.

1.1 Literature Review

A number of important results and observations regarding the capacity-cost function and the rate-distortion function direct and guide our present research goals:

- to analyze the capacity-cost function of channels with memory,
- to implement tight bounds on the capacity-cost function for channels with additive Markov noise,
- and to demonstrate that feedback can increase the capacity-cost function of such channels.

Our first aim is aided by the work of Shannon, Wyner, Ziv, Blahut and Alajaji. Shannon first remarked in [24] that 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 functions Shannon describes have no closed form in general, and their specification is limited to bounding techniques. Wyner and Ziv discovered two key bounds. Their lower bound on $R(D)$ ([26]) helps our search indirectly by inspiring a dual upper bound by Alajaji on $C(\beta)$ ([2]). Wyner and Ziv also derived a lower bound known as *Mrs. Gerber's Lemma* [27][1] to the capacity-cost function of binary channels with independent input and noise sequences. Blahut's contribution is an algorithm for the computation of the constrained channel capacity, for a memoryless channel with finite input and output symbol sets [8]. This result was also discovered independently by Arimoto in [4].

[1] In this thesis we actually present a more general proof due to Shamai and Wyner [20] of which Mrs. Gerber's Lemma is a consequence.

Our analytical and numerical results on feedback channels draw on the ideas of Shannon and Alajaji, but are also encouraged by the results of Cover and Pombra. Shannon proved in [22] that feedback does not increase the capacity (or the capacity-cost function) of discrete memoryless channels. In [23] he goes on to prove the same result for discrete memoryless channels with *side information* (generalized feedback). Alajaji improved the scope of this result by extending it to discrete channels with arbitrary (not necessarily stationary ergodic) additive noise [1]. A question remained as to the effect of feedback on the capacity-cost function. Cover and Pombra showed in [10] that for continuous power constrained channels with non-white Gaussian noise feedback does help.

1.2 Contribution

The contribution of this thesis has both numeric and analytic significance. The numeric results flow from the implementation of Blahut's algorithm for the n^{th} capacity-cost function, and from algebraic manipulations on discrete alphabet Markov chains. A C++ program in the thesis performs the following tasks:

- computes Blahut's lower bound and Alajaji's upper bound for the capacity-cost function of non-feedback mod q Markov noise channels,

- computes Mrs. Gerber's Lemma for binary addition channels,

- solves for the stationary distribution of an arbitrary discrete stationary ergodic Markov process of order k and alphabet q such that $q^k \leq 16$.

As well, we derive analytically the following results for modulo channels with additive Markov noise.

- We establish a lower bound to the capacity-cost function of Markov noise channels with feedback. This bound can be numerically evaluated using Blahut's algorithm.

- We present a feedback strategy and a set of Markov noise sources for which the capacity-cost function with feedback is greater than the non-feedback capacity-cost function.

1.3 Thesis Overview

This thesis is organized in the following manner.

Chapter 2 presents background analysis needed to understand the arguments presented later. Our results concern channels with additive Markov noise; so we start by introducing the properties of a discrete Markov process. We proceed by defining the discrete channel in general and the notion of an average cost constraint on channel input symbols. The capacity-cost function $C(\beta)$ as well as the n^{th} capacity-cost function are defined and some of their properties are stated. We close by formulating the different channel models that will be used in the subsequent chapters. We also derive an expression for the actual capacity C of a mod q channel corrupted by a finite state Markov process.

Chapter 3 is devoted to the derivation and computation of existing bounds on the capacity-cost function. A tight lower bound which can be computed by an algorithm due to Blahut [8] converges slowly as the channel block length n increases, and is improved when used jointly with Mrs. Gerber's Lemma, a special case of the Binary Analog to the Power Entropy Inequality [20]. The upper bound, due to Alajaji [2], is a direct analogy to the rate-distortion function lower bound by Wyner and Ziv [26]. In the instance where the memory in the noise process is computable for both finite and infinite random sequences (Markov chains fit this description), Alajaji's upper bound converges to Blahut's lower bound showing that both bounds are tight in the limit as $n \to \infty$. These bounds are then implemented on a number of discrete Markov channels.

Chapter 4 focuses on additive noise channels with feedback, and develops a class of channels for which nonlinear feedback increases the capacity-cost function. To allow the previous definitions and algorithms to apply to feedback channels, some interpretation of terms is essential. Feedback can be viewed either as changing the noise process on the channel or as re-encoding the message sequence. To fully understand feedback in relation to the capacity-cost function, we must take both of these views simultaneously. For instance, the per letter cost in the non-feedback case is determined solely by the source probabilities. Under a feedback rule, the costs of channel input codewords can only be determined probabilistically using both the input probabilities and the channel noise process. Using the new costs, Blahut's algorithm can still be employed to evaluate a lower bound to the feedback capacity-cost function. By choosing an appropriate feedback scheme, a class of channels is provided, whose capacity-cost function is increased by feedback. The examples from previous chapters are incorporated to numerically illustrate this result.

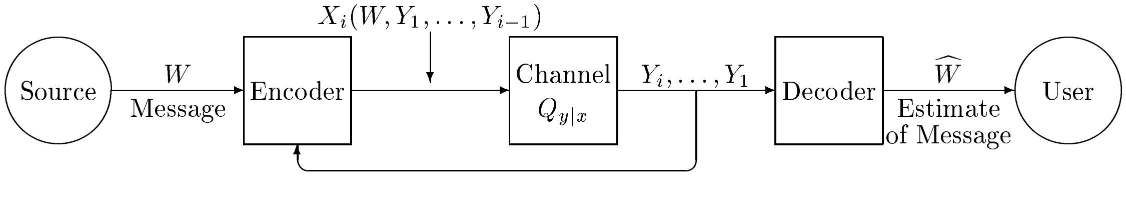


Figure 1.2: Discrete Channel Model with Feedback.

Chapter 5 summarizes our findings and points to future work and related problems that may be addressed by a similar approach.

The thesis concludes with two appendices. Appendix A presents the necessary background in information theory for a reader to follow the text. Appendix B is a complete proof of the Blahut-Arimoto algorithm for the n^{th} capacity-cost function $C_n(\beta)$.

Chapter 2

The Capacity-Cost Function $C(\beta)$

Before embarking on presenting our results, covering a certain amount of background is necessary. We start by defining and characterizing the notion of a Markov process, of a discrete addition channel, of a cost constraint and of a capacity-cost function.

The capacity-cost $(C(\beta))$ itself was first proposed by Shannon in his inaugural papers on information theory [21, 24]. Not only did Shannon invent the language of modern communications, he also proposed and solved many of the fundamental problems. Among them are the limit on the compressibility of a data source, the limit on the transmission rate of information across a noisy channel without memory, and the ability to separately design source and channel coding schemes without loss of optimality. Since Shannon's work, other researchers have extended his results by deriving the capacity formula - i.e. the maximum rate at which information can be reliably transmitted - for more general channel models. In this chapter, we introduce the concepts necessary for discussing a *constrained capacity*, i.e., the *capacity* subject to a maximum allowable *average cost* on the input sequence. As implementation is also an important aspect of this work, we will also introduce some examples of additive channels with memory that will be used to illustrate previous results and our new findings.

2.1 Markov Sources of Order k

A k^{th} order Markov process is one whose present value depends only on the previous k values, and is conditionally independent of any *older* ones. Exploring questions of channels with memory calls for a selection of noise sources that are both general and easy to analyze. Finite order Markov chains fit this description.

We now introduce the properties of discrete k^{th} order Markov sources with alphabet size q. The following definitions are a synthesis of those found in the books by Cover and Thomas [11], and Grimmett and Stirzaker [14]. The formulas for solving the stationary distribution were derived independently, and are necessary for the computer implementation required in the thesis.

Definition 2.1 A discrete stochastic process $\{Z_i\}_{i=1}^{\infty}$ with finite state space $\mathcal{Z} = \{0, 1, \ldots, q-1\}$ is said to be a *Markov process of order* k if,

$$\begin{aligned}\Pr(Z_n = z_n | Z_{n-1}{=}z_{n-1}, Z_{n-2}{=}z_{n-2}, \ldots, Z_1{=}z_1) = \\ \Pr(Z_n = z_n | Z_{n-1}{=}z_{n-1}, Z_{n-2}{=}z_{n-2}, \ldots, Z_{n-k}{=}z_{n-k}),\end{aligned} \tag{2.1}$$

for all $z_n \in \mathcal{Z} = \{0, 1, \ldots, q-1\}$, and for all $n \in \{k+1, k+2, \ldots\}$.

Definition 2.2 A stochastic process is said to be *stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index, i.e.,

$$\Pr(Z_1{=}z_1, Z_2{=}z_2, \ldots, Z_n{=}z_n) = \Pr(Z_{1+l}{=}z_1, Z_{2+l}{=}z_2, \ldots, Z_{n+l}{=}z_n), \tag{2.2}$$

for every time shift l and for all $z_i \in \mathcal{Z},\ i \in \{1, 2, \ldots\}$.

If we denote the random k-tuple $(Z_{n-k}, Z_{n-k+1}, \ldots, Z_{n-2}, Z_{n-1})$ by S_n, the state at time n, then the state transition probabilities are related to $\Pr(\cdot|\cdot)$ as follows

$$\Pr(S_{n+1}|S_n) \;=\; \Pr(Z_n, Z_{n-1}, \ldots, Z_{n-k+1}|Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k}) \tag{2.3}$$

$$= \Pr(Z_n|Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k}). \tag{2.4}$$

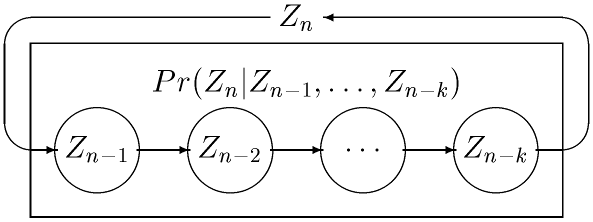


Figure 2.1: k^{th} Order Markov Process in State $S_n = (Z_{n-1}, \ldots, Z_{n-k})$.

It is easily seen that the states $\{S_n\}$, form a new first order Markov process with $S_n = (Z_{n-k}, \ldots, Z_{n-1}) \in \mathcal{Z}^k$, indexed for time $n > k$. If the cardinality of the noise process is q (denoted $|\mathcal{Z}| = q$) we get a $q^k \times q^k$ state transition matrix $\mathbf{\Pi}^{(n)}$ built from the q^{k+1} conditional probabilities, $\Pr(Z_n|Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k})^1$,

$$\mathbf{\Pi}^{(n)} = [\Pr(S_{n+1}|S_n)] = [\Pr(Z_n|Z_{n-1}, \ldots, Z_{n-k})], \tag{2.5}$$

indexed along its rows by vectors in $\mathcal{Z}^k$ representing the present channel state and along its columns by vectors in $\mathcal{Z}^k$ representing the next channel state. Since the cardinality of the set of all noise states is q^k, we can assign an integer value l to any state $s_n = (z_{n-k}, \ldots, z_{n-1})$ according to the following rule

$$\begin{aligned} l(s_n) &\stackrel{\triangle}{=} z_{n-k}q^{k-1} + z_{n-k+1}q^{k-2} + \cdots + z_{n-2}q^1 + z_{n-1} \\ &= \sum_{i=1}^{k} z_{n-i}q^{i-1}, \end{aligned} \tag{2.6}$$

where $z_i \in \{0, 1, \ldots, q-1\}$ for all integers $i > 0$. In the computer implementation of Blahut's algorithm all n-tuples are indexed in this manner.

[1] The remaining $q^k(q^k - q)$ entries in the matrix are set to zero, as they represent impossible state transitions of the form

$$\Pr(Z_n{=}z_n, Z_{n-1}{=}\hat{z}_{n-1}, \ldots, Z_{n-k+1}{=}\hat{z}_{n-k+1}|Z_{n-1}{=}z_{n-1}, \ldots, Z_{n-k}{=}z_{n-k}),$$

where at least one of the $\hat{z}_i \neq z_i$ for $i = n-k+1, \ldots, n-1$.

Definition 2.3 A Markov chain is said to be *time invariant* (or *homogeneous*) if the conditional probability $\Pr(S_{n+1}|S_n)$ of the Markov states does not depend on n. More specifically,

$$\Pr(S_{n+1} = b|S_n = a) = \Pr(S_2 = b|S_1 = a) \tag{2.7}$$

for all $a, b \in \mathcal{Z}^k$ and for all times $n \in \{1, 2, \ldots\}$, which implies that

$$\mathbf{\Pi}^{(n)} = \mathbf{\Pi}. \tag{2.8}$$

We will assume throughout this thesis that the conditional probabilities are homogeneous.

We now classify the Markov chain defined by $\mathbf{\Pi}$ on the state space $\mathcal{S} = \mathcal{Z}^k$, according to the conventions given in [14]. Our immediate goal is a general algorithm for the stationary probability distribution of the states.

Definition 2.4 State i is called *persistent* (or *recurrent*) if

$$\Pr(S_n = i \text{ for some } n > 1|S_1 = i) = 1. \tag{2.9}$$

Literally, the probability of eventual return to i, having started from i, is 1. If this probability is strictly less than 1, i is called a *transient* state.

As we are interested in state transitions, let

$$f_{ij}(n) = \Pr(S_2 \neq j, \ldots, S_{n-1} \neq j, S_n = j|S_1 = i) \tag{2.10}$$

be the probability that the first visit to state j, starting from state i, takes place at time n. Let

$$p_{ij}(n) = \Pr(S_n = j|S_1 = i) \tag{2.11}$$

be the probability that state j occurs $n - 1$ steps after state i.

Definition 2.5 ([14]) A persistent state i is called *null* if

$$\sum_n n f_{ii}(n) = \infty. \tag{2.12}$$

Otherwise we say that the persistent state is *non-null* (or positive).

Theorem 2.1 ([14]) A persistent state is *null* iff $p_{ii}(n) \to 0$ as $n \to \infty$; if this holds then $p_{ji}(n) \to 0$ for all j.

Definition 2.6 The *period* $d(i)$ of a state i is defined by

$$d(i) = \gcd\{n : p_{ii}(n) > 0\}, \tag{2.13}$$

the greatest common divisor of the epochs at which return is possible. We call i *periodic* if $d(i) > 1$ and *aperiodic* if $d(i) = 1$. That is to say, $p_{ii}(n) = 0$ unless n is a multiple of $d(i)$.

Definition 2.7 We say i *communicates* with j, written $i \to j$ if the chain may ever visit state j with positive probability, starting from state i. That is, $i \to j$ if $p_{ij}(m) > 0$ for some $m \geq 1$. We say i and j *intercommunicate* if $i \to j$ and $j \to i$, in which case we write $i \leftrightarrow j$. If $i \neq j$, then $i \to j$ iff $\sum_n f_{ij}(n) > 0$.

Remark: it is simple to demonstrate that $i \leftrightarrow j$ is an equivalence relation.

- Clearly $i \leftrightarrow i$ since $p_{ii}(1) = 1$.
- By definition, $i \leftrightarrow j$ implies $j \leftrightarrow i$.
- If $i \leftrightarrow j$ at step m and $j \leftrightarrow k$ at step n then $p_{ik}(m+n) \geq p_{ij}(m)p_{jk}(n) > 0$. So $i \leftrightarrow k$, and the relation is transitive and the class is closed.

Therefore, the state space $\mathcal{S}$ can be partitioned into the equivalence classes of $\leftrightarrow$.

Theorem 2.2 ([14]) If $i \leftrightarrow j$ then

a) i and j have the same period,

b) i is transient iff j is transient,

c) i is null persistent iff j is null persistent.

Definition 2.8 A set of states B is called

a) *closed* if $p_{ij} = 0$ for all $i \in B$, $j \notin B$,

b) *irreducible* if $i \leftrightarrow j$ for all $i, j \in B$.

A finite-alphabet stationary Markov process $\{Z_i\}_{i=1}^{\infty}$ is *ergodic* iff it is irreducible (i.e. any state is achievable with positive probability from any other state in a finite number of steps); furthermore, it is *mixing* iff it is irreducible and aperiodic [6].

In this thesis, we require stationary mixing noise. Therefore, our Markov processes must be at least irreducible and aperiodic. We will now show that these are sufficient conditions to guarantee the existence of a stationary distribution as well.

Definition 2.9 The vector $\boldsymbol{\pi}$ is called a *stationary distribution* of the chain if $\boldsymbol{\pi}$ has components $(\pi_j : j \in \mathcal{S})$ such that

a) $\pi_j \geq 0$ for all j, and $\sum_j \pi_j = 1$,

b) $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{\Pi}$, which is to say that $\pi_j = \sum_i \pi_i \mathbf{\Pi}_{ij}$ for all j.

Theorem 2.3 (Perron-Frobenius) If $\mathbf{\Pi}$ is the transition matrix of a finite irreducible chain with period d then

a) $\lambda_1 = 1$ is an eigenvalue of $\mathbf{\Pi}$,

b) the d complex roots of unity

$$\lambda_1 = \omega^0, \lambda_2 = \omega^1, \ldots, \lambda_d = \omega^{d-1}, \tag{2.14}$$

where $\omega = \exp(2\pi \mathrm{i}/d)$, are eigenvalues of $\mathbf{\Pi}$,

c) the remaining eigenvalues $\lambda_{d+1}, \ldots, \lambda_N$ satisfy $|\lambda_j| < 1$.

Theorem 2.4 For a finite-alphabet, irreducible and aperiodic Markov chain with transition matrix $\mathbf{\Pi}$, the stationary probability distribution $\boldsymbol{\pi}$ on the states $j \in \mathcal{S}$ is unique, and solved by $\boldsymbol{\pi}(I_{|\mathcal{S}|} - \mathbf{\Pi}) = 0$ with the normalizing constraint $\sum_j \pi_j = 1$.

Proof of Theorem 2.4 Applying Theorem 2.3 to an aperiodic process (i.e., $d = 1$) results in a single left eigenvector solution to

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{\Pi}, \tag{2.15}$$

$$\boldsymbol{\pi}(\lambda_1 I_{|\mathcal{S}|} - \mathbf{\Pi}) = 0, \tag{2.16}$$

where $\lambda_1 = 1$ and I is the identity matrix of appropriate size. The solution to (2.16) is unique under the normalizing constraint $\sum_j \pi_j = 1$. □

The computer algorithm that solves for the steady-state probabilities in $\boldsymbol{\pi}$, makes the following observations for an irreducible aperiodic Markov chain on the state space $\mathcal{S} = \mathcal{Z}^k$.

- $\mathbf{\Pi}$ has a unique left eigenvector solution, therefore $K = (I_{q^k} - \mathbf{\Pi})$ has rank[2] equal to the number of columns minus one: $\text{Rank}(K) = q^k - 1$.
- If any proper subset of columns $C \subset K$ were linearly dependent, then the states corresponding to columns in C would constitute a separate equivalence class under $\leftrightarrow$. This implies reducibility, but our chain is irreducible. Therefore,

[2] The rank of a matrix is defined as the number of linearly independent rows or columns. For more precise definitions please see any text on linear algebra , for example [17].

any subset containing $q^k - 1$ columns of $K = (I_{q^k} - \mathbf{\Pi})$ is a linearly independent set.

We now explain the algorithm for the computation of $\boldsymbol{\pi}$. Deleting any column of $K = I_{q^k} - \mathbf{\Pi}$ above results in a reduced system of equations that is linearly independent. Let us delete the right most column of K to form matrix $\widehat{K}$. Substituting $\widehat{K}$ into (2.16) yields

$$[\pi_0, \ldots, \pi_{q^k-1}]\widehat{K} = \underline{0} \tag{2.17}$$

where $\widehat{K}$ equals[3]

$$\begin{bmatrix} 1 - P_{S|S}(0|0) & -P_{S|S}(1|0) & \cdots & -P_{S|S}(q^k-2|0) \\ -P_{S|S}(0|1) & 1 - P_{S|S}(1|1) & \cdots & -P_{S|S}(q^k-2|1) \\ \vdots & \vdots & \ddots & \vdots \\ -P_{S|S}(0|q^k-2) & -P_{S|S}(1|q^k-2) & \cdots & 1 - P_{S|S}(q^k-2|q^k-2) \\ -P_{S|S}(0|q^k-1) & -P_{S|S}(1|q^k-1) & \cdots & -P_{S|S}(q^k-2|q^k-1) \end{bmatrix}. \tag{2.18}$$

Column reduction on this new matrix leads to:

$$[\pi_0, \ldots, \pi_{q^k-1}] \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{q^k-2} \end{bmatrix} = \underline{0} \tag{2.19}$$

where it can be seen that

$$\pi_i = \alpha_i \pi_{q^k-1} \tag{2.20}$$

for $i \in \{0, 1, \ldots, q^k - 2\}$. Notice, however, that

$$\sum_{i=0}^{q^k-1} \pi_i = \pi_{q^k-1} + \sum_{i=0}^{q^k-2} \alpha_i \, \pi_{q^k-1} = 1. \tag{2.21}$$

3We now define $P_{S|S}(i|j) \overset{\triangle}{=} \Pr(S = i | S = j)$.

Therefore,

$$\pi_{q^k-1} = \left(1 + \sum_{i=0}^{q^k-2} \alpha_i\right)^{-1}. \tag{2.22}$$

This value can then be substituted back into Equation (2.20) for all $i \in \{0, 1, \ldots, q^k - 2\}$ to complete the computation of the stationary distribution.

In the above algorithm we could have rearranged terms to delete any particular column, i, and solved the equations for π_i instead of π_{q^k-1} without loss of generality.

Now that the stationary distribution $\boldsymbol{\pi}$ of the Markov chain has been determined, we can compute the entropy of the chain assuming stationary initial conditions. The following derivation assumes a rudimentary knowledge of information theory. For a description of basic information theory concepts please refer to [11] or [18].

Theorem 2.5 The entropy rate of a stationary Markov process $\{Z_i\}_{i=1}^{\infty}$ is

$$\begin{aligned} H(Z_\infty) &= H(Z_{k+1}|Z_k, \ldots, Z_1) && (2.23) \\ &= -\sum_{l,m \in \mathcal{Z}^k} \pi_l \, \mathbf{\Pi}_{l,m} \, \log(\mathbf{\Pi}_{l,m}), && (2.24) \end{aligned}$$

where $\mathbf{\Pi}_{l,m}$ represents the state transition probability $Pr(S_2 = m|S_1 = l)$, and where π_l is the l^{th} component of the stationary distribution vector $\boldsymbol{\pi}$.

Proof of Theorem 2.5

$$\begin{aligned} H(Z_\infty) &= \lim_{n\to\infty} \frac{1}{n} H(Z_1, Z_2, \ldots, Z_n) && (2.25) \\ &= \lim_{n\to\infty} H(Z_n|Z_{n-1}, \ldots, Z_1) && (2.26) \\ &= H(Z_{k+1}|Z_k, \ldots, Z_1). && (2.27) \end{aligned}$$

The first equality is simply the definition of entropy rate. The second equality follows from Theorem 4.2.1 in [11], which is a well known result for the entropy rate of any stationary stochastic process. The final step is simply an application of the definition of a stationary Markov chain. We express this in terms of our state notation and compute the conditional entropy as

$$H(Z_\infty) = H(S_2|S_1) \tag{2.28}$$

$$= -\sum_{l,m\in\mathcal{Z}^k} \pi_l \mathbf{\Pi}_{l,m} \log(\mathbf{\Pi}_{l,m}). \tag{2.29}$$

□

2.2 Discrete Channels with Cost Constraints

A discrete channel is characterized by an input process $\{X_i\}_{i=1}^{\infty}$ with finite alphabet $\mathcal{X}$, an output process $\{Y_i\}_{i=1}^{\infty}$ with finite alphabet $\mathcal{Y}$, and a set of block transition distributions

$$\Big\{P_{Y^n|X^n}(y_1, y_2, \ldots, y_n|x_1, x_2, \ldots, x_n) \triangleq \\ \Pr(Y_1{=}y_1, Y_2{=}y_2, \ldots, Y_n{=}y_n|X_1{=}x_1, X_2{=}x_2, \ldots, X_n{=}x_n)\big\}_{n=1}^{\infty}, \tag{2.30}$$

where $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$ and $i = 1, 2, \ldots, n$. When $|\mathcal{X}| = r$ and $|\mathcal{Y}| = t$ (i.e. the cardinality of $\mathcal{X}$ and $\mathcal{Y}$ are r and t) we denote each set as $\{0, 1, \ldots, r-1\}$ and $\{0, 1, \ldots, t-1\}$ respectively.

Furthermore, letting $x^n \triangleq (x_1, x_2, \ldots, x_n)$ represent a block of inputs to the channel and $y^n \triangleq (y_1, y_2, \ldots, y_n)$ a block of outputs, the channel transition distributions satisfy

$$P_{Y^n|X^n}(y^n|x^n) \geq 0, \qquad \forall\ x^n \in \mathcal{X}^n,\ y^n \in \mathcal{Y}^n; \tag{2.31}$$

$$\sum_{y^n\in\mathcal{Y}^n} P_{Y^n|X^n}(y^n|x^n) = 1, \qquad \forall\ x^n \in \mathcal{X}^n. \tag{2.32}$$

For convenience, the transition probabilities are written in matrix form. Setting the size of the input alphabet $|\mathcal{X}| = r$ and the output alphabet $|\mathcal{Y}| = t$, we can define

$$Q \triangleq \Big[P_{Y^n|X^n}(y^n|x^n)\Big] \tag{2.33}$$

as an $r^n \times t^n$ matrix indexed along its rows by vectors in $\mathcal{X}^n$ and along its columns by vectors in $\mathcal{Y}^n$. Note that throughout this thesis almost all alphabets will have cardinality q. The only exception being the super-alphabets over which we will index either the input, noise and output n-tuples, or the order k Markov states, which will have cardinality q^n and q^k respectively.

The 'n' inputs are instances of the input process $\{X_i\}_{i=1}^{\infty}$, $X_i \in \mathcal{X}$, and are represented probabilistically by the random vector $X^n \triangleq (X_1, X_2, \ldots, X_n)$ with joint probability mass function

$$P_{X^n}(x^n) \triangleq \Pr(X_1{=}x_1, X_2{=}x_2, \ldots, X_n{=}x_n). \tag{2.34}$$

Outputs are similarly represented with probability distribution

$$P_{Y^n}(y^n) \triangleq \Pr(Y_1{=}y_1, Y_2{=}y_2, \ldots, Y_n{=}y_n). \tag{2.35}$$

The discrete channel is further subject to a constraint imposed on the inputs. For each $x \in \mathcal{X}$, let $b(x)$ denote the non-negative cost associated with transmitting letter x. We assume that

$$b_{max} \triangleq \max_{x \in \mathcal{X}} b(x) \tag{2.36}$$

is finite. As we intend to make use of the channel many times in succession, we interpret the cost of sending $x^n = (x_1, x_2, \ldots, x_n)$ as an additive cost; i.e.,

$$b(x^n) = \sum_{i=1}^{n} b(x_i). \tag{2.37}$$

In general, the probability of sending each possible input vector, x^n, need not be equal. We define the "average" or "expected" cost of a given input distribution to be

$$E\left[b(X^n)\right] = \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) b(x^n) \tag{2.38}$$

$$= \sum_{i=1}^{n} E\left[b(X_i)\right]. \tag{2.39}$$

A means of comparing block codes of varying length is achieved by scaling the average cost by $1/n$, which leads to the following definition.

Definition 2.10 An n-dimensional random vector $X^n = (X_1, X_2, \ldots, X_n)$ is said to be *β-admissible* if it satisfies

$$\frac{1}{n} E\left[b(X^n)\right] \le \beta \tag{2.40}$$

The set of all n-dimensional β-admissible input distributions is denoted by $\tau_n(\beta)$:

$$\tau_n(\beta) = \{P_{X^n}(x^n) : \frac{1}{n} E\left[b(X^n)\right] \le \beta\}. \tag{2.41}$$

Relating this to our channel model in Fig 1.1, the inputs to the channel are themselves outputs of the source-channel coding block. In this sense, x^n can be viewed as a channel codeword.

Definition 2.11 A *channel block code* of length n over $\mathcal{X}$ is a subset of $\mathcal{X}^n$, $\mathcal{C} = \{c_{(1)}, \ldots, c_{(|\mathcal{C}|)}\}$, where each $c_{(i)}$ is an n-tuple. The rate of the code is $R = \frac{1}{n}\log_2|\mathcal{C}|$ bits per channel symbol.[4] The code is β-admissible if $b(c_{(i)}) \le n\beta$ for $i = 1, 2, \ldots, |\mathcal{C}|$. If the encoder wants to transmit message W, where W is uniform over $\{1, 2, \ldots, |\mathcal{C}|\}$, it sends the codeword $c_{(W)}$. (This process is represented in Figure 1.1 by the function $X^n = f(W) = f_2(f_1(W))$.) At the channel output, the decoder receives Y^n and chooses as estimate of the message $\widehat{W} = g(Y^n)$, where $g(\cdot)$ is a decoding rule. The (average) probability of decoding error is then $P_e^{(n)} = \Pr\{g(Y^n) \ne W\}$.

2.3 Capacity-Cost Function

A discrete channel is stationary if for every stationary input process $\{X_i\}_{i=1}^{\infty}$, the resultant input-output process $\{X_i, Y_i\}_{i=1}^{\infty}$ is stationary. Similarly, a discrete channel

[4] We assume throughout this thesis that the logarithms log are in base 2.

is ergodic if for every ergodic input process $\{X_i\}_{i=1}^{\infty}$ an ergodic input-output process $\{X_i, Y_i\}_{i=1}^{\infty}$ results.

Definition 2.12 The capacity-cost function $C(\beta)$ and the capacity C of discrete stationary channels with memory are respectively defined by [18]

$$C(\beta) = \sup_n \; C_n(\beta) = \lim_{n\to\infty} \; C_n(\beta), \tag{2.42}$$

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

$$C_n(\beta) \triangleq \max_{P_{X^n}(x^n) \in \tau_n(\beta)} \frac{1}{n} \, I(X^n; Y^n), \tag{2.43}$$

and

$$C = \sup_n \; C_n = \lim_{n\to\infty} \; C_n, \tag{2.44}$$

where C_n is the is the n^{th} *capacity* given by

$$C_n \triangleq \max_{P_{X^n}(x^n)} \frac{1}{n} \, I(X^n; Y^n), \tag{2.45}$$

where $I(X^n; Y^n)$ is the block mutual information between the input vector X^n and the output vector Y^n.

Strictly speaking, the fact that the limits are equal to the supremums in Equations (2.42) and (2.44) is a non-trivial result. In Theorem A.3 of Appendix A we demonstrate this result.

The capacity-cost function $C(\beta)$ has an *operational* significance for channels satisfying certain regularity conditions (e.g., a stationary ergodic channel, a discrete channel with stationary mixing additive noise, or an information stable channel [13, 19, 25]). More specifically, $C(\beta)$ represents 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). Equivalently, $C(\beta)$ is the maximum amount of information that can be transmitted reliably over the channel such that the expected cost per symbol is $\leq \beta$. If $b(x) = 0$ for every letter $x \in \mathcal{X}$, $C(\beta)$ reduces to the channel capacity C.

We next discuss the properties of $C_n(\beta)$ and $C(\beta)$ for a given discrete channel and a cost function [18]. We first observe that if we define

$$\beta_{min} = \min_{x \in \mathcal{X}} \; b(x), \tag{2.46}$$

then $\frac{1}{n} E[b(X^n)] \geq \beta_{min}$; this implies that $C_n(\beta)$ and $C(\beta)$ are defined only for $\beta \geq \beta_{min}$.

Lemma 2.1 $C_n(\beta)$ and $C(\beta)$ are *concave* and *non-decreasing* functions of β, for $\beta \geq \beta_{min}$.

Proof of Lemma 2.1 See Appendix B.

Since $C_n(\beta)$ and $C(\beta)$ are concave, they are also continuous for $\beta \geq \beta_{min}$. If we define

$$\beta_{max}^{(n)} \triangleq \min \left\{ \frac{1}{n} E[b(X^n)] : \frac{1}{n} I(X^n; Y^n) = C_n \right\}, \tag{2.47}$$

and

$$\beta_{max} \triangleq \lim_{n \to \infty} \beta_{max}^{(n)} \tag{2.48}$$

then clearly

$$C_n(\beta) = C_n \qquad \text{for all } \beta \geq \beta_{max}^{(n)} \tag{2.49}$$

and

$$C(\beta) = C \qquad \text{for all } \beta \geq \beta_{max}. \tag{2.50}$$

Remark: For a discrete q-ary channel with additive stationary noise $\{Z_i\}_{i=1}^{\infty}$ and linear cost function on the input – i.e. $b(i) = i$, $i \in \{0, 1, \ldots, q-1\}$ – we get that

$$\beta_{min} = 0,$$

$$\beta_{max}^{(n)} = \beta_{max} = \frac{q-1}{2},$$

$$C\left(\beta_{min}\right) = 0,$$

$$C\left(\beta_{max}^{(n)}\right) = C_n = \log_2(q) - \frac{1}{n}H(Z^n),$$

and

$$C\left(\beta_{max}\right) = C = \log_2(q) - H(Z_\infty),$$

where $\frac{1}{n}H(Z^n)$ is the normalized block noise entropy and $H(Z_\infty) = \lim_{n\to\infty}\frac{1}{n}H(Z^n)$ is the noise entropy rate.

Corollary 2.1 $C_n\left(\beta\right)$ is *strictly increasing* in β for $\beta_{min} \leq \beta \leq \beta_{max}^{(n)}$. Therefore, $C\left(\beta\right)$ is *strictly increasing* in β for $\beta_{min} \leq \beta \leq \beta_{max}$.

Proof of Corollary 2.1 See Appendix B.

2.4 Computation of Capacity for Discrete Markov Channels

A number of relatively general additive noise channels with memory are now developed to illustrate the above results and definitions, and to demonstrate the computation of channel capacity. These examples will be referred to in the following chapters.

All of our examples use stationary ergodic additive Markov noise as the corrupting process. The examples are differentiated by the order of the Markov chain and by the cardinality of the input alphabet. The first three additive channels are all examples of modulo-q channels with first order Markov memory in the noise. The final example uses a binary channel with a 2$^{\text{nd}}$ order Markov noise. Our additive noise channels are described by

$$Y_i = X_i \oplus Z_i, \tag{2.51}$$

where Z_j is independent of X_i for all $i, j \in \{1, 2, \ldots\}$.

The independence of the noise process and the input sequence induce a symmetry in the channel transition matrix $Q \triangleq \left[P_{Y^n|X^n}(y^n|x^n)\right]$. By the invertibility of $\oplus$, the conditional probabilities become

$$P_{Y^n|X^n}(y^n|x^n) = P_{Z^n}(y^n \ominus x^n). \tag{2.52}$$

Thus all rows or columns of Q are simply permutations of each other. Channels conforming to this pattern are known as symmetric channels. We now state the following definition from [11].

Definition 2.13 A channel is said to be *symmetric* if the rows of the channel transition matrix $\left[P_{Y^n|X^n}(y^n|x^n)\right]$ are permutations of each other, and the columns are permutations of each other. A channel is said to be *weakly symmetric* if every row of the transition matrix $\left[P_{Y^n|X^n}(\cdot|x^n)\right]$ is a permutation of every other row, and all the column sums $\sum_{x^n} P_{Y^n|X^n}(y^n|x^n)$ are equal.

For such symmetric additive channels the conditional entropy $H(Y^n|X^n)$ is equal to $H(Z^n)$, and the capacity, $C = \sup_n \, C_n$ is achieved by an iid uniformly distributed input process. The following is a generalization of the theorem in [18].

Theorem 2.6 If a weakly symmetric channel has r^n inputs and s^n outputs, its n^{th} capacity is achieved with equiprobable inputs, i.e., $P_{X^n}(x^n) = 1/r^n$, for all $x^n \in \mathcal{X}^n$, and the n^{th} capacity is

$$C_n = \log(s) - \frac{1}{n} H(p_0, p_1, \ldots, p_{s^n-1}), \tag{2.53}$$

where $(p_0, p_1, \ldots, p_{s^n-1})$ is any row of the transition matrix. For modulo addition channels, $(p_0, p_1, \ldots, p_{s^n-1})$ consists of the probability distribution on the set of all possible noise n-tuples $P_{Z^n}(z^n)$.

Proof of Theorem 2.6 By the definition of block mutual information,

$$\frac{1}{n}I(X^n;Y^n) = \frac{1}{n}H(Y^n) - \frac{1}{n}H(Y^n|X^n). \tag{2.54}$$

We can expand the second term in the sum as

$$\frac{1}{n}H(Y^n|X^n) = \frac{1}{n}\sum_{x^n} H(Y^n|X^n = x^n)P_{X^n}(x^n). \tag{2.55}$$

But since every row is a permutation of every other row,

$$\begin{aligned} H(Y^n|X^n = x^n) &= \sum_{y^n} P_{Y^n|X^n}(y^n|x^n)\log P_{Y^n|X^n}(y^n|x^n) & (2.56)\\ &= H(p_0, p_1, \ldots, p_{s^n-1}), & (2.57)\end{aligned}$$

and is independent of x^n. A well known result (c.f. Theorem 1.1 [18]) tells us that $\frac{1}{n}H(Y^n) \leq \log s$, with equality iff $P_{Y^n}(y^n) = \frac{1}{s^n}$ for all y^n. Fortunately, the condition on the columns of the transition matrix guarantees that, if $P_{X^n}(x^n) = \frac{1}{r^n}$ for all x^n, then $P_{Y^n}(y^n) = \frac{1}{s^n}$ for all y^n. Substituting this observation back into Equation (2.54) maximizes the block mutual information, so

$$C_n = \max_{P_{X^n}(x^n)} \frac{1}{n}I(X^n;Y^n) = \log(s) - \frac{1}{n}H(p_0, p_1, \ldots, p_{s^n-1}). \tag{2.58}$$

To convince ourselves that $(p_0, p_1, \ldots, p_{s^n-1})$ really is the distribution of $P_{Z^n}(z^n)$, let us recall that $\oplus$ is a one-to-one operation and that

$$H(Y^n|X^n) = H(Z^n \oplus X^n|X^n) = H(Z^n). \tag{2.59}$$

□

We next determine the capacity of mod q additive Markov noise channels.

$$\begin{aligned} C &= \lim_{n\to\infty} C_n & (2.60)\\ &= \lim_{n\to\infty}\left\{\max_{P_{X^n}(x^n)} \frac{1}{n}\, I(X^n;Y^n)\right\} & (2.61)\end{aligned}$$

$$= \lim_{n\to\infty} \frac{1}{n}\left\{\max_{P_{X^n}(x^n)} [H(Y^n) - H(Y^n|X^n)]\right\} \tag{2.62}$$

$$= \lim_{n\to\infty} \frac{1}{n}\left\{\max_{P_{X^n}(x^n)} [H(Y^n) - H(X^n \oplus Z^n|X^n)]\right\} \tag{2.63}$$

$$= \lim_{n\to\infty} \frac{1}{n}\left\{\max_{P_{X^n}(x^n)} [H(Y^n) - H(Z^n)]\right\} \tag{2.64}$$

$$= \lim_{n\to\infty} \frac{1}{n}\left\{\max_{P_{X^n}(x^n)} H(Y^n)\right\} - \lim_{n\to\infty} H(Z_\infty). \tag{2.65}$$

We next observe that since mod q additive noise channels are symmetric, $H(Y^n)$ is maximized for a uniform input distribution which also yields a uniform output distribution. From Theorem 2.5 on the entropy rate of stationary Markov processes of order k, $H(Z_\infty)$ is nothing more than $H(Z_n|Z_{n-1},\ldots,Z_{n-k})$. Therefore, for all q-ary channels with this type of additive noise,

$$\begin{aligned} C &= \lim_{n\to\infty} C_n = \lim_{n\to\infty} \frac{1}{n}\left\{-\sum_{i=0}^{q^n-1} \frac{1}{q^n}\log\frac{1}{q^n}\right\} - H(Z_n|Z_{n-1},\ldots,Z_{n-k}) \\ &= \log(q) + \sum_{l,m=0}^{q^k-1} \pi_l\, \mathbf{\Pi}_{l,m}\, \log(\mathbf{\Pi}_{l,m}), \end{aligned} \tag{2.66}$$

where, π_l is the stationary probability that state l occurs and $\mathbf{\Pi}_{l,m}$ is the noise state transition probability from state l to state m, using the indexing function defined in (2.6).

Example 2.1 *Binary Alphabet Channel with* 1^{st} *Order Markov Noise.*

The mod 2 first order Markov channel, is characterized by two states $(0, 1)$ and the noise state transition probabilities α and ϱ from $0 \to 1$ and $1 \to 0$ respectively, where $0 < \alpha < 1$ and $0 < \varrho < 1$ (c.f. Figure 2.2). Solving for the stationary distribution $\boldsymbol{\pi} = [P_Z(0), P_Z(1)]$ is straightforward.

$$(1-\alpha)\,\pi_0 + \varrho\,\pi_1 = \pi_0 \tag{2.67}$$

$$\varrho\,\pi_1 = \alpha\,\pi_0, \tag{2.68}$$

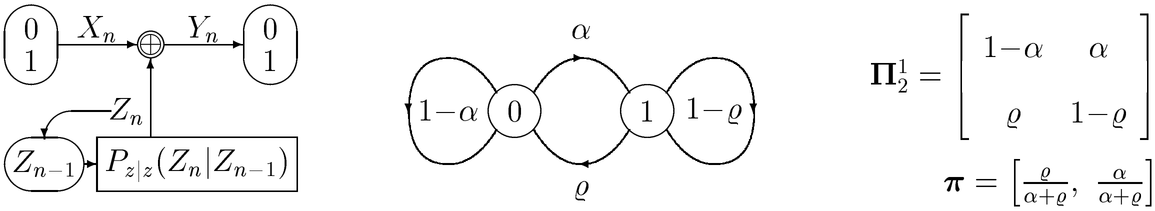

0
1
X_n
Y_n
0
1
Z_n
Z_{n-1}
$P_{z|z}(Z_n|Z_{n-1})$
α
$1-\alpha$
0
1
$1-\varrho$
ϱ
$\mathbf{\Pi}_2^1 = \begin{bmatrix} 1-\alpha & \alpha \\ \varrho & 1-\varrho \end{bmatrix}$
$\boldsymbol{\pi} = \left[\frac{\varrho}{\alpha+\varrho}, \frac{\alpha}{\alpha+\varrho}\right]$

Figure 2.2: Two-State Markov Process.

together with $\pi_0 + \pi_1 = 1$ yields

$$\boldsymbol{\pi} = \left[\frac{\varrho}{\alpha+\varrho}, \frac{\alpha}{\alpha+\varrho}\right]. \tag{2.69}$$

Using (2.66) on this particular example, the general capacity is derived as

$$\begin{aligned} C &= \log(2) + \sum_{l=0}^{1} \pi_l \sum_{m=0}^{1} \mathbf{\Pi}_{l,m} \log \mathbf{\Pi}_{l,m} && (2.70) \\ &= \log(2) - \frac{\varrho}{\alpha+\varrho} h(\alpha) + \frac{\alpha}{\alpha+\varrho} h(\varrho) && (2.71) \\ &= 1 - \frac{\varrho\, h(\alpha) + \alpha\, h(\varrho)}{\alpha+\varrho}, && (2.72) \end{aligned}$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the binary entropy function.

Example 2.2 *Ternary Alphabet Channel with 1^{st} order Markov Noise*

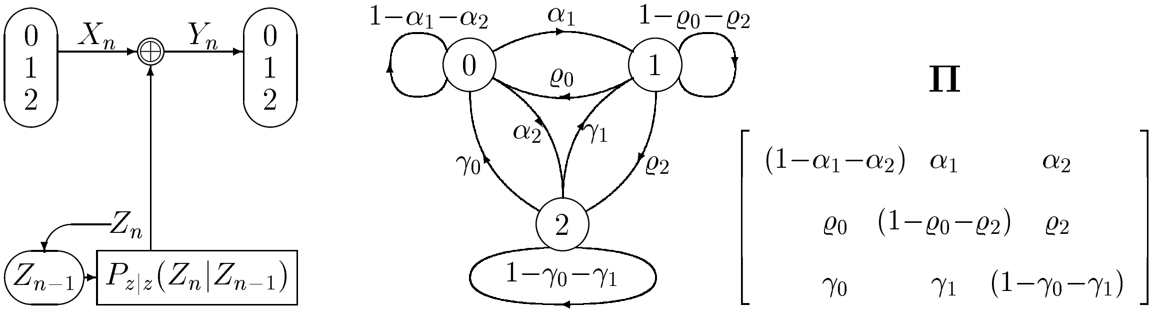


Figure 2.3: Three-State Markov Process.

We can see from Figure 2.3 that the level of variability is dramatically increased from the binary case. The result of Equation (2.66) can be applied to the ternary (three letter) alphabet channel as well. Solving for $\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2]$ using column reduction on $I_3 - \mathbf{\Pi}$ with the third column removed occurs as follows,

$$[\pi_0, \pi_1, \pi_2] \begin{bmatrix} P_{Z|Z}(0|0) & P_{Z|Z}(1|0) & P_{Z|Z}(2|0) \\ P_{Z|Z}(0|1) & P_{Z|Z}(1|1) & P_{Z|Z}(2|1) \\ P_{z|Z}(0|2) & P_{Z|Z}(1|2) & P_{Z|Z}(2|2) \end{bmatrix} = [\pi_0, \pi_1, \pi_2]. \qquad (2.73)$$

For convenience, label the probabilities in the first row as $\alpha_0, \alpha_1, \alpha_2$, those in the second row as $\varrho_0, \varrho_1, \varrho_2$ and those in the third row as $\gamma_0, \gamma_1, \gamma_2$. We then conveniently employ the fact that each row sums to 1, thereby obtaining the matrix of Figure 2.3. Bring all terms onto the left hand side gives

$$[\pi_0, \pi_1, \pi_2] \begin{bmatrix} (\alpha_1{+}\alpha_2) & -\alpha_1 & -\alpha_2 \\ -\varrho_0 & (\varrho_0{+}\varrho_2) & \varrho_2 \\ -\gamma_0 & -\gamma_1 & (\gamma_0{+}\gamma_1) \end{bmatrix} = [0, 0, 0]. \qquad (2.74)$$

Now let us remove the third column of the matrix and column reduce to get

$$[\pi_0, \pi_1, \pi_2] \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -A_0 & -A_1 \end{bmatrix} = [0, 0], \qquad (2.75)$$

where

$$\begin{aligned} -A_0 &= -\left(\frac{\gamma_0}{\alpha_1{+}\alpha_2} + \frac{\varrho_0}{\alpha_1{+}\alpha_2} \times A_1\right) && (2.76) \\ -A_1 &= -\left(\frac{\alpha_1(\gamma_0 + \gamma_1) \ + \ \alpha_2\gamma_1}{\alpha_2(\varrho_0 + \varrho_2) \ + \ \alpha_1\varrho_2}\right). && (2.77) \end{aligned}$$

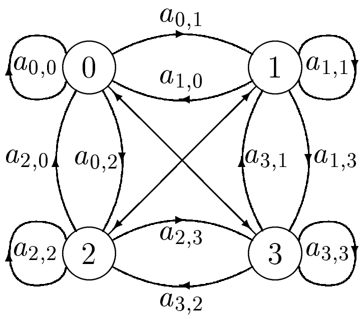
We can already see how complicated the equations get as the alphabet size increases.

Solving for $P_z(Z)$ using Equation (2.22) and Equation (2.20):

$$\begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} A_0/(1+A_0+A_1) \\ A_1/(1+A_0+A_1) \\ 1/(1+A_0+A_1) \end{bmatrix}. \tag{2.78}$$

As we use this type of channel in future chapters, the formula in Equation (2.66) will provide a $C\left(\beta_{max}\right)$ value that we can compare with the results of our bounding algorithms.

Example 2.3 *Quaternary Alphabet Channel with First order Markov Noise*



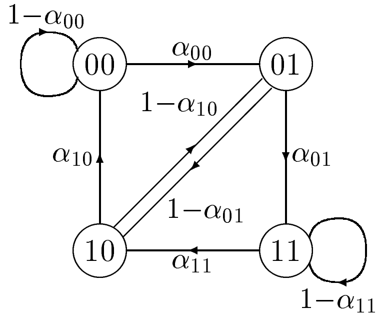
$$\mathbf{\Pi}_4^1$$

$$\begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & a_{0,3} \\ a_{1,0} & a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,0} & a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,0} & a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}$$

Figure 2.4: Four-State Markov Process.

The diagram in Figure 2.4 clearly shows that all transitions are possible between the four states. From the calculations in Example 2.2 we can see that the quaternary channel also has a known *capacity* C but that the form of the stationary probabilities as a function of parameters of $\mathbf{\Pi}$ is difficult to calculate by hand. Still, the algorithm for computing the stationary distribution is valid, and the computer implementation component of this thesis uses the generalized algorithm on all user specified noise processes.

Example 2.4 *Binary Alphabet Channel with 2^{nd} Order Markov Noise.*



$$\mathbf{\Pi}_2^2$$

$$\begin{bmatrix} 1-\alpha_{00} & \alpha_{00} & 0 & 0 \\ 0 & 0 & \alpha_{01} & 1-\alpha_{01} \\ 1-\alpha_{10} & \alpha_{10} & 0 & 0 \\ 0 & 0 & \alpha_{11} & 1-\alpha_{11} \end{bmatrix}$$

Figure 2.5: Second Order Binary (Four-State) Markov Process.

As is apparent from the diagrams and matrices in Figure 2.4 and Figure 2.5, the second order binary Markov process can be seen as a special case of the first order quaternary Markov process. To ease comparison with the first order binary model, we have labeled the 4 variables in $\mathbf{\Pi}_2^2$ using a simple symmetry. We label the state at time t by $S_t = (Z_{t-2}, Z_{t-1})$. Since this is a stationary time invariant system, we need only concern ourselves with S_1 and S_2. Notice that $(0,0)$ and $(1,1)$ are states with return loops (a 1-loop) that are not mutually joined (a two-loop), while $(0,1)$ and $(1,0)$ are states with a 2-loop but no 1-loop. Let $1-\alpha_{ij}$ be the probability that state (i,j) might enter a 1- or 2-loop, and let α_{ij} be the probability that the state (i,j) will not reoccur in the next two time steps. This notation helps simplify the computation of the stationary distribution $P_S(s)$. At first it may seem as though there is a problem with the periodicity of this chain, but the greatest common divisor of return times is indeed 1 for all states, since 2 and 3 are relatively prime, so long as $0 < \alpha_{ij} < 1$ for all $i, j \in \{0,1\}$.

While we have previously seemed interested in $P_Z(z)$, for a stationary first order Markov process, the distribution on the noise samples is identical to the distribution on the states $P_Z(z) = P_S(s) = \boldsymbol{\pi}$. Now that we are dealing with higher order chains, a small amount of additional work is required to get both. Applying Equation (2.18) to the matrix in Figure 2.5, we form the reduced matrix

$$\begin{bmatrix} \alpha_{00} & -\alpha_{00} & 0 \\ 0 & 1 & -\alpha_{01} \\ \alpha_{10}-1 & -\alpha_{10} & 1 \\ 0 & 0 & -\alpha_{11} \end{bmatrix}, \tag{2.79}$$

which we then column reduce to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -A_{00} & -A_{01} & -A_{10} \end{bmatrix}, \tag{2.80}$$

where

$$-A_{00} = \frac{-\alpha_{10}\alpha_{11}}{\alpha_{00}(1-\alpha_{01}+\alpha_{01}\alpha_{10})} \tag{2.81}$$

and

$$-A_{01} = -A_{10} = \frac{-\alpha_{11}}{1-\alpha_{01}+\alpha_{01}\alpha_{10}}. \tag{2.82}$$

The stationary distribution on the states is then solved as

$$\begin{bmatrix} \pi_{00} \\ \pi_{01} \\ \pi_{10} \\ \pi_{11} \end{bmatrix} = \frac{1}{v} \begin{bmatrix} \alpha_{10}\alpha_{11} \\ \alpha_{00}\alpha_{11} \\ \alpha_{00}\alpha_{11} \\ \alpha_{00}(1-\alpha_{01}+\alpha_{01}\alpha_{01}) \end{bmatrix}, \tag{2.83}$$

where

$$v = \alpha_{00}(1-\alpha_{01}+\alpha_{01}\alpha_{10}) + 2\alpha_{11}\alpha_{00} + \alpha_{10}\alpha_{11}. \tag{2.84}$$

The formulas for computing $P_Z(z)$ given a stationary distribution on the states are not complicated:

$$\begin{aligned} P_Z(0) &= \sum_{i,j=0}^{1} \Pr(Z_n = 0|Z_{n-2} = i, Z_{n-1} = j) P_{Z^2}(i,j) && (2.85) \\ &= \sum_{i,j=0}^{1} P_{S|S}(j0|ij)\pi_{ij}) && (2.86) \\ &= (1-\alpha_{00})\pi_{00} + (1-\alpha_{01})\pi_{01} \\ &\quad +\alpha 10\pi_{10} + \alpha 11\pi_{11} && (2.87) \\ P_Z(1) &= \sum_{i,j=0}^{1} P_{S|S}(j1|ij)\pi_{ij}. && (2.88) \end{aligned}$$

In subsequent chapters, we use the analytically determined value for C in (2.66) to verify the recursively determined value for $C_n\ (\beta_{max}) + M_n$ in the C++ implementation of Alajaji's Bound. In the next chapter we will define M_n as the difference between the per symbol entropy of the noise process and the entropy rate of the noise process.

The next chapter focuses on defining and computing bounds for the types of channels discussed in this section.

Chapter 3

Additive Noise Channels without Feedback

We consider q-ary additive noise channels of the type discussed in Chapter 2, where the noise process is stationary. Shannon first noticed that an interesting and fundamental duality exists between the rate-distortion function of a source ($R(D)$) and the capacity-cost function of a channel ($C(\beta)$) [24]. Using methods analogous to those of Wyner and Ziv, who found a lower bound to the Rate-Distortion function, Alajaji established a tight upper bound to the capacity of input constrained additive noise channels with memory [2]. Using the asymptotically tight lower bound for the capacity-cost function, given by direct computation of Blahut's algorithm, in conjunction with Alajaji's upper bound, numerical results indicate that a tight envelope is formed on the channel capacity-cost function. In the case of binary alphabet channels an additional bound can also be employed to better estimate the true capacity-cost function. This bound is an application of Mrs. Gerber's Lemma [20], which lower bounds $C\left(\beta\right)$ using some clever algebraic manipulations and the pseudo-invertibility of the binary entropy function.

In this chapter, we fully develop the mathematical framework behind these three bounding techniques before implementing them (where applicable) on the examples introduced in Section 2.4.

3.1 Existing Lower Bounds to the Capacity-Cost Function

Lower bounds on the capacity-cost function can be formed in two ways. By simply finding the per letter capacity-cost for the *associated block memoryless channel*[1] with input blocks of a fixed block length n, we can bound $C(\beta)$ from below:

$$C(\beta) = \sup_{n \geq 1} C_n(\beta) \geq C_n(\beta). \tag{3.1}$$

Blahut's algorithm is ideally suited to this type of computation, where a channel transition matrix Q provides the probability of receiving Y^n given that X^n was transmitted. Another lower bound exists for the capacity-cost function in the binary case. If we take the inputs to be iid and the alphabet to be binary, we can apply Mrs. Gerber's Lemma in [20] to obtain a lower bound on $C(\beta)$. As we are dealing primarily with q-ary channels, we use the $C_n(\beta)$ lower bound in all cases except the binary case where we also apply Mrs. Gerber's bound. The $C(\beta)$ lower bound is in fact dual to the upper bound on the rate-distortion function $R(D)$ also computed by Blahut in [8].

3.1.1 Blahut's Algorithm

As part of his 1971 doctoral dissertation at Cornell University, R. Blahut reformulated the computation of both $R_n(D)$ and $C_n(\beta)$ in nats, as a convex programming problem on the block mutual information between source and receiver, $I(X^n; Y^n)$.

[1] By this we mean that the memory between noise blocks is ignored.

Conversion from nats (information measured with natural logarithms) to bits (information measured with base 2 logarithms) is performed using division by a factor of $\ln 2$. We will now state Blahut's results concerning the constrained channel capacity as found in [8]. The proofs of all the theorems appear in Appendix B.

To facilitate reference to Blahut's very general paper, let us examine a block length n channel with differing source and receiver alphabets. The *per symbol* input alphabet $\mathcal{X} = \{0, 1, \ldots, r-1\}$ is expanded to interpret the entire block as a single entity (called a word for convenience) $j \in \{0, 1, \ldots, r^n - 1\}$. Let $N = r^n$ be the size of the new input dictionary. Similarly, the *per symbol* output alphabet $\mathcal{Y} = \{0, 1, \ldots, t-1\}$ is expanded to interpret the entire block as a single word $k \in \{0, 1, \ldots, M-1\}$, where $M = t^n$ is the size of the new output dictionary. The cost function is also reinterpreted according to this new indexing set on possible input words. If $x^n = (x_1, x_2, \ldots, x_n)$ is represented by word j in the new dictionary, then let

$$e_j \triangleq \frac{1}{n} b\left(x^n\right) = \frac{1}{n} \sum_{i=1}^{n} b\left(x_i\right). \tag{3.2}$$

As before, the $N \times M$ forward channel transition matrix Q contains, in row j and column k, the conditional probability $Q_{k|j} = \Pr(Y^n = k | X^n = j)$. We also denote by p_j the probability that block $X^n = j$ is transmitted: $p_j = \Pr(X^n = j)$.

We can now rewrite the block mutual information in nats under this new notation.

$$\begin{aligned} I(X^n; Y^n) &= H(X^n) - H(X^n|Y^n) && (3.3) \\ &= \sum_{j,k} P_{X^n,Y^n}(j,k) \ln \left(\frac{P_{X^n|Y^n}(j|k)}{P_{X^n}(j)} \right) && (3.4) \\ &= \sum_{j,k} p_j Q_{k|j} \ln \frac{P_{j|k}}{p_j} = I(p; Q) && (3.5) \end{aligned}$$

where $P_{j|k}$ is an element of the reverse channel transition matrix from the output word $Y^n = k$ to the input word $X^n = j$. A reformulation of $C_n\left(\beta\right)$ is also possible:

$$C_n\left(\beta\right) = \max_{p\in\tau_n(\beta)} \frac{1}{n} I(p;Q) = \max_{p\in\tau_n(\beta)} \frac{1}{n}\sum_{j,k} p_j Q_{k|j} \ln \frac{P_{j|k}}{p_j}, \tag{3.6}$$

where $\tau_n(\beta)$ is redefined in an updated version of Equation (2.41) as

$$\tau_n(\beta) = \{p \in \mathbf{P}^N : \sum_j p_j e_j \leq \beta\},$$

where

$$\mathbf{P}^N \triangleq \left\{ p = [p_1, p_2, \ldots, p_N] \in \mathbf{R}^n : p_j \geq 0,\ \sum_{j=1}^{N} p_j = 1 \right\}. \tag{3.7}$$

Applying the results in Appendix B[2], this can be rewritten as

$$C_n\left(\beta\right) = \max_p \frac{1}{n}\left[\sum_{j,k} p_j Q_{k|j} \log \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s(\sum_j p_j e_j - \beta)\right] \tag{3.8}$$

where

$$\beta = \sum_j p_j^* e_j, \tag{3.9}$$

where p^* is the distribution on the input words that achieves the above maximum, and where s is the first derivative (slope) in nats per unit cost of $C_n\left(\beta\right)$ with respect to β.

Under this restatement of the problem, the n^{th} capacity-cost function is a maximization over all possible input vectors p, parameterized by s. A recursive algorithm for obtaining p found independently by Arimoto and Blahut, is given in the following theorem.

Theorem 3.1 Let $s \in [0, \infty)$ be given, and for any $p \in \mathbf{P}^N$ let

$$c_j(p) = \exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j\right). \tag{3.10}$$

[2] Appendix B contains proofs for all subsequent theorems in this section, using this notation. In [8] some proofs were omitted and others were cited from texts.

Then if p^0 is any probability vector in $\mathbf{P}^N$ with all components strictly positive, the sequence of vectors resulting from

$$p_j^{r+1} = p_j^r \frac{c_j^r}{\sum_j p_j^r c_j^r} \tag{3.11}$$

has the properties that

$$\frac{1}{n} I(p^r; Q) \to C_n\left(\beta_s\right), \qquad \text{as } r \to \infty, \tag{3.12}$$

$$e(p^r) \to \beta_s, \qquad \text{as } r \to \infty, \tag{3.13}$$

where β_s is the average per letter cost of the point parametrized by s, and $I(\cdot;\cdot)$ and $C_n\left(\cdot\right)$ are measured in nats.

The stopping rule for the algorithm bounds $C_n\left(\beta\right)$ above and below, using convergent functions of c_j.

Theorem 3.2 Let the left derivative of a point on $C_n\left(\beta\right)$ be specified by parameter s. Assuming p is any probability vector, we let

$$c_j = \exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j\right). \tag{3.14}$$

Then, for cost $\beta = \sum_j p_j e_j$,

a)

$$C_n\left(\beta\right) \geq C_n^L\left(\beta\right) \triangleq s\beta + \sum_j p_j \ln c_j, \qquad \text{in nats,} \tag{3.15}$$

b)

$$C_n\left(\beta\right) \leq C_n^U\left(\beta\right) \triangleq s\beta + \ln \max_j c_j, \qquad \text{in nats.} \tag{3.16}$$

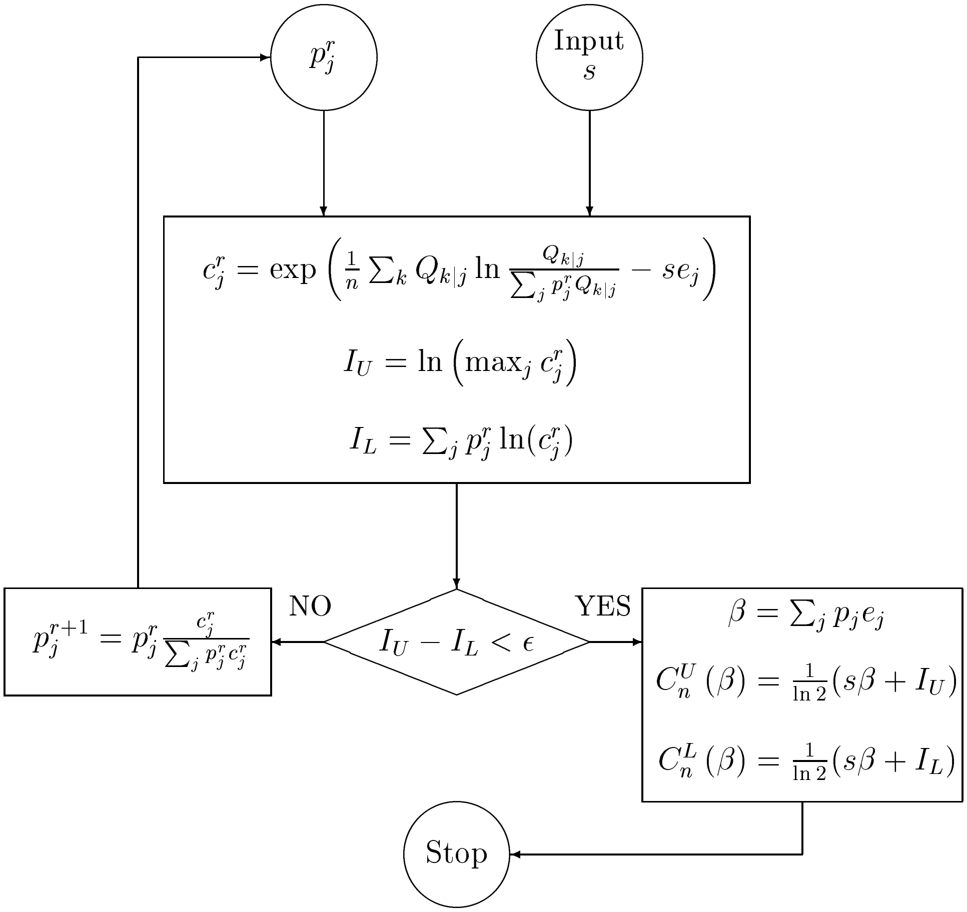

p_j^r
Input s
$c_j^r = \exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j^r Q_{k|j}} - s e_j\right)$
$I_U = \ln\left(\max_j c_j^r\right)$
$I_L = \sum_j p_j^r \ln(c_j^r)$
NO
$I_U - I_L < \epsilon$
YES
$p_j^{r+1} = p_j^r \frac{c_j^r}{\sum_j p_j^r c_j^r}$
$\beta = \sum_j p_j e_j$
$C_n^U(\beta) = \frac{1}{\ln 2}(s\beta + I_U)$
$C_n^L(\beta) = \frac{1}{\ln 2}(s\beta + I_L)$
Stop

Figure 3.1: Constrained Capacity Algorithm from [8] for $C_n(\beta)$ in bits.

Notice that the n^{th} capacity-cost function algorithm is only valid for a discrete block memoryless channel with conditional probability mass function $P_{Y^n|X^n}(y^n|x^n)$, where each input or output block is treated as a single channel symbol. Within a block $y^n = x^n + z^n$, for instance, the inter-symbol dependence of the noise letters is accounted for explicitly, but the inter-word dependence of the noise process is not taken into account. Let $(\beta, C_n^L(\beta))$ be a pair determined by the algorithm for slope s. Now recall that $(\beta, C_n^L(\beta))$ is within ϵ of the n^{th} capacity-cost function in nats.

$$C(\beta) \geq C_n(\beta) \geq C_n^L(\beta). \tag{3.17}$$

We can see then, that Blahut's algorithm provides a block length lower bound to

$C_n(\beta)$ (and therefore, $C(\beta)$) for an additive channel with a Markov noise process.

3.1.1.1 Implementation Issues

As can be seen from Figure 3.1, there are a number of fixed parameters that must be set in addition to the input parameter s, before the algorithm is operational.

- *Specifying the Conditional Probability Distribution Matrix Q*

  For additive channels, we see that a non-zero entry at $Q(y^n, x^n)$ is equal to $P_{Z^n}(z^n = y^n - x^n)$. We have already shown how to determine the stationary distribution of $\{Z_i\}_{i=1}^{\infty}$ given an ergodic time invariant Markov process with state transition matrix $\mathbf{\Pi}$. The program implemented in this thesis allows the user to specify the input alphabet size q, the order k of the Markov chain, and a Markov transition probability matrix $\mathbf{\Pi}$ on states $(z_{n-k}, \ldots, z_{n-1}) \in \mathcal{A}_q^k$. This matrix can be specified either during execution or from an input file.

- *Stopping Criteria*

  The halting rule computes an upper bound $C_n^U(\beta)$ and lower bound $C_n^L(\beta)$ on the n^{th} capacity-cost function $C_n(\beta)$ after each iteration of the algorithm. If the difference between the bounds is less than a fixed amount ϵ, then $C_n(\beta) - C_n^L(\beta) < \epsilon$ also. The program implements an additional check as well. Setting $\epsilon = 10^{-6}$, the program executes until $|I_L - I_U| < \epsilon$ or the number of iterations equals 10 000. For mod q channels, for instance, each loop of the algorithm requires $2q^{2n-1}$ additions, $2q^{2n} + q^n$ multiplications, $2q^n$ divisions, and q^n subtractions, logs and exponents. This has the potential to introduce numerical error into the loop. For this reason, the algorithm is stopped after 10 000 iterations and the result of the subtraction is returned. If the difference is an order of magnitude smaller than what is required, the point has acceptable error for plotting. There have been few instances thus far, where the error is greater than $5 \cdot 10^{-5}$ after 10 000 iterations. Such instances are expected to be more probable with increasing block length and alphabet size. The increases in the capacity-cost function with feedback are all at least on the order of 10^{-3}, so accuracy of our results is assured.

- *Cost Function*

  For all results presented in this thesis, the cost function used is a linear cost constraint – $b(x_i) = x_i$, where $x_i \in \{0, 1, \ldots, q-1\}$ – on the input letters. The C++ code as written, however, supports any cost constraint on either letters or blocks. In the case of the binary channel, the linear cost constraint is identical to the power cost constraint – $b(x_i) = x_i^2$.

- *Parameter s*

  Our choice of parameterization over input blocks of length n requires that we specify the slope s, in nats per unit cost, of $C_n(\beta)$ at a desired cost β. The program inputs various s and then scales the output into bits by multiplication of $C_n^L(\beta)$ and $C_n^U(\beta)$ by $\frac{1}{\ln 2}$. These values are plotted by gnuplot for various Markov chains, alphabet sizes and block lengths.

Used alone, Blahut's lower bound is sometimes loose for high values of β. The next section discusses another bound on binary channels, that squeezes the envelope for β close to $\beta_{max}^{(n)}$.

3.1.2 Mrs. Gerber's Lemma

One advantage of Blahut's lower bound for additive Markov channels is its agreement with $C(\beta)$ at β_{min}, i.e. $C_n^L(\beta_{min}) = C(\beta_{min})$. If we were to use an additional lower bound $C^{L_1}(\beta)$ whose performance at large β was superior, the maximization

$$C_n^{L*}(\beta) = \max(C_n^L(\beta), C^{L_1}(\beta)) \tag{3.18}$$

over both lower bounds would help form a much tighter envelope on the capacity-cost function at block length n. One such additional bound, due to Shamai and Wyner [20], and valid for mod 2 additive *binary* channels only, is a direct analogy of the Entropy Power Inequality.

Theorem 3.3 ([20]) Let $\{X_i\}_{i=1}^{\infty}$ and $\{Z_i\}_{i=1}^{\infty}$ be independent stationary binary random sources with entropy rates $H(X_\infty)$ and $H(Z_\infty)$ respectively. Denote the binary entropy function as

$$h(\phi) = -\phi \log \phi - (1-\phi)\log(1-\phi), \tag{3.19}$$

where $0 \leq \phi \leq 1$. Let $\sigma(X) = h^{-1}(H(X_\infty))$ and let $\sigma(Z) = h^{-1}(H(Z_\infty))$, where

$$h^{-1}(\omega) = min\{\phi : \ \omega = h(\phi)\}. \tag{3.20}$$

Let $Y_i = X_i \oplus Z_i$. Then

$$\sigma(Y) \geq \sigma(X) * \sigma(Z), \tag{3.21}$$

where $\sigma(Y) = h^{-1}(H(Y_\infty))$, and $a * b \overset{\triangle}{=} a(1-b) + (1-a)b$.

For the case when $\{Z_i\}_{i=1}^{\infty}$ is independent identically distributed (iid), this theorem reduces to "Mrs. Gerber's Lemma" due to Wyner and Ziv [27]. We will now state (with slight modifications) the proof of the above binary analog to the Entropy-Power Inequality [20].

Proof of Theorem 3.3 We express the conditional entropy of output Y_{n+1} as

$$\begin{aligned} H(Y_{n+1}|Y^n) &\geq H(Y_{n+1}|Y^n, X^n, Z^n) && (3.22) \\ &= H(Y_{n+1}|X^n, Z^n) && (3.23) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x^n \in \mathcal{A}_2^n} P_{X^n}(x^n) \sum_{z^n \in \mathcal{A}_2^n} P_{Z^n}(z^n) \\
&\qquad \cdot H(Y_{n+1}|X^n = x^n,\ Z^n = z^n),
\end{aligned} \tag{3.24}
$$

where the inequality results from the fact that conditioning reduces entropy, the first equality follows from $Y^n = X^n \oplus Z^n$, and the second equality is due to the independence of $\{X_i\}_{i=1}^{\infty}$ and $\{Z_i\}_{i=1}^{\infty}$. Let us now expand the final term of (3.24) as

$$
H(Y_{n+1}|x^n, z^n) = -\sum_{i \in \{0,1\}} \Pr(Y_{n+1} = i|x^n, z^n) \log \Pr(Y_{n+1} = i|x^n, z^n). \tag{3.25}
$$

We can remove all reference to Y_{n+1} from the right hand side of Equation (3.24) by rewriting $\Pr(Y_{n+1} = 1|x^n, z^n)$ as

$$
\Pr(Y_{n+1} = 1|x^n, z^n) = \hat{\alpha}(x^n)(1 - \hat{\gamma}(z^n)) + (1 - \hat{\alpha}(x^n))\hat{\gamma}(z^n) = \hat{\alpha}(x^n) * \hat{\gamma}(z^n), \tag{3.26}
$$

where

$$
\begin{aligned}
\hat{\alpha}(x^n) &\triangleq P_{X_{n+1}|X^n}(x_{n+1} = 1|x^n), &(3.27)\\
\hat{\gamma}(z^n) &\triangleq P_{Z_{n+1}|Z^n}(z_{n+1} = 1|z^n), &(3.28)
\end{aligned}
$$

and then substituting in (3.25) yields

$$
H(Y_{n+1}|X^n = x^n,\ Z^n = z^n) = h(\hat{\alpha}(x^n) * \hat{\gamma}(z^n)). \tag{3.29}
$$

Now define

$$
\begin{aligned}
\alpha(x^n) &\triangleq \min[\hat{\alpha}(x^n), 1 - \hat{\alpha}(x^n)] &(3.30)\\
\gamma(z^n) &\triangleq \min[\hat{\gamma}(z^n), 1 - \hat{\gamma}(z^n)] &(3.31)
\end{aligned}
$$

and notice that the binary entropy is not affected:

$$
h(\hat{\alpha}(x^n) * \hat{\gamma}(z^n)) = h(\alpha(x^n) * \gamma(z^n)). \tag{3.32}
$$

Also

$$\alpha(x^n) = h^{-1}(H(X_{n+1}|X^n = x^n)), \tag{3.33}$$

$$\gamma(z^n) = h^{-1}(H(Z_{n+1}|Z^n = z^n)). \tag{3.34}$$

Substituting these observations into (3.24) yields

$$\begin{aligned} H(Y_{n+1}|Y^n) \geq & \sum_{z^n \in \mathcal{A}_2^n} P_{Z^n}(z^n) \\ & \sum_{x^n \in \mathcal{A}_2^n} P_{X^n}(x^n) h\left[\gamma(z^n) * h^{-1}(H(X_{n+1}|X^n = x^n))\right]. \end{aligned} \tag{3.35}$$

From Lemma 2 of [27] we know that the function

$$f(u) = h(p_0 * h^{-1}(u)), \quad 0 \leq u \leq 1, \tag{3.36}$$

with $p_0 \in (0, \frac{1}{2}]$ fixed, is convex in u. The proof itself is lengthy, but demonstrates that the convexity of $p_0 * h^{-1}(u)$ is sufficient to overcome the concavity of $h(\cdot)$. Hence if we apply Jensen's inequality to the inner sum in (3.35), we obtain

$$\begin{aligned} & \sum_{x^n} P_{X^n}(x^n) \; h\left[\gamma(z^n) * h^{-1}(H(X_{n+1}|X^n = x^n))\right] \\ & \geq h\left[\gamma(z^n) * h^{-1}(H(X_{n+1}|X^n))\right], \end{aligned} \tag{3.37}$$

since

$$\sum_{x^n} P_{X^n}(x^n) H(X_{n+1}|X^n = x^n) = H(X_{n+1}|X^n). \tag{3.38}$$

Summing over z^n and applying Jensen's inequality again, for $H(X_{n+1}|X^n)$ fixed, results in

$$H(Y_{n+1}|Y^n) \geq h\left[h^{-1}(H(X_{n+1}|X^n)) * h^{-1}(H(Z_{n+1}|Z^n))\right]. \tag{3.39}$$

Letting $n \to \infty$ and taking the limit gives us the useful form

$$H(Y_\infty) = \lim_{n \to \infty} H(Y_{n+1}|Y^n) \geq h\left[h^{-1}(H(X_\infty)) * h^{-1}(H(Z_\infty))\right]. \tag{3.40}$$

Taking h^{-1} of both sides completes the proof. □

We apply this result to our bounding problem. Let $P_X(x_i = 1) \triangleq \hat{\alpha}$ be the marginal distribution of an iid input process such that $E[b(X_i)] = \beta$, then

$$C(\beta) \geq C^{L_1}(\beta) \triangleq h\left[\alpha * h^{-1}(\gamma)\right] - H(Z_\infty), \tag{3.41}$$

where $h(\cdot)$ is the binary entropy function, $a*b \triangleq a(1-b)+(1-a)b$, $\alpha \triangleq \min\{\hat{\alpha}, 1-\hat{\alpha}\}$ and $\gamma \triangleq \min\{H(Z_\infty), 1 - H(Z_\infty)\}$. We have already explained how to compute the entropy rate of the noise and so this generalization of Mrs. Gerber's Lemma can be readily applied to our binary $\oplus$ examples.

3.2 An Upper Bound to $C(\beta)$

Consider a discrete channel with memory, with common input and noise q-ary alphabet and an output alphabet described by the following equation: $Y_i = X_i \oplus Z_i$, for $i \in \{1, 2, \ldots\}$ where:

- $\oplus$ represents modulo q addition.
- The random variables X_i, Z_i and Y_i are respectively the input, noise and output of the channel.
- $\{X_i\} \perp \{Z_i\}$, i.e. the input and noise sequences are independent from each other.
- The noise process $\{Z_i\}_{i=1}^{\infty}$ is stationary.

The channel is non-symmetric for $\beta < \beta_{max}$. Thus the formula of $C(\beta)$ given by Equation (2.42) will not have a closed form. We will then use the results of Alajaji [2] to derive an upper bound to $C(\beta)$.

In [26], Wyner and Ziv derived a lower bound to the rate-distortion function ($R(D)$) of stationary sources:

$$R(D) \geq R_1(D) - \mu_1, \tag{3.42}$$

where

- $R_1(D)$ is the rate-distortion function of an *associated memoryless source* with distribution equal to the marginal distribution $P_X(\cdot)$ of the stationary source.
- $\mu_1 \triangleq H(X_1) - H(X_\infty)$, is the amount of memory in the source. $H(X_1)$ is the entropy of the associated memoryless source with distribution $P_X^{(1)}(\cdot)$ and $H(X_\infty)$ is the entropy rate of the original stationary source.

This lower bound was later tightened by Berger [5]:

$$R(D) \geq R_n(D) - \mu_n \geq R_1(D) - \mu_1, \tag{3.43}$$

where $R_n(D)$ is the nth 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 $C(\beta)$, Alajaji proved an equivalent upper bound to the capacity-cost function of a discrete additive channel.

Theorem 3.4 ([2]) Consider a discrete channel with additive stationary noise process $\{Z_i\}_{i=1}^{\infty}$. Let $P_{Z^n}(\cdot)$ denote the n-fold probability distribution function of the noise process. Then for $N = kn$, $k, n \in \{1, 2, \ldots\}$,

$$C_N(\beta) \leq C_n(\beta) + \Delta_{nN} \leq C_1(\beta) + \Delta_{1N}, \tag{3.44}$$

where

- $C_n(\beta)$ is the n-fold capacity-cost function of the channel as defined in (2.43).

- $C_1(\beta)$ 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 $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_{q-1})$ of the stationary noise process.

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

For the sake of completeness, we reproduce the proof of the above theorem.

Proof of Theorem 3.4 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

$$I(X^N; Y^N) \le \sum_{i=1}^{k} I(X^n_{(i)}; Y^n_{(i)}) + N\Delta_{nN}, \tag{3.45}$$

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

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

and

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

Proving the above inequality goes as follows:

$$\begin{aligned}
&\sum_{i=1}^{k} I(X^n_{(i)}; Y^n_{(i)}) + N\Delta_{nN} - I(X^N; Y^N) \\
&\quad = \sum_{i=1}^{k} \left[H(Y^n_{(i)}) - H(Y^n_{(i)}|X^n_{(i)})\right] + \frac{N}{n}H(Z^n) - H(Z^N) \\
&\quad - H(Y^N) + H(Y^N|X^N) \qquad (3.46) \\
&\quad = \sum_{i=1}^{k} \left[H(Y^n_{(i)}) - H(Z^n_{(i)})\right] + kH(Z^n) - H(Z^N) - H(Y^N) + H(Z^N) \qquad (3.47)
\end{aligned}$$

$$= \sum_{i=1}^{k} H(Y_{(i)}^n) - H(Y^N) \tag{3.48}$$

$$= \sum_{i=1}^{k} H(Y_{(i)}^n) - \sum_{i=1}^{k} H(Y_{(i)}^n | Y_{(i-1)}^n, Y_{(i-2)}^n, \ldots, Y_{(1)}^n) \tag{3.49}$$

$$\geq \sum_{i=1}^{k} H(Y_{(i)}^n) - \sum_{i=1}^{k} H(Y_{(i)}^n) \tag{3.50}$$

$$= 0, \tag{3.51}$$

where, the first equality is true by the definitions of $N\Delta_{nN}$ and $I(X^n; Y^n)$. Both the third and final steps are algebraic manipulations and the fourth is merely the chain rule for entropy. The inequality arises since conditioning reduces entropy, leaving the second equality as the only place to limit the scope of our expression. Therefore, the inequality is valid whenever the noise entropy is exactly equal to the conditional entropy of the input-output process:

$$H(Y|X) = H(Z).$$

This is true for all invertible noise processes, including those in our examples.

Let $P_{X^N}(x^N) \in \tau_N(\beta)$ where $\tau_N(\beta)$ is described in (2.41). For this input distribution, we denote $\beta_i \triangleq \frac{1}{n} E\left[b\left(X_{(i)}^n\right)\right]$ for $i = 1, 2, \ldots, k$; thus

$$\frac{1}{k}\sum_{i=1}^{k} \beta_i = \frac{1}{N} E\left[b\left(X^N\right)\right] \leq \beta. \tag{3.52}$$

By (3.45), we obtain with this $P_{X^N}(x^N)$:

$$\frac{1}{N}\ I(X^N; Y^N) \leq \frac{1}{N}\ \sum_{i=1}^{k} I(X_{(i)}^n; Y_{(i)}^n) + \Delta_{nN}; \tag{3.53}$$

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

$$\frac{1}{N}\ I(X^N; Y^N) \leq \frac{1}{k}\ \sum_{i=1}^{k} C_n\left(\beta_i\right) + \Delta_{nN}. \tag{3.54}$$

By concavity of $C_n(\cdot)$, we have

$$\frac{1}{k}\sum_{i=1}^{k} C_n(\beta_i) \leq C_n\left(\frac{1}{k}\sum_{i=1}^{k}\beta_i\right) \tag{3.55}$$

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

$$\frac{1}{N}\, I(X^N;Y^N) \leq C_n(\beta) + \Delta_{nN}, \tag{3.56}$$

or

$$\max_{P_{X^N}(x^N)\in\tau_N(\beta)} \frac{1}{N}\, I(X^N;Y^N) = C_N(\beta) \leq C_n(\beta) + \Delta_{nN}. \tag{3.57}$$

Thus the first inequality in (3.44) is proved. To prove the second inequality in (3.44), 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 Equation (3.44) and letting $n = 1$. □

Using Equations (3.44) and (2.42), we obtain the following tight upper bound on $C(\beta)$.

Corollary 3.1 ([2]) Consider the channel described in Theorem 3.4, with the assumption that the noise process is stationary. Then

$$C(\beta) \leq C_n(\beta) + M_n \leq C_1(\beta) + M_1, \tag{3.58}$$

where

- $C_n(\beta)$ and $C_1(\beta)$ are as defined in Theorem 3.4.
- $M_n \triangleq \Delta_{n\infty} = \frac{1}{n}H(Z^n) - H(Z_\infty)$, and $M_1 \triangleq \Delta_{1\infty} = H(Z) - H(Z_\infty)$ is the amount of memory in the noise process.

The bound given above is asymptotically tight with n, and we can see that by applying Blahut's algorithm to determine the upper bound $C_n^U(\beta)$ on $C_n(\beta)$,

$$C_n^{ub}(\beta) \triangleq C_n^U(\beta) + M_n \geq C_n(\beta_s) + M_n \geq C(\beta), \tag{3.59}$$

where s is the slope of both the upper and lower bounds at β_s. This will become apparent through experimentation as well. Blahut's lower bound will converge to Alajaji's upper bound since $M_n \to 0$ as $n \to \infty$.

3.3 Numerical Results

In this section we implement the bounding techniques of this chapter in C++ code on a SUN ULTRA SPARC 1 workstation. It soon becomes obvious that the computation time required for the evaluation of $C_n(\beta)$ grows extremely fast both in the block length n of the input vectors and in the input alphabet size q. While this effect is being explored we also witness the convergence of the envelope between the Blahut-Arimoto, and the Alajaji bounds. This distance becomes important in Chapter 4 where we use the Blahut-Arimoto lower bound on feedback channels and the Alajaji upper bound on non-feedback channels (both responding to the same noise process) to show improvement of $C(\beta)$ with feedback.

Let us now compute these bounds for the examples introduced in Chapter 2.

Example 3.1 *Binary Alphabet Channel with 1^{st} Order Markov Noise.* We wish to evaluate the memory element, M_n, and $C_n^L(\beta)$ for the mod 2 channel with first order noise state transitions give by

$$\mathbf{\Pi} = \begin{bmatrix} 1-\alpha & \alpha \\ \varrho & 1-\varrho \end{bmatrix} = \begin{bmatrix} .95 & .05 \\ .20 & .80 \end{bmatrix} \tag{3.60}$$

and with stationary distribution given by

$$[\pi_0, \pi_1] = \left[\frac{\varrho}{\alpha+\varrho}, \frac{\alpha}{\alpha+\varrho}\right] = [.8, .2]. \tag{3.61}$$

Applying the results of Equation (2.24) in Theorem 2.5, we can express $H(Z_\infty)$ as $H(Z_2|Z_1)$. This yields

$$\begin{aligned} M_n &= \frac{1}{n}H(Z^n) - H(Z_2|Z_1) && (3.62) \\ &= -\frac{1}{n}\sum_{z^n=0}^{2^n-1} P_{Z^n}(z^n)\log P_{Z^n}(z^n) + \sum_{l=0}^{1}\sum_{m=0}^{1}\pi_l \mathbf{\Pi}_{l,m}\log \mathbf{\Pi}_{l,m} && (3.63) \\ &= -\frac{1}{n}\sum_{z^n=0}^{2^n-1} P_{Z^n}(z^n)\log P_{Z^n}(z^n) - \frac{\varrho h(\alpha) + \alpha h(\varrho)}{\alpha+\varrho}, && (3.64) \end{aligned}$$

where

$$P_{Z^n}(z^n) = \pi_{z_1}\prod_{i=1}^{n-1}\mathbf{\Pi}_{z_i,z_{i+1}}. \qquad (3.65)$$

This closed form solution to the memory remaining after blocking n input symbols is a key aspect to our analysis. Table 3.1 shows the decrease in M_n for a number of different binary Markov matrices $\mathbf{\Pi}$.

(α, ϱ)	π_1	$n = 1$	$n = 2$	$n = 3$	$n = 6$	$n = 9$	$n = 12$
$(.05, .2)$	.2	.348425	.174212	.116142	.058071	.038714	.029035
$(.1, .4)$	.2	.152542	.076271	.050847	.025424	.016949	.012712
$(.15, .6)$	.2	.039866	.019933	.013289	.006644	.004430	.003322
$(.175, .7)$	.2	.010457	.005229	.003486	.001743	.001162	.000871
$(.225, .9)$	.2	.012775	.006387	.004258	.002129	.001419	.001065
$(.05, .45)$	.1	.111961	.055980	.037320	.018660	.012440	.009330
$(.05, .075)$	.4	.645388	.322694	.215129	.107565	.071710	.053782

Table 3.1: First-order binary Markov noise: M_n in bits for different values of n and (α, ϱ); $\pi_1 = \frac{\alpha}{\alpha+\varrho}$.

This memory element M_n plus the difference $(C_n^U(\beta) - C_n^L(\beta))$ distinguishes Blahut's lower bound $C_n^L(\beta)$ from Alajaji's upper bound $C_n^{ub}(\beta)$. Computation of both bounds requires forming the corresponding memoryless channel $Y^n = X^n \oplus Z^n$ with channel transition matrix $Q = [P_{Y^n|X^n}(y^n|x^n)]$. Let us now explicitly demonstrate how this is done by the simulation program.

If $n = 5$, then let $x^n = (x_1, x_2, \ldots, x_n) = (1, 0, 0, 1, 1)$ be the input word to the channel. This is just one of 32 possible input words for the binary channel with block length 5. At the receiver, any output could be observed. For the moment, let $y^n = (1, 0, 1, 0, 1)$ be the output of the channel, and let us decipher what occurred.

$$\begin{aligned} y^n &= x^n \oplus z^n, && (3.66) \\ (1,0,1,0,1) &= (1,0,0,1,1) \oplus (z_1, z_2, z_3, z_4, z_5), && (3.67) \\ (z_1, z_2, z_3, z_4, z_5) &= (1,0,1,0,1) \ominus (1,0,0,1,1), && (3.68) \\ (z_1, z_2, z_3, z_4, z_5) &= (0,0,1,1,0). && (3.69) \end{aligned}$$

Note that in the binary case $\oplus$ is equivalent to $\ominus$, but this is not true for a general q-ary channel. The noise sequence is uniquely determined by the inputs and outputs for additive channels, and we can determine this probability as we did above. Using the chain rule for probability and the Markovity of $\{Z_i\}_{i=1}^{\infty}$;

$$P_{Z^5}(0,0,1,1,0) = \pi_0 \mathbf{\Pi}_{0,0} \mathbf{\Pi}_{0,1} \mathbf{\Pi}_{1,1} \mathbf{\Pi}_{1,0}. \quad (3.70)$$

To facilitate the indexing of these probability mass functions in the computer program we label them as $j = z_1 q^{n-1} + \cdots + z_n q^0$. For $q = 2$ this allows us to write

$$x^n = 19, \quad y^n = 21, \quad z^n = 6, \quad (3.71)$$

in the above equation. Therefore,

$$\begin{aligned} P_{Y^5|X^5}(21/19) &= P_{Y^5|X^5}((1,0,1,0,1)|(1,0,0,1,1)) && (3.72) \\ &= P_{Z^5}((0,0,1,1,0)) && (3.73) \\ &= P_{Z^5}(6). && (3.74) \end{aligned}$$

Note also that the computer defines

$$P_{Y^5|X^5}((0,0,1,1,0)|(0,0,0,0,0)) = P_{Z^5}(6). \tag{3.75}$$

The computer program forms the matrix Q in exactly this fashion: by taking the block probabilities P_{Z^n} and cycling through all possible $x^n \in \mathcal{A}_q^n$.

All that remains for the implementation of Blahut's algorithm is a choice of cost function. For simplicity, we chose a linear cost function that assigns to each input letter its nominative cost. For binary alphabets the costs are 0 and 1. For a ternary alphabet the costs are 0, 1 and 2, and so on.

Figures 3.2 and 3.3 demonstrate the convergence (as n increases) of Blahut's bound to $C\left(\beta\right)$ from below, and the convergence of Alajaji's bound from above, respectively. In Figure 3.4, we make use of Mrs. Gerber's Lemma ($C^{L_1}\left(\beta\right)$) to improve the lower bound, and also to show the tightness of the envelope on the capacity-cost function. Unfortunately, this additional bound is only valid for binary noise processes. We can use our example to illustrate the computation of Equation (3.41). Using the linear cost constraint,

$$\begin{aligned} \beta = E\left[b(X_i)\right] &= b(0)P_X(0) + b(1)P_X(1) && (3.76) \\ \beta &= P_X(1) && (3.77) \end{aligned}$$

for $\beta \leq \beta_{max}^{(n)} = \beta_{max} = \frac{1}{2}$. Substituting into (3.41) yields

$$\begin{aligned} C\left(\beta\right) &= \lim_{n\to\infty} \max_{P_{X^n}(x^n)\in\tau_n(\beta)} \frac{1}{n}H(Y^n) - H(Z_\infty) && (3.78) \\ &\geq h\left[\beta * h^{-1}(\gamma)\right] - H(Z_\infty), && (3.79) \end{aligned}$$

where $\gamma = \min\{H(Z_\infty), 1 - H(Z_\infty)\}$ is given by

$$H(Z_\infty) = \frac{\varrho h(\alpha) + \alpha h(\varrho)}{\alpha + \varrho}.$$

Example 3.2 *Ternary Alphabet Channel with* 1^{st} *Order Markov Noise.* As in the previous example we calculate the memory element, M_n, and $C_n^L(\beta)$. The noise process that corrupts our mod 3 channel has state transition probabilities given by

$$\mathbf{\Pi} = \begin{bmatrix} 1-\alpha_1-\alpha_2 & \alpha_1 & \alpha_2 \\ \varrho_0 & 1-\varrho_0-\varrho_2 & \varrho_2 \\ \gamma_0 & \gamma_1 & 1-\gamma_0-\gamma_1 \end{bmatrix} = \begin{bmatrix} .8 & .15 & .05 \\ .3 & .5 & .2 \\ .3 & .1 & .6 \end{bmatrix} \tag{3.80}$$

and with stationary distribution given by

$$\boldsymbol{\pi} = [P_Z(0), P_Z(1), P_Z(2)] = [.6, .2167, .1833]. \tag{3.81}$$

In keeping with the previous example, if $n = 5$ and $x^n = (x_1, x_2, \ldots, x_n)$ is an input word to the channel, we can index it as $j = 3^4 x_1 + \cdots + 3^1 x_4 + x_5$. Recall that there were 32 possible words at this stage in the binary example compared with 243 words here. It is for this reason that we limit n to 5 in our graph for this example. Figure 3.5 shows that Blahut's algorithm can be successfully applied to a ternary alphabet as well, but with reduced effectiveness due to computation time requirements.

Example 3.3 *Quaternary Alphabet Channel with* 1^{st} *Order Markov Noise.* Calculation of the memory element, M_n, and $C_n^L(\beta)$ is sufficient to bound $C(\beta)$ from both sides. The cost is still a linear cost constraint, $b(i) = i$, for $i \in \{0, 1, 2, 3\}$, and the noise process across our mod 4 channel has state transition probabilities given by

$$\mathbf{\Pi} = \begin{bmatrix} \alpha_{00} & \alpha_{01} & \alpha_{02} & \alpha_{03} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{30} & \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} = \begin{bmatrix} .8 & .1 & .05 & .05 \\ .3 & .5 & .1 & .1 \\ .3 & .05 & .05 & .6 \\ .5 & .05 & .4 & .05 \end{bmatrix}, \tag{3.82}$$

where α_{jk} is the conditional probability $P_{Z_i|Z_{i-1}}(k|j)$. The stationary probabilities for this particular example are

$$\boldsymbol{\pi} = [P_Z(0), P_Z(1), P_Z(2), P_Z(3)] = [.6441, .1495, .0961, .1103]. \tag{3.83}$$

The graph in Figure 3.6 shows the distance between the upper and lower bounds using block length $n = 4$. The number of iterations of Blahut's algorithm required for each point increases in n and q as does the number of computations within each iteration.

Example 3.4 *Binary Alphabet Channel with 2^{nd} Order Markov Noise.* In this example we are again able to use Mrs. Gerber's Lemma to improve the lower bound provided by Blahut's algorithm, $C_n^L(\beta)$. The upper bound, $C_n^U(\beta)$, is computed by adding the memory element M_n. The complication of higher order Markov noise is not great. Rather than simply calculating the stationary distribution on the noise symbols from the state transition matrix, we must assume this stationary distribution for an initial state. Then for the first $k-1$ noise samples, we use $\mathbf{\Pi}$ to generate the required probabilities.

We use the following transition matrix in this example:

$$\mathbf{\Pi} = \begin{bmatrix} 1-\alpha_{00} & \alpha_{00} & 0 & 0 \\ 0 & 0 & 1-\alpha_{01} & \alpha_{01} \\ \alpha_{10} & 1-\alpha_{10} & 0 & 0 \\ 0 & 0 & \alpha_{11} & 1-\alpha_{11} \end{bmatrix} = \begin{bmatrix} .95 & .05 & 0 & 0 \\ 0 & 0 & .25 & .75 \\ .90 & .10 & 0 & 0 \\ 0 & 0 & .1 & .9 \end{bmatrix}. \tag{3.84}$$

The stationary distributions on the noise states $s_i = (z_{i-2}, z_{i-1})$, where $z_i \in \{0, 1\}$ for all $j \in \{1, 2, \ldots\}$, are given by[3]

$$\boldsymbol{\pi} = [\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}] = [.6\overline{54}, .0\overline{36}, .0\overline{36}, .2\overline{72}], \tag{3.85}$$

[3] The overline on a decimal number implies repeating digits unto infinity (i.e., $0.0\overline{28} = 0.028282828282\ldots$).

and the initial stationary distribution on the letters is

$$[P_Z(0), P_Z(1)] = [.6\overline{90}, .3\overline{09}], \tag{3.86}$$

respectively. The computation of these values is explained in Example 2.4.

Applying the results of Equation (2.24) in Theorem 2.5, we can express $H(Z_\infty)$ as $H(S_2|S_1)$. Thus

$$\begin{aligned} M_n &= \frac{1}{n}H(Z^n) - H(S_2|S_1) && (3.87) \\ &= -\frac{1}{n}\sum_{z^n=0}^{2^n-1} P_{Z^n}(z^n)\log P_{Z^n}(z^n) + \sum_{l,m\in\mathcal{A}_2^2} \pi_l \mathbf{\Pi}_{l,m}\log\mathbf{\Pi}_{l,m}, && (3.88) \end{aligned}$$

where

$$\begin{aligned} P_{Z^n}^{(n)}(z^n) &= \pi_{z_1,z_2}\prod_{i=3}^{n}\Pr(Z_i{=}z_i|Z_{i-1}{=}z_{i-1}, Z_{i-2}{=}z_{i-2}) && (3.89) \\ &= \pi_{z_1,z_2}\prod_{i=3}^{n}\mathbf{\Pi}[(z_{i-2}z_{i-1}),(z_{i-1}z_i)] && (3.90) \end{aligned}$$

given the noise syndrome $z^n = (z_1, z_2, \ldots, z_n)$, $n \geq 2$.

These terms and linear cost constraints are used as parameters in Blahut's algorithm and Mrs. Gerber's Lemma to bound $C(\beta)$ in the range of $0 \leq \beta \leq 1/2$ (see Figure 3.7).

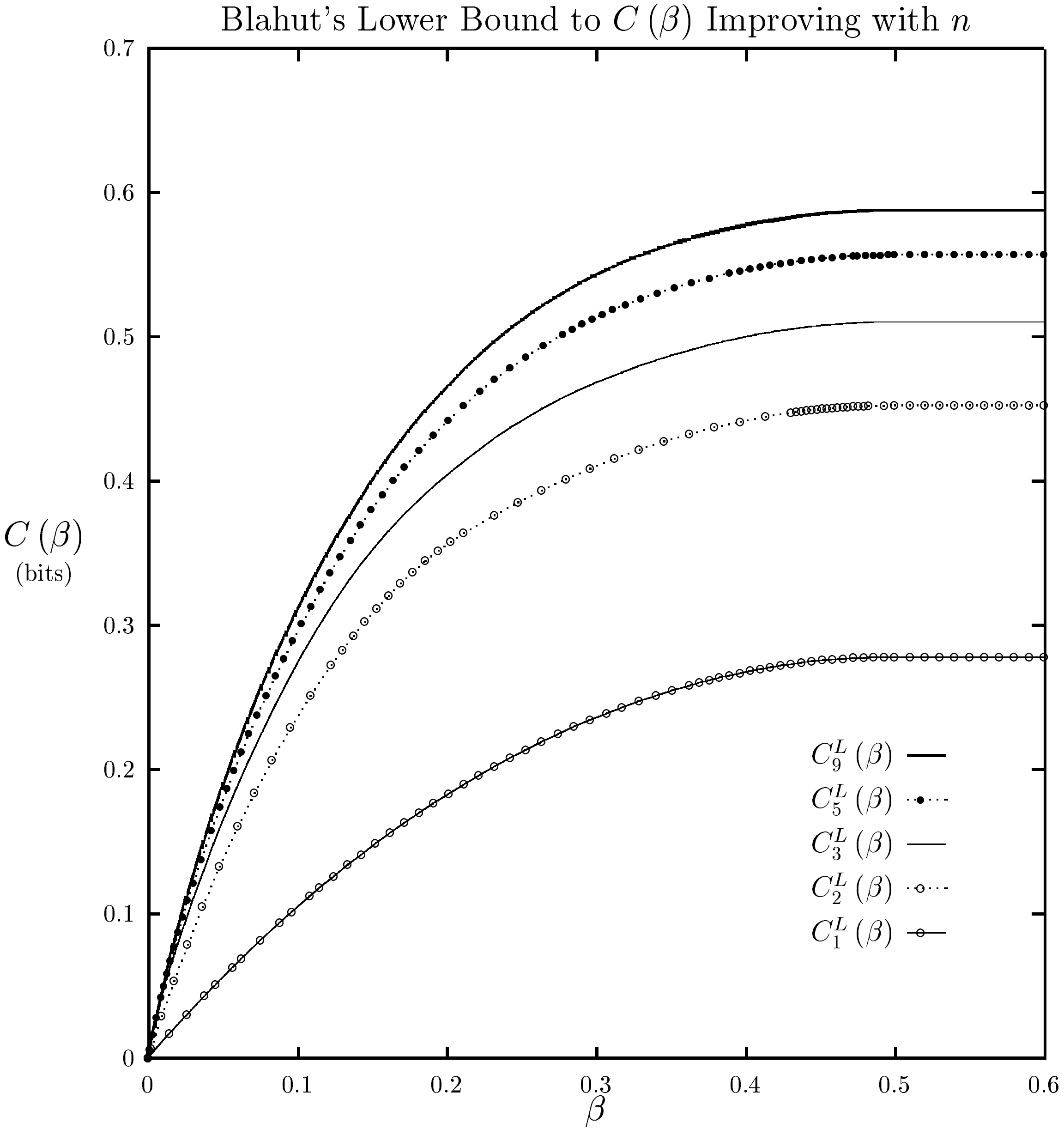


Figure 3.2: $C_n^L(\beta)$ for $n = 1, 2, 3, 5$ and 9 using 1st order binary Markov noise with $\alpha = .05$ and $\varrho = 0.2$, $\beta_{max} = 1/2$. Cost function: $b(0) = 0, b(1) = 1$.

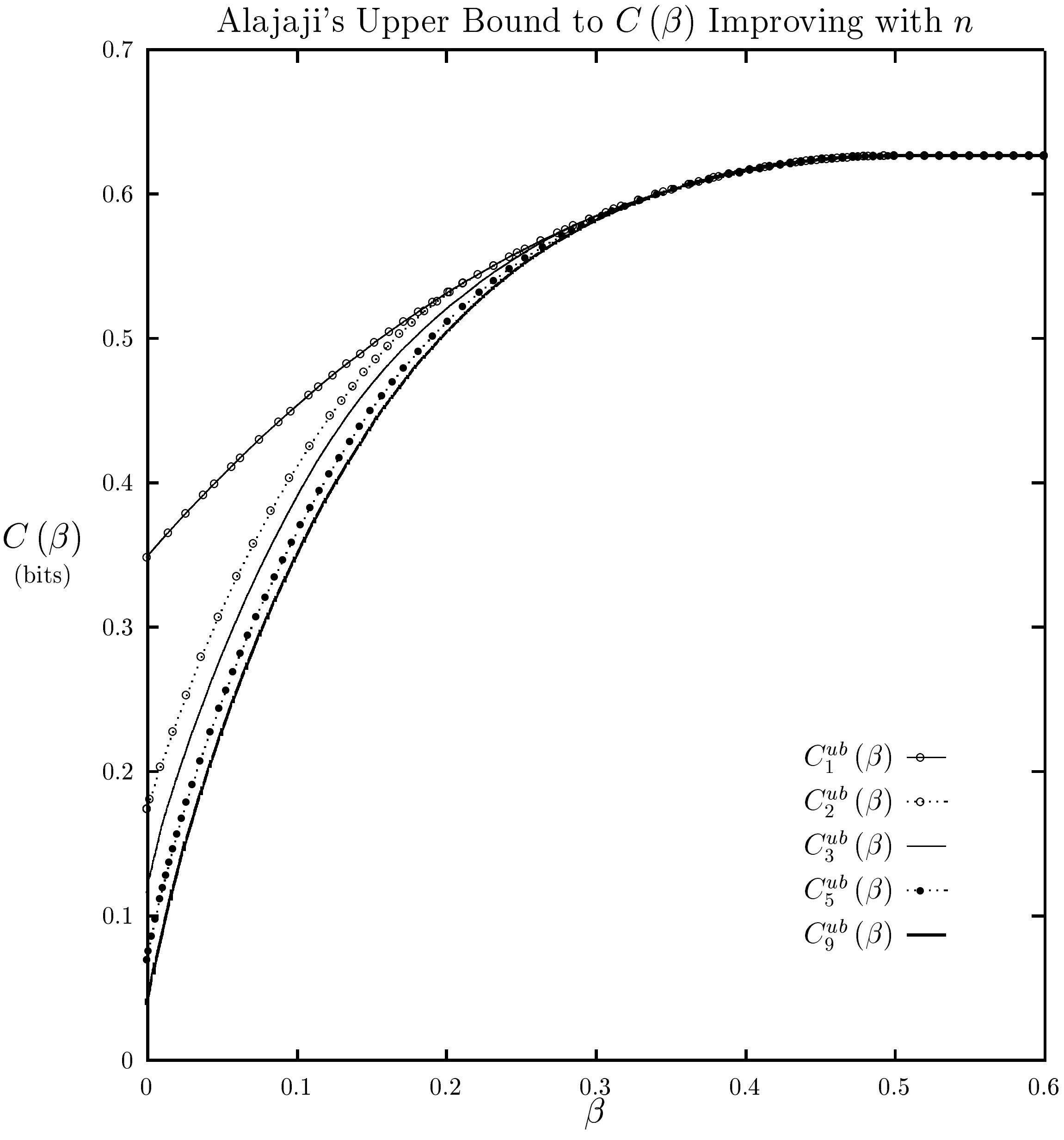


Figure 3.3: $C_n^{ub}(\beta)$ for $n = 1, 2, 3, 5$ and 9 using 1st order binary Markov noise with $\alpha = .05$ and $\varrho = 0.2$, $\beta_{max} = 1/2$. Cost function: $b(0) = 0, b(1) = 1$.

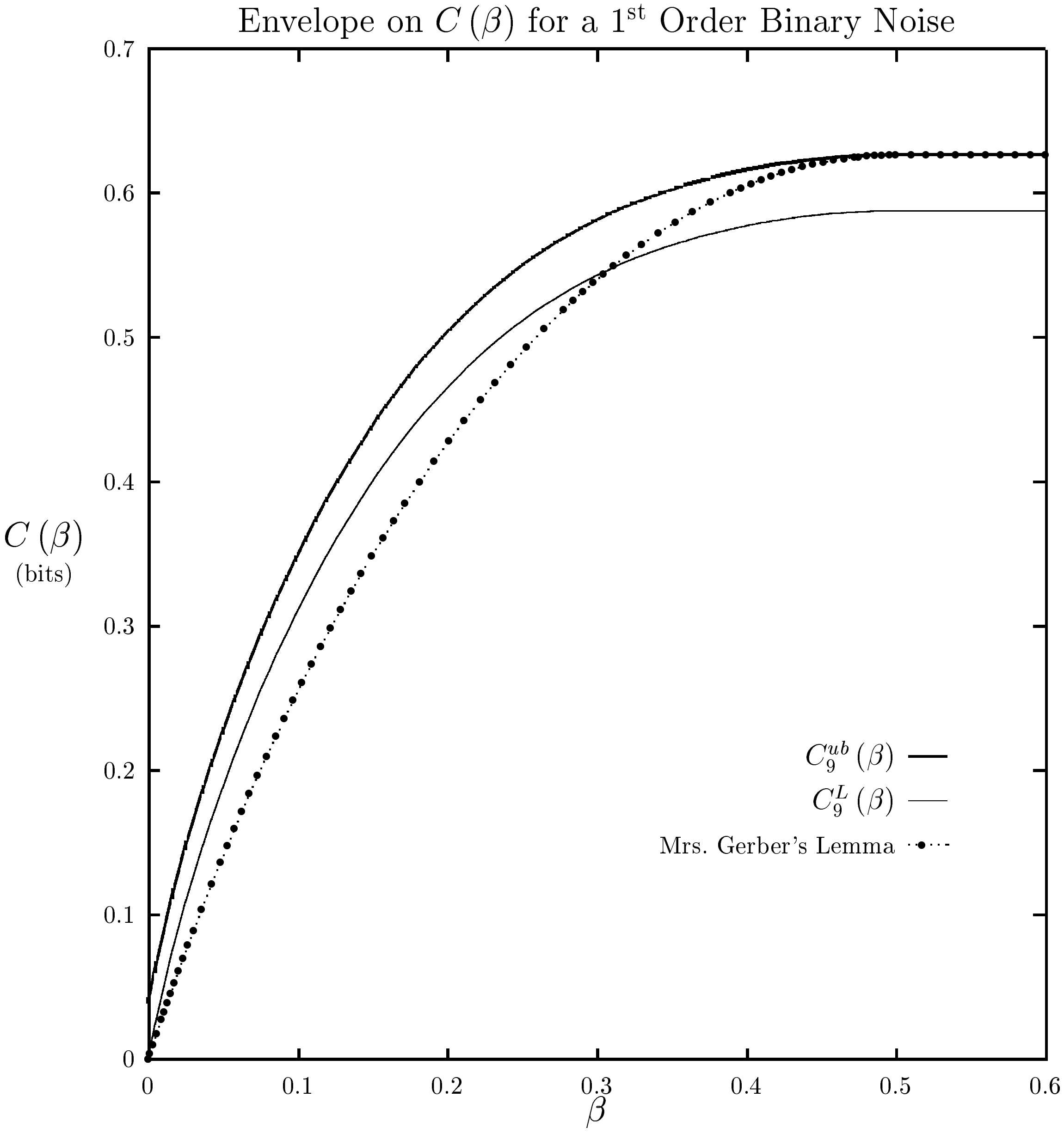


Figure 3.4: Comparison of $C_9^{ub}(\beta)$ with $C_9^L(\beta)$ and Mrs. Gerber's lower bound for a first order binary Markov noise with $\alpha = .05$ and $\varrho = .2$. $\beta_{max} = 1/2$. Cost function: $b(0) = 0, b(1) = 1$.

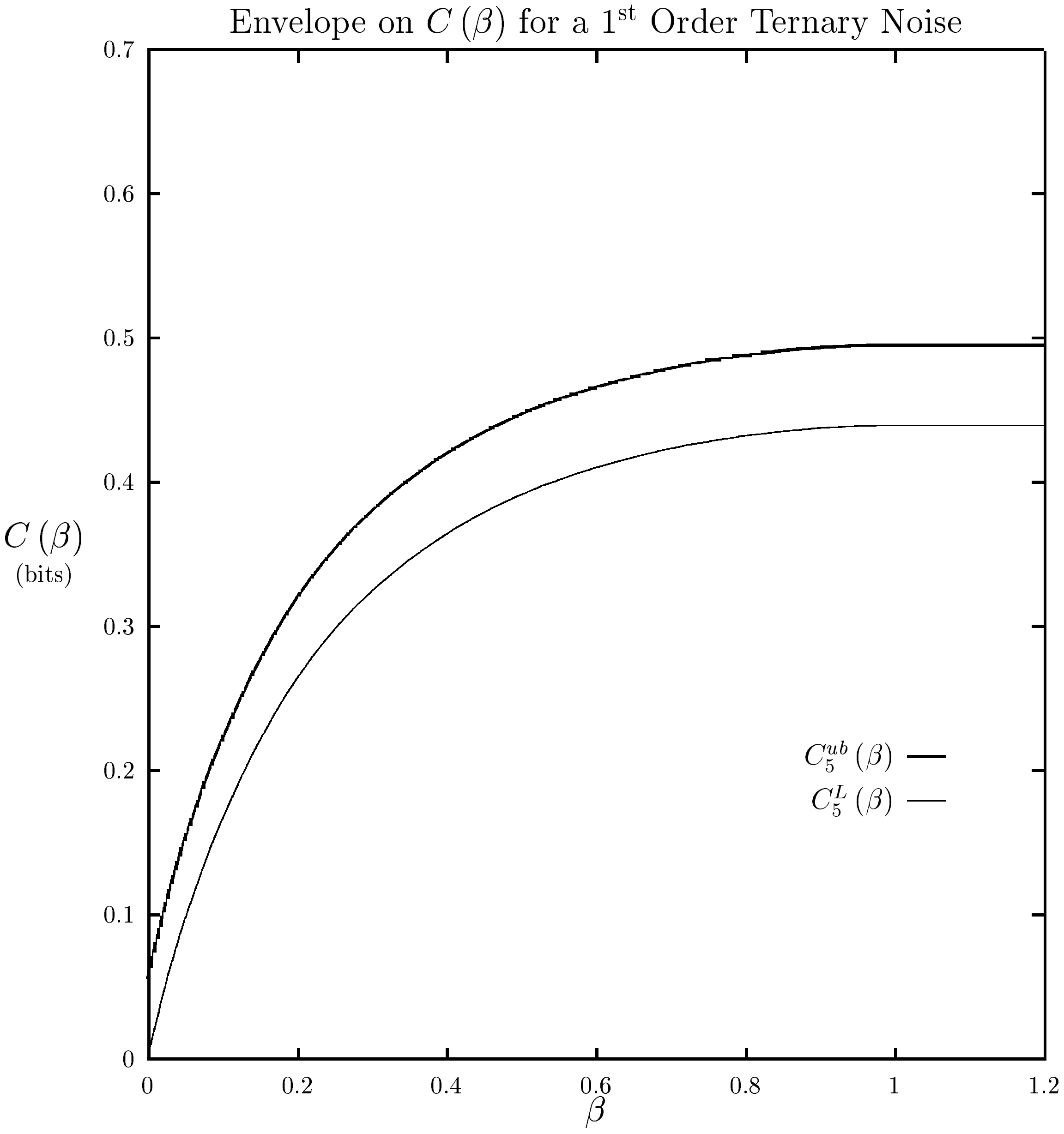


Figure 3.5: Convergence of $C_n^L(\beta)$ to $C_n^{ub}(\beta)$ for a 3-ary channel with 1st order Markov noise defined in Example 3.2. $\beta_{max} = 1$. Cost function: $b(i) = i,\ i \in \{0, 1, 2\}$.

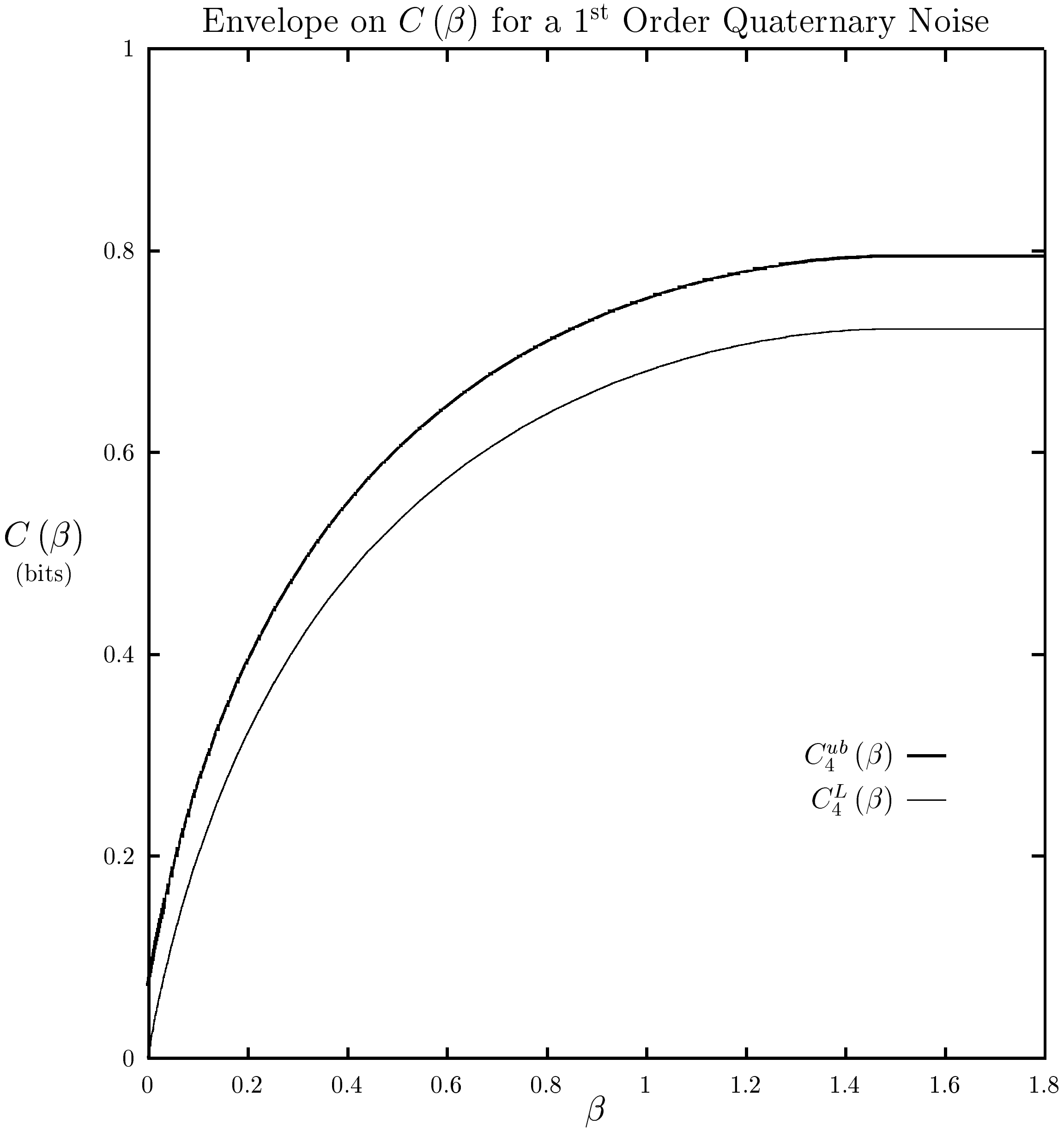


Figure 3.6: Convergence of $C_n^L(\beta)$ to $C_n^{ub}(\beta)$ for a 4-ary channel with 1st order Markov noise defined in Example 3.3. $\beta_{max} = 1.5$. Cost function: $b(i) = i, \ i \in \{0, 1, 2, 3\}$.

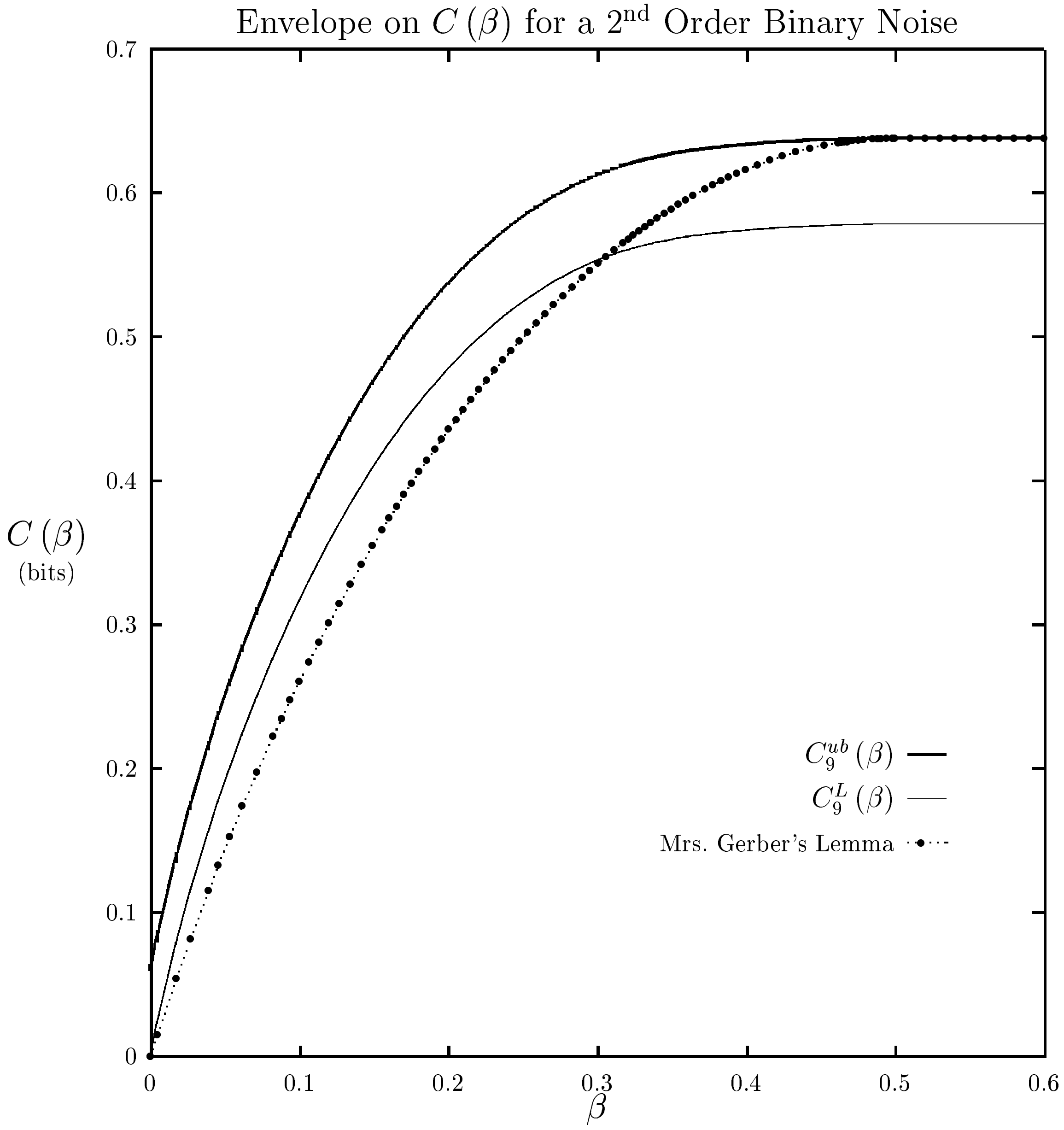


Figure 3.7: Comparison of $C_9^{ub}(\beta)$ with $C_9^L(\beta)$ and Mrs. Gerber's Lemma for a 2^{nd} order binary Markov noise whose parameters are given in Example 3.4, $\beta_{max} = 1/2$. Cost function: $b\left(0\right) = 0, b\left(1\right) = 1$.

Chapter 4

Additive Noise Channels with Feedback

Consider a q-ary alphabet channel with a large capacity, noiseless, delayless feedback loop. The additive noise process is assumed to be a stationary mixing (hence ergodic) Markov process of order k. The encoder is also assumed to have both sufficient storage power for the received and transmitted symbols and sufficient computation power for the analysis of the data. The question then becomes: does this additional information help to increase the amount of information that can be transmitted subject to an input cost constraint? We have already analyzed the computation of the capacity-cost function for channels without feedback. We herein investigate the effect of feedback on the capacity-cost function.

The first section of this chapter states some preliminary results to point out the direction of our search for capacity-cost function increasing feedback schemes. We establish a lower bound ($C^{lb}(\beta)$) to the capacity-cost function with feedback ($C_{FB}(\beta)$) which can be computed via Blahut's algorithm. We next propose a non-linear feedback scheme and characterize a set of noise processes for which $C_{FB}(\beta) > C(\beta)$. This is demonstrated analytically and via numerical bounds.

4.1 Existing Feedback Results

A number of results concerning the usefulness or uselessness of feedback already instruct us in our search. Shannon showed in [22] that side information (feedback being a special case) does not increase the capacity of discrete memoryless channels. For continuous channels, Cover and Pombra [10] showed that linear feedback over non-white additive Gaussian noise channels allows for an increase of at most $\frac{1}{2}$ bits per transmission over the non-feedback capacity, $C \leq C_{FB} \leq C + \frac{1}{2}$. No increase is possible for white Gaussian noise channels. Ihara also provides general conditions on the noise and average power of non-white Gaussian channels under which the capacity is increased by feedback ([15, 16]). In [1], Alajaji extends Shannon's result to discrete modulo addition channels with arbitrary noise (not necessarily stationary ergodic). He also conjectures that while an increase in *capacity* is impossible with feedback, an increase in the *capacity-cost function* may be evident.

This thesis is motivated by Alajaji's hypothesis, and the search initially involved an attempt to extend Cover's linear feedback strategy to our discrete channels. Unfortunately, the complexity of high order Markov systems makes them difficult to study, and linear feedback strategies mod q are too few in number for low order systems to approximate the methods of Cover and Pombra. Nonlinear feedback techniques are herein employed to show an increase in the constrained capacity.

4.2 A Lower Bound to the Feedback Capacity-Cost Function

We now expand our description of a discrete channel from Chapter 2 to include feedback. This description matches very closely with the notation in [2, 11].

Definition 4.1 A feedback channel block code with block length n and rate R consists of the following

- An index set $\Psi = \{1, 2, \ldots, 2^{nR}\}$ on the messages $\{W\}$.
- A sequence of encoding functions
$$f_i : \Psi \times \mathcal{Y}^{i-1} \rightarrow \mathcal{X} \tag{4.1}$$
for $i = 1, 2, \ldots, n$.
- A decoding function,
$$g : \mathcal{Y}^n \rightarrow \Psi, \tag{4.2}$$
which is a deterministic rule assigning an estimate, $\hat{W}$, to each output vector.

Transmitting message $W \in \{1, 2, \ldots, 2^{nR}\}$, implies sending $X^n = (X_1, X_2, \ldots, X_n)$ symbol by symbol, where $X_i = f_i(W, Y_1, Y_2, \ldots, Y_{i-1})$ for $i = 1, 2, \ldots, n$. The decoder receives $Y^n = (Y_1, Y_2, \ldots, Y_n)$ and guesses the original message to be $g(Y^n)$. We assume the messages to be uniformly distributed over the indexing set.

The average probability of decoding error is then

$$\begin{aligned} P_e^{(n)} &= \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} \Pr\{g(Y^n) \neq k | W = k\} \qquad (4.3) \\ &= \Pr\{g(Y^n) \neq k | W = k\}. \qquad (4.4) \end{aligned}$$

In Section 2.2 we associated a cost with each channel input symbol, and required that the average cost be less than the cost constraint β.

Definition 4.2 A code rate R is said to be achievable at cost β if there exists a sequence of β-admissible (cf. Definition 2.11) feedback codes of block length n and rate R such that

$$\lim_{n \rightarrow \infty} P_e^{(n)} = 0. \tag{4.5}$$

The capacity-cost function with feedback $C_{FB}(\beta)$ is defined to be the supremum of all such rates R.

A β-admissible feedback random vector has expected per letter costs given by

$$\frac{1}{n}E\left[b\left(X^n\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{w}\sum_{y^{i-1}} P_{W,Y^{i-1}}(w, y_1, \ldots, y_{i-1})\ b\left(f_i(w, y_1, \ldots, y_{i-1})\right) \tag{4.6}$$

As our goal is the identification of feedback strategies capable of increasing the constrained capacity, we should characterize the various possibilities for the encoding sequence $\{f_i\}$, $i = 1, 2, \ldots, n$.

We begin by representing the random input message W by a random n-tuple $V^n = (V_1, V_2, \ldots, V_n)$. We assume that W and $\{Z_i\}$ are independent from each other. At time i we transmit symbol

$$X_i = f_i(V_i, Y_1, Y_2, \ldots, Y_{i-1}), \qquad X_i \in \mathcal{A}_q,\ i = 1, \ldots, n. \tag{4.7}$$

For our additive noise channels, the Y_i are given by

$$Y_i = X_i \oplus Z_i, \tag{4.8}$$

and where $Z_i \in \mathcal{A}_q$ is drawn from a stationary ergodic Markov process $\{Z_i\}_{i=1}^{\infty}$. Since $\oplus$ is an invertible operation, and since feedback provides the encoder with knowledge of $Y_1, \ldots, Y_{i-1}$ at time i, it can deduce $Z_1, Z_2, \ldots, Z_{i-1}$ from the equation

$$Z_j = Y_j \ominus X_j, \qquad j = 1, \ldots, i-1.$$

Note as well that for a finite memory system of order k, the feedback of terms more than k time steps old provides no new information. Therefore, we can express the feedback function in terms of the input components and noise state as

$$X_i = f_i(V_i, Z_{i-k}, \ldots, Z_{i-1}). \tag{4.9}$$

In general, the feedback rule $f_i(V_i, Z_{i-k}, \ldots, Z_{i-1})$ is time varying and dependent on both the input and the noise. In this thesis we restrict our study to time invariant feedback rules.

Observe that the cost constraint is imposed on the input vector X^n; i.e., it is required that

$$\frac{1}{n}E[b(X^n)] \leq \beta. \tag{4.10}$$

For this mod q channel with feedback, we define $C^{lb}(\beta)$ using a *fixed* encoding rule f^* as

$$C^{lb}(\beta) = \sup_n C_n^{lb}(\beta) = \lim_{n\to\infty} C_n^{lb}(\beta), \tag{4.11}$$

where

$$C_n^{lb}(\beta) = \max_{P_{V^n}(v^n)\in\tilde{\tau}_n(\beta)} \frac{1}{n} I(V^n; Y^n), \tag{4.12}$$

where

$$\tilde{\tau}_n(\beta) = \left\{ P_{V^n}(v^n) : \frac{1}{n}\sum_{i=1}^{n} E\left[b\left(X_i\right)\right] \leq \beta \right\}, \tag{4.13}$$

and where $X_i = f^*(V_i, Z_{i-k}, \ldots, Z_{i-1})$.

Let us now demonstrate the achievability of $C^{lb}(\beta)$, thereby justifying its use as a lower bound on $C_{FB}(\beta)$. The proof involves a use of the asymptotic equipartition property, and requires a definition of *jointly ϵ-typical* sequences [11].

Definition 4.3 ([11]) Given jointly distributed random vectors (V^n, Y^n) with distribution $P_{V^n,Y^n}(v^n, y^n)$, and $\epsilon > 0$, the set $A_\epsilon^{(n)}$ of *jointly ϵ-typical* sequences (V^n, Y^n) is defined by

$$A_\epsilon^{(n)} \triangleq \left\{ (v^n, y^n) \in \mathcal{V}^n \times \mathcal{Y}^n : \left| -\frac{1}{n}\log P_{V^n}(v^n) - H(V_\infty) \right| \leq \epsilon \right. \tag{4.14}$$

$$\left| -\frac{1}{n}\log P_{Y^n}(y^n) - H(Y_\infty) \right| \leq \epsilon \tag{4.15}$$

$$\left. \left| -\frac{1}{n}\log P_{V^n,Y^n}(v^n, y^n) - H(V_\infty, Y_\infty) \right| \leq \epsilon \right\}, \tag{4.16}$$

where $P_{V^n}(v^n)$ and $P_{Y^n}(y^n)$ are the n-fold marginal distributions corresponding to the joint p.m.f. $P_{V^n,Y^n}(v^n, y^n)$. The set $A_\epsilon^{(n)}$ is then the set of n-sequences whose empirical entropies are within ϵ of the true entropies.

Theorem 4.1 (Achievability of $C^{lb}(\beta)$: $C_{FB}(\beta) \geq C^{lb}(\beta)$) Consider a q-ary k^{th} order additive Markov noise channel defined above with a fixed time invariant feedback function f^*. If $C_n^{lb}(\beta)$ is as defined in Equations (4.11) to (4.13), then there exists a sequence of β-admissible feedback codes of block length n and rate R such that $P_e^{(n)} \to 0$ as $n \to \infty$ for all rates $R < C^{lb}(\beta)$.

Proof of Theorem 4.1 We use random coding by letting the indexed q-ary representation of W, given by $V^n(1), V^n(2), \ldots, V^n(2^{nR})$, be independent identically distributed n-vectors drawn from $P_{V^n}(v^n)$, where $P_{V^n}(v^n)$ achieves $C^{lb}(\beta)$ and is an n-fold distribution of a stationary ergodic process.

Encoding: To send message W, where W is uniform over $\{1, 2, \ldots, 2^{nR}\}$, the transmitter sends $X^n(W, Z^{n-1}) = (X_1, X_2, \ldots, X_n)$, where $X_i = f^*(V_i, Z_{i-k}, \ldots, Z_{i-1})$. Since $f^*(\cdot)$ is assumed to be a time invariant function of stationary ergodic processes $\{V_i\}$ and $\{Z_i\}$, then $\{X_i\}$ is also stationary ergodic.

Decoding: The receiver must process $Y^n = (Y_1, Y_2, \ldots, Y_n)$, where $Y_i = X_i \oplus Z_i$ for $i = 1, 2, \ldots, n$. It decides that $\hat{W} \in \{1, 2, \ldots, 2^{nR}\}$ must have been sent if $\left(V^n(\hat{W}), Y^n\right)$ is the only jointly ϵ-typical pair. Due to the random coding, we assume without loss of generality that $W = 1$ was sent. Let

$$E_i = \left\{(V^n(i), Y^n) \in A_\epsilon^{(n)}\right\}, \qquad i = 1, 2, \ldots, 2^{nR}; \tag{4.17}$$

and let E_i^c denote the complement of E_i.

The receiver declares an error if any of the four following events occur:

- there is no jointly ϵ-typical pair $\left(V^n(\hat{W}), Y^n\right)$,

- there is more than one such pair,
- $\hat{W} \neq W = 1$,
- or the β-admissibility constraint is violated. We define this possibility by the event

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} b\left(X_i(1)\right) > \beta \right\}. \tag{4.18}$$

The probability of the first three events can be simplified into

$$\Pr(\hat{W} \neq 1 | W = 1) = \Pr(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}} | W = 1). \tag{4.19}$$

Cover ([11]) explains how this allows us to write the overall probability of error as

$$\begin{aligned} P_e^{(n)} &= \Pr(E_0) + \Pr(\hat{W} \neq 1 | W = 1) & (4.20) \\ &= \Pr(E_0) + \Pr(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}} | W = 1) & (4.21) \\ &\leq \Pr(E_0) + \Pr(E_1^c | W = 1) + \sum_{i=2}^{2^{nR}} \Pr(E_i | W = 1). & (4.22) \end{aligned}$$

We intend to show that each of the three terms in Equation (4.22) becomes negligibly small as $n \to \infty$. Since the codewords are drawn according to a stationary ergodic distribution that achieves $C^{lb}(\beta)$, $\Pr(E_0) < \epsilon$ for n sufficiently large by the law of large numbers. We now prove as a lemma, a variation of Theorem 8.6.1 from [11].

Lemma 4.1 (Joint AEP) Given the joint sequence (V^n, Y^n) of length n defined above, the following results hold.

a)

$$|A_\epsilon^{(n)}| \leq 2^{n[H(V_\infty, Y_\infty) + \epsilon]}. \tag{4.23}$$

b) For n sufficiently large,

$$\Pr\left((V^n, Y^n) \in A_\epsilon^{(n)}\right) \geq 1 - \epsilon. \tag{4.24}$$

c) If $\tilde{V}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $P_{V^n,Y^n}(v^n, y^n)$, then for n sufficiently large

$$(1-\epsilon)2^{-n[\frac{1}{n}I(V_\infty;Y_\infty)+3\epsilon]} \leq \Pr\left((\tilde{V}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq 2^{-n[\frac{1}{n}I(V_\infty;Y_\infty)-3\epsilon]} \tag{4.25}$$

Proof of Lemma 4.1 We refer to [11] for the proofs of parts a) and c) above, and prove the result in b) as follows. Since

$$Y_i = f^*(V_i, Z_{i-k}, \ldots, Z_{i-1}) \oplus Z_i, \tag{4.26}$$

$\{Y_i\}$ is a time-invariant function of the stationary ergodic processes $\{V_i\}$ and $\{Z_i\}$. Hence $\{Y_i\}$ is stationary ergodic.

$$\Pr\left((V^n, Y^n) \notin A_\epsilon^{(n)}\right) \tag{4.27}$$

$$\begin{aligned} &\leq \Pr\left(\left|-\frac{1}{n}\log P_{V^n}(V^n) - H(V_\infty)\right| > \epsilon\right) \\ &\quad + \Pr\left(\left|-\frac{1}{n}\log P_{Y^n}(Y^n) - H(Y_\infty)\right| > \epsilon\right) \\ &\quad + \Pr\left(\left|-\frac{1}{n}\log P_{V^n,Y^n}(V^n, Y^n) - H(V_\infty, Y_\infty)\right| > \epsilon\right). \end{aligned} \tag{4.28}$$

The first term on the right-hand side of (4.28) is $\leq \epsilon/3$ for n sufficiently large since $\{V_i\}$ is stationary ergodic. Similarly, the second term is $\leq \epsilon/3$ for n sufficiently large since $\{Y_i\}$ is stationary ergodic. Let us now examine the third term. First note that

$$P_{V^n,Y^n}(v^n, y^n) \tag{4.29}$$

$$= P_{Y^n|V^n}(y^n|v^n)P_{V^n}(v^n) \tag{4.30}$$

$$= P_{V^n}(v^n)\prod_{i=1}^{n} P_{Y_i|Y^{i-1},V^n}(y_i|y^{i-1}, v^n) \tag{4.31}$$

$$
\begin{aligned}
&= P_{V^n}(v^n) \prod_{i=1}^{n} P_{Z_i|Z_{i-k},\ldots,Z_{i-1}}(z_i|z_{i-k},\ldots,z_{i-1}) && (4.32)\\
&= P_{V^n}(v^n) P_{Z^n}(z^n), && (4.33)
\end{aligned}
$$

where $z^n = (z_1, \ldots, z_n)$ and $z_i = y_i \ominus f^*(v_i, y^{i-1})$, $i = 1, \ldots, n$. Therefore,

$$-\frac{1}{n}\log P_{V^n,Y^n}(v^n, y^n) = -\frac{1}{n}\log P_{V^n}(v^n) - \frac{1}{n}\log P_{Z^n}(z^n). \tag{4.34}$$

By the same rationale,

$$
\begin{aligned}
H(V_\infty, Y_\infty) &= H(V_\infty) + H(Y_\infty|V_\infty) && (4.35)\\
&= H(V_\infty) + H(Z_\infty). && (4.36)
\end{aligned}
$$

However,

$$-\frac{1}{n}\log P_{V^n}(V^n) \overset{n\to\infty}{\longrightarrow} H(V_\infty) \tag{4.37}$$

in probability, and

$$-\frac{1}{n}\log P_{Z^n}(Z^n) \overset{n\to\infty}{\longrightarrow} H(Z_\infty) \tag{4.38}$$

in probability, which implies that

$$-\frac{1}{n}\log P_{V^n}(V^n) - \frac{1}{n}\log P_{Z^n}(Z^n) \overset{n\to\infty}{\longrightarrow} H(V_\infty) + H(Z_\infty). \tag{4.39}$$

Therefore,

$$\Pr\left(\left|-\frac{1}{n}\log P_{V^n,Y^n}(V^n, Y^n) - H(V_\infty, Y_\infty)\right| > \epsilon\right) < \frac{\epsilon}{3} \tag{4.40}$$

for n sufficiently large. Substituting back into Equation (4.28) yields

$$\Pr\left((V^n, Y^n) \notin A_\epsilon^{(n)}\right) < \epsilon \tag{4.41}$$

for n sufficiently large; this completes the proof of b). □

Now, by Part b) of Lemma 4.1

$$\Pr(E_1^c) = \Pr\left((V^n, Y^n) \notin A_\epsilon^{(n)}\right) < \epsilon, \tag{4.42}$$

for n sufficiently large.

Part c) of Lemma 4.1 refers to the probability that an independent distribution will also generate pairs in the jointly ϵ-typical set. Therefore, for $i \neq 1$

$$\begin{aligned} \Pr(E_i|W=1) &= \Pr\left((\tilde{V}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) && (4.43) \\ &\leq 2^{-n[\frac{1}{n}I(V_\infty;Y_\infty)-3\epsilon]} && (4.44) \\ &= 2^{-n[C^{lb}(\beta)+3\epsilon]}, && (4.45) \end{aligned}$$

since $P_{V^n}(v^n)$ is drawn from the stationary ergodic distribution that achieves $C^{lb}(\beta)$.

Therefore, Equation (4.22) reduces to

$$P_e^{(n)} \leq 2\epsilon + (2^{nR}-2)2^{-n[C^{lb}(\beta)+3\epsilon]} \leq 3\epsilon \qquad (4.46)$$

if n is sufficiently large and $R < C^{lb}(\beta)$. Therefore, $C^{lb}(\beta)$ is achievable and constitutes a lower bound to $C_{FB}(\beta)$. □

4.3 Nonlinear Feedback where $C_{FB}(\beta) > C(\beta)$

In this section we develop a nonlinear feedback scheme and a set of noise processes for which a uniform input distribution results in a uniform output distribution, and feedback capacity is achieved at an expected cost $\tilde{\beta}^{lb} < \beta_{max}$. This implies that feedback increases $C(\beta)$ for such channels.

We begin by describing the particular feedback encoding function f^* that we intend to use. We then show that for a particular type of Markov noise, the feedback channel transition probabilities are equal to the non-feedback channel transition probabilities. This allows us to verify that

$$C^{lb}(\beta) = C_{FB}(\beta) = \log q - H(Z_\infty) \qquad (4.47)$$

for $\beta \geq \beta_{max}$.

For a channel with q-ary k^{th} order additive Markov noise, consider the following time invariant feedback encoding function $f^*(\cdot)$.

$$\begin{aligned} X_i &= V_i \quad \text{if } i \leq k; \qquad (4.48) \\ X_i &= f^*(V_i, S_i) \triangleq \begin{cases} V_i, & S_i \neq \tilde{s} \\ 0, & S_i = \tilde{s}, \end{cases} \quad \text{if } i > k, \qquad (4.49) \end{aligned}$$

where $S_i \triangleq (Z_{i-k}, Z_{i-k+1}, \ldots, Z_{i-1})$ is the random vector describing the state of the noise process at time i, $\tilde{s}$ is some preselected state.

Under linear or power cost constraints this feedback strategy asks the transmitter to monitor the noise state, S. If the encoder detects a particular *bad* state $\tilde{s}$ (i.e., one whose transition probabilities are nearly uniform) at step i then the transmitter is instructed to send the least expensive word irrespective of the current message symbol V_i. In our examples the least expensive letter has $b(0) = 0$.

Let us now apply this feedback strategy to a q-ary channel with a particular additive Markov noise of order k.

Lemma 4.2 Consider a q-ary channel with stationary ergodic additive Markov noise of order k with and without the feedback rule given in Equation (4.48) above. If for a particular noise state $\tilde{s}$ the conditional probabilities of the current noise sample are uniformly distributed, i.e.,

$$P_{Z_i|S_i}(z_i|\tilde{s}) = \frac{1}{q}, \qquad \forall\ z_i\ \in \mathcal{A}_q, \qquad (4.50)$$

then the conditional probabilities of y^n given v^n are equal for both the feedback and non-feedback channels, i.e.,

$$P_{Y^n|V^n}(y^n|v^n) = P^{FB}_{Y^n|V^n}(y^n|v^n), \qquad (4.51)$$

for all $y^n, v^n \in \mathcal{A}_q^k$.

Proof of Lemma 4.2 The transition probabilities for the non-feedback channel are given by

$$\begin{aligned} P_{Y^n|V^n}(y^n|v^n) &= P_{Z^n}(z^n = y^n \ominus v^n) && (4.52) \\ &= P_{Z^k}(y_1 \ominus v_1, y_2 \ominus v_2, \ldots, y_k \ominus v_k) \\ &\quad \cdot \prod_{i=k+1}^{n} P_{Z_i|S_i}(y_i \ominus v_i|s_i), && (4.53) \end{aligned}$$

where $s_i = (z_{i-k}, z_{i-k+1}, \ldots, z_{i-1})$ is the state of the Markov chain at step i for a given input-output pair (v^n, y^n). Using the same notation but with a superscript to denote the feedback channel, the transition probabilities are given by

$$\begin{aligned} P^{FB}_{Y^n|V^n}(y^n|v^n) &= P_{Z^k}(y_1 \ominus v_1, y_2 \ominus v_2, \ldots, y_k \ominus v_k) \\ &\quad \cdot \prod_{i=k+1}^{n} P^*_{Z_i|S_i}(y_i \ominus f^*(v_i, s_i)|s_i), && (4.54) \end{aligned}$$

where

$$P^*_{Z_i|S_i}(y_i \ominus f^*(v_i, s_i)|s_i) = \begin{cases} P_{Z_i|S_i}(y_i \ominus v_i|s_i), & \text{if } s_i \neq \tilde{s} \\ P_{Z_i|S_i}(y_i|s_i), & \text{if } s_i = \tilde{s}. \end{cases} \tag{4.55}$$

Notice that Equations (4.53) and (4.54) are identical except possibly when noise state $\tilde{s}$ occurs. But $P_{Z|S}(z|\tilde{s}) = \frac{1}{q}$ for all $z \in \{0, 1, \ldots, q-1\}$, which implies that

$$P_{Z_i|S_i}(y_i|\tilde{s}) = P_{Z_i|S_i}(y_i \ominus v_i|\tilde{s}) = \frac{1}{q}. \tag{4.56}$$

Therefore, $P_{Y^n|V^n}(y^n|v^n) = P^{FB}_{Y^n|V^n}(y^n|v^n)$ for the feedback encoding scheme in (4.48) if the conditional probabilities of Z_i given $S_i = \tilde{s}$ are uniform. □

Lemma 4.2 implies that since the non-feedback channel is symmetric, then so is the feedback channel. From Theorem 2.6[1] we can infer that a uniform distribution on the input blocks V^n induces a uniform distribution on the output blocks Y^n.

[1] Actually, Theorem 2.6 only requires a weakly symmetric channel.

We have now shown that for a particular type of Markov noise, our feedback rule has no effect on the conditional distribution. It does, however, affect the cost of individual input blocks. The following lemma compares the expected cost of non-feedback distributions with feedback distributions using our strategy.

Lemma 4.3 Consider the non-feedback and feedback channels described above, with feedback strategy given in Equation (4.48), and $P_{Z|S}(z|\tilde{s}) = \frac{1}{q}$ for all z. Let $P^*_{V^n}(v^n)$ be a stationary input distribution that achieves $C_n(\beta)$ for $\beta > \beta_{min}{}^2$. Then

$$C_n^{lb}\left(\beta_n^{lb}\right) \geq C_n(\beta) \tag{4.57}$$

where β_n^{lb} is the expected per letter cost of $P^*_{V^n}(v^n)$ under the feedback encoding strategy given by

$$\beta_n^{lb} = \left[1 - \frac{n-k}{n} P_S(\tilde{s})\right] \beta \tag{4.58}$$

Proof of Lemma 4.3 By Corollary B.3, for the non-feedback channel

$$\begin{aligned} \beta &= \frac{1}{n} \sum_{v^n} P^*_{V^n}(v^n) b(v^n) && (4.59) \\ &= \sum_{v} P_V(v) b(v) && (4.60) \end{aligned}$$

since $P^*_{V^n}(v^n)$ is a stationary input distribution that achieves the non-feedback capacity-cost function. For the feedback channel we charge costs to the channel input letters after applying the feedback rule f^*. Thus

$$\begin{aligned} \beta_n^{lb} &= \frac{1}{n} E\left[b(X^n)\right] = \frac{1}{n} \sum_{i=1}^{n} E\left[b(X_i)\right] && (4.61) \\ &= \frac{1}{n} \sum_{i=1}^{k} E\left[b(X_i)\right] + \frac{1}{n} \sum_{i=k+1}^{n} E\left[b(X_i)\right] && (4.62) \end{aligned}$$

2The existence of a stationary input distribution that achieves the capacity-cost is shown in [9].

$$= \frac{1}{n}\sum_{i=1}^{k}\sum_{v} P_V(v_i) b\left(X_i\right) + \frac{1}{n}\sum_{i=k+1}^{n} E\left[b\left(f^*(v_i, s_i)\right)\right] \tag{4.63}$$

$$= \frac{k}{n}\beta + \frac{n-k}{n}\sum_{v}\sum_{s} P_S(s)P_V(v)b\left(f^*(v,s)\right) \tag{4.64}$$

$$= \frac{k}{n}\beta + \frac{n-k}{n}\left[\sum_{v} P_S(\tilde{s})P_V(v)b\left(0\right)\right.$$

$$\left. + \sum_{s\neq\tilde{s}}\sum_{v} P_S(s)P_V(v)b\left(v\right)\right] \tag{4.65}$$

$$= \frac{k}{n}\beta + \frac{n-k}{n}\sum_{s\neq\tilde{s}} P_S(s)\beta \tag{4.66}$$

$$= \left[1 - \frac{n-k}{n}P_S(\tilde{s})\right]\beta, \tag{4.67}$$

where $s_i = (z_{i-k}, \ldots, z_{i-1})$. Note that, since we are dealing with stationary mixing noise processes, $P_S(\tilde{s}) > 0$ and thus $\beta_n^{lb} < \beta$. Now, by Lemma 4.2 above, the channel transition probabilities are identical for the feedback and non-feedback channel. Using $P^*_{V^n}(v^n)$ as a particular input distribution,

$$C_n^{lb}\left(\beta_n^{lb}\right) \geq \frac{1}{n}\sum_{v^n,y^n} P^*_{V^n}(v^n)P^{FB}_{Y^n|V^n}(y^n|v^n)\log\frac{P^{FB}_{Y^n|V^n}(y^n|v^n)}{\sum_{v^n} P^*_{V^n}(v^n)P^{FB}_{Y^n|V^n}(y^n|v^n)} \tag{4.68}$$

$$= \frac{1}{n}\sum_{v^n,y^n} P^*_{V^n}(v^n)P_{Y^n|V^n}(y^n|v^n)\log\frac{P_{Y^n|V^n}(y^n|v^n)}{\sum_{v^n} P^*_{V^n}(v^n)P_{Y^n|V^n}(y^n|v^n)} \tag{4.69}$$

$$= C_n\left(\beta\right). \tag{4.70}$$

□

Let us define $\tilde{\beta}_n^{lb}$ such that

$$C_n^{lb}\left(\tilde{\beta}_n^{lb}\right) = C_n\left(\beta_{max}\right) = \log q - \frac{1}{n}H(Z^n). \tag{4.71}$$

Alajaji has shown in [1] that feedback cannot increase channel capacity. Since we already know that a uniform input distribution achieves capacity for the non-feedback channel, Lemmas 4.2 and 4.3 imply that a uniform input distribution also achieves capacity for the feedback channel. We can now find β_{max} for the non-feedback channel, and use the above lemma to determine $\tilde{\beta}^{lb}$. Using the linear cost function $b(i) = i$, we already know from Section 2.3 that for the non-feedback channel:

$$\beta_{max} = \beta_{max}^{(n)} = \frac{q-1}{2}, \qquad \forall\ n. \tag{4.72}$$

Therefore, for the channel with feedback,

$$\begin{aligned}
\tilde{\beta}^{lb} &\triangleq \lim_{n\to\infty} \tilde{\beta}_n^{lb} && (4.73)\\
&= \lim_{n\to\infty} \left\{ \frac{k}{n}\beta_{max} + \frac{n-k}{n}[1 - P_S(\tilde{s})]\beta_{max} \right\} && (4.74)\\
&= \lim_{n\to\infty} \left\{ \frac{k}{n} + (1 - P_S(\tilde{s}))\frac{n-k}{n} \right\} \frac{q-1}{2} && (4.75)\\
&= (1 - P_S(\tilde{s}))\frac{q-1}{2}. && (4.76)
\end{aligned}$$

Since the Markov noise is irreducible and aperiodic, $P_S(\tilde{s}) > 0$ which implies that $\tilde{\beta}^{lb} < \beta_{max}$. We observe here, as in Lemma 4.3 that our feedback channel attains capacity at a β strictly less than that for the non-feedback channel. We summarize these observations in the following theorem.

Theorem 4.2 Consider the q-ary non-feedback and feedback channels with stationary mixing additive Markov noise and feedback rule described above. Let $P_{Z_i|S_i}(z_i|\tilde{s}_i) = \frac{1}{q}$ for all $z_i \in \mathcal{A}_q$. Then for $0 < \beta < \beta_{max}$,

$$C_{FB}(\beta) > C(\beta). \tag{4.77}$$

Proof of Theorem 4.2 This theorem follows easily from Lemmas 4.2 and 4.3. Recall that $C_n(\beta)$ is strictly increasing on $\beta < \beta_{max}$ (c.f. Corollary 2.1). For input distribution $P_{V^n}^*(v^n)$ that achieves $C_n(\beta)$ for the non-feedback channel,

$$\beta_n^{lb} = \left(\frac{k}{n} + \frac{n-k}{n}[1 - P_S(\tilde{s})]\right)\beta \tag{4.78}$$

and if $P_{\tilde{s}} > 0$, then by Lemma 4.3

$$C_n(\beta) \leq C_n^{lb}\left(\beta_n^{lb}\right). \tag{4.79}$$

Therefore, taking the limit as $n \to \infty$ on (4.79), and using the fact that the limit of a concave function is concave and thus continuous, yields

$$C^{lb}\left(\beta^{lb}\right) \geq C(\beta), \tag{4.80}$$

where $\beta^{lb} = \lim_{n\to\infty} \beta_n^{lb} = (1 - P_S(\tilde{s}))\beta$. Since $C(\beta)$ is strictly increasing in β, we obtain

$$C^{lb}(\beta^{lb}) > C\left(\beta^{lb}\right), \tag{4.81}$$

which implies that

$$C_{FB}(\beta^{lb}) > C(\beta^{lb}) \tag{4.82}$$

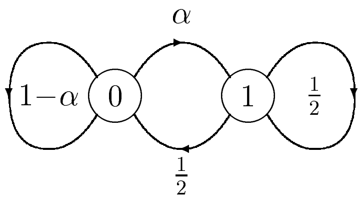
for $0 < \beta^{lb} < \beta_{max}$. □

This analytic result proves that, for some types of Markov noise, the capacity-cost function of mod q channels can be increased by feedback. In the next section, we apply our strategy to demonstrate the result numerically.

4.4 Numerical Examples

We continue with the four examples of the previous chapters, to demonstrate our new results. In some instances we use a channel with a uniformly poor state $\tilde{s}$, and in others we use a nearly uniformly poor state. In both instances we see an increase in the capacity-cost function with feedback for some range of the costs.

Example 4.1 *Binary Alphabet Channel with 1^{st} Order Markov Noise.* We introduce a channel with a particular type of binary Markov noise. As can be seen in Figure 4.1, the probability of witnessing $Z_i = 1$ given that $Z_{i-1} = 1$ is exactly $\frac{1}{2}$.



$$\mathbf{\Pi} = \begin{bmatrix} 1-\alpha & \alpha \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$P_z(Z) = \left[\tfrac{1}{2\alpha+1}, \; \tfrac{2\alpha}{2\alpha+1}\right]$$

Figure 4.1: Two-State Markov Chain: state 1 is uniformly bad.

As we proved in the previous section, there is nothing to be gained by transmitting V_i if the present channel state uniformly poor. Let $\tilde{s} = 1$. By the nonlinear feedback rule given in (4.48) we save a factor of $[1 - P_S(\tilde{s})]$ on the expected cost of sending message $W = (V_1, V_2, \ldots, V_n)$ as $n \to \infty$. We now implement this strategy and observe in Figure 4.2 that when $\alpha = .2$, a very noisy channel, the gains in $C_{FB}(\beta)$ over $C(\beta)$ are very apparent at block length $n = 8$. Notice that $\tilde{\beta}_n^{lb}$ for the feedback channel is precisely,

$$\tilde{\beta}_n^{lb} = \frac{(q-1)}{2}\left[\frac{1}{n} + \frac{(n-1)}{n}(1 - \pi_{\tilde{s}})\right] \tag{4.83}$$

which for this channel is

$$\begin{aligned} \tilde{\beta}_8^{lb} &= \frac{1}{2}\left[\frac{1}{8} + \frac{7}{8}(1 - \frac{2}{7})\right] && (4.84) \\ &= \frac{3}{8} = .375, && (4.85) \end{aligned}$$

a cost savings of 25 percent.

We now repeat this experiment for a binary channel with first order Markov noise defined by $\alpha = .18$ and $\varrho = .45$, using the same $\tilde{s}$ and feedback strategy. This noise process has stationary distribution equal to the one above. Figure 4.3 shows that we obtain a numerical increase in $C_8^{lb}(\beta)$ over $C_8^{ub}(\beta)$ even if the state $\tilde{s}$ is not uniformly corrupting.

Example 4.2 *Ternary Alphabet Channel with* 1^{st} *Order Markov Noise.* For the case where state $\tilde{s}$ is uniformly corrupting, we have analytically shown a strict increase in the capacity-cost function as $n \to \infty$ using our simple feedback scheme as the lower bound. We use the following Markov chain and block length $n = 5$ to create the graphs in Figure 4.4 and Figure 4.5:

$$\mathbf{\Pi} = \begin{bmatrix} .70 & .05 & .25 \\ .71 & .04 & .25 \\ .333334 & .333333 & .333333 \end{bmatrix}. \tag{4.86}$$

The state probabilities for this example are

$$\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2] = \left[0.\overline{6012}, 0.\overline{1260}, 0.\overline{27}\right], \tag{4.87}$$

where state $\tilde{s}$ is defined by the event $Z_{i-1} = 2$. The resultant noise process given that $Z_{i-1} = 2$ is not exactly uniform, but we do see a distinct increase in $C_5^{lb}(\beta)$ over $C_5^{ub}(\beta)$. This increase grows as $n \to \infty$. As well, $\tilde{\beta}_5^{lb} = 0.7818$ while $\tilde{\beta}^{lb} = 0.\overline{72}$, which would also serve to increase the gap between the bounds.

Example 4.3 *Quaternary Alphabet Channel with* 1^{st} *Order Markov Noise.* The maximum block length for which Blahut's algorithm would converge in a reasonable amount of time (2 days for 100 data points) was $n = 4$. We use the following Markov chain to create the graphs in Figure 4.6 and Figure 4.7:

$$\mathbf{\Pi} = \begin{bmatrix} .55 & .10 & .10 & .25 \\ .55 & .10 & .10 & .25 \\ .55 & .10 & .10 & .25 \\ .25 & .25 & .25 & .25 \end{bmatrix}. \tag{4.88}$$

The state probabilities for this example are

$$\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2, \pi_3] = [0.475, 0.1375, 0.1375, 0.25], \tag{4.89}$$

where state $\tilde{s}$ is defined by the event $Z_{i-1} = 3$. As $n \to \infty$ we experience additional increase in $C^{lb}(\beta)$ due to memory decrease to zero ($M_4 = .017$), and due to cost decrease from $\tilde{\beta}_4^{lb} = 1.219$ down to $\tilde{\beta}^{lb} = 1.125$.

Example 4.4 *Binary Alphabet Channel with 2^{nd} Order Markov Noise.* As with the first order mod 2 case, in the second order case we are also able to compute a high enough block length to see a substantial increase in the feedback capacity-cost function. As such, we wish to demonstrate this increase for two channels not explicitly suggested by our feedback rule in Section 4.3. In both channels we apply the feedback rule to two states $\tilde{s}'$ and $\tilde{s}''$ as follows.

$$X_i = f^*(V_i, S_i) \triangleq \begin{cases} V_i, & S_i \in \{(00), (10)\} \\ 0, & S_i \in \{(01), (11)\}, \end{cases} \tag{4.90}$$

where $\tilde{s}_1 = (01)$ and $\tilde{s}_2 = (11)$. Applying this feedback strategy to the channel with state transition matrix

$$\mathbf{\Pi} = \begin{bmatrix} .80 & .20 & 0 & 0 \\ 0 & 0 & .50 & .50 \\ .78 & .22 & 0 & 0 \\ 0 & 0 & .50 & .50 \end{bmatrix} \tag{4.91}$$

and with stationary probabilities

$$\boldsymbol{\pi} = [.565, .145, .145, .145], \tag{4.92}$$

results in the strict increase seen in Figure 4.8. We now alter the state transition probabilities slightly, so that we no longer have a uniform distribution on the probabilities

$P_{Z_i|S_i}(z_i|\tilde{s}')$. More specifically,

$$\mathbf{\Pi}^* = \begin{bmatrix} .80 & .20 & 0 & 0 \\ 0 & 0 & .45 & .55 \\ .80 & .20 & 0 & 0 \\ 0 & 0 & .50 & .50 \end{bmatrix} \tag{4.93}$$

and with stationary probabilities

$$\boldsymbol{\pi}^* = [.563, .141, .141, .155]. \tag{4.94}$$

We have not proven analytically that this channel must experience an increase in constrained capacity with feedback, but we can see in Figure 4.9, as in Figure 4.3 for the first order case, that a numerical increase is achieved despite the fact that no increase is guaranteed.

Remark: These numerical examples are not the best we can do. Using the closed form expressions for $\tilde{\beta}^{lb}$ and the capacity C, we can improve the lower bound to $C_{FB}(\beta)$. In fact, for $\beta > \tilde{\beta}^{lb} = (1 - P_S(\tilde{s}))\beta_{max}$ we have that $C_{FB}(\beta) = C$. Furthermore, by the strict concavity of the capacity-cost function, the tangent line from the point $(\tilde{\beta}^{lb}, C)$ to the graph of $C_n^{lb}(\beta)$ also represents an improvement in the lower bound. As we achieved our goal of demonstrating the increase in the capacity-cost function with feedback, we need not further complicate the following graphs.

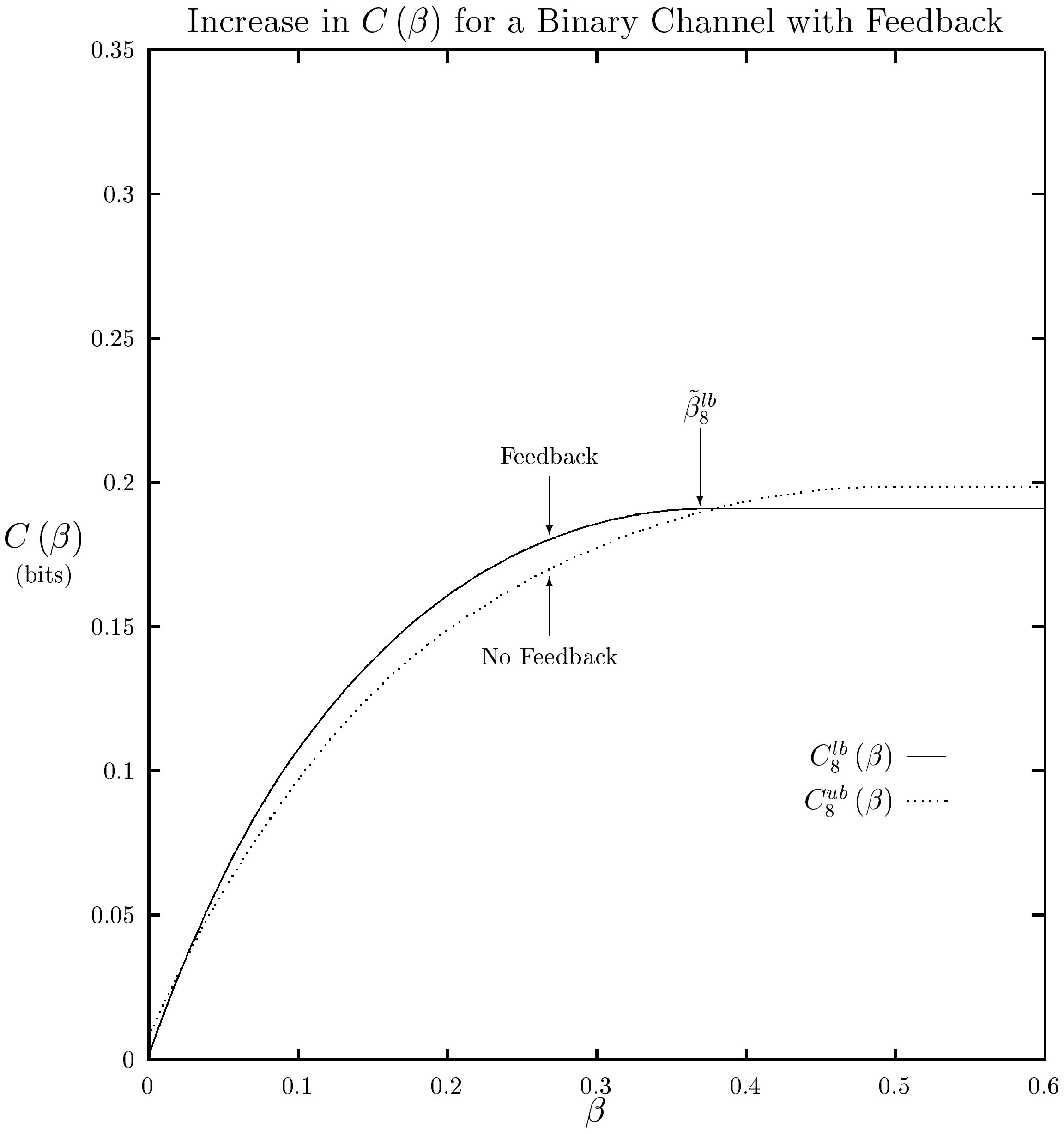


Figure 4.2: Nonlinear Feedback Increase: $C_8^{lb}(\beta) > C_8^{ub}(\beta)$ for $\beta < \tilde{\beta}_8^{lb} = .375$, using first order binary Markov noise ($\alpha = .2$ and $\varrho = .5$) and cost function: $b(0) = 0, b(1) = 1$.

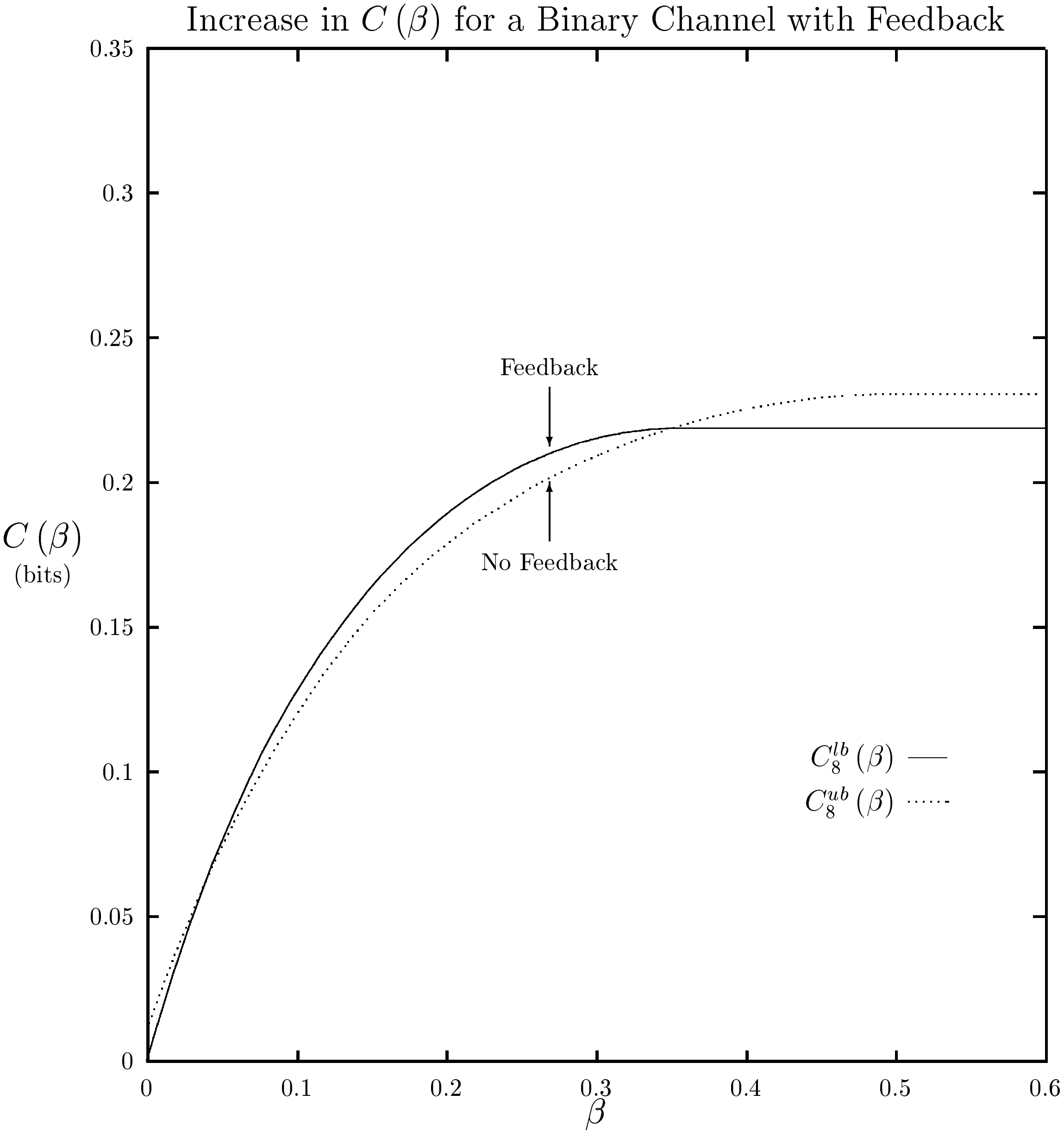


Figure 4.3: Nonlinear Feedback Increase: $C_8^{lb}(\beta) > C_8^{ub}(\beta)$ for $\beta < \beta_{max}^{lb(8)} = .375$, using first order binary Markov noise ($\alpha = .18$ and $\varrho = .45$) and cost function: $b(0) = 0, b(1) = 1$.

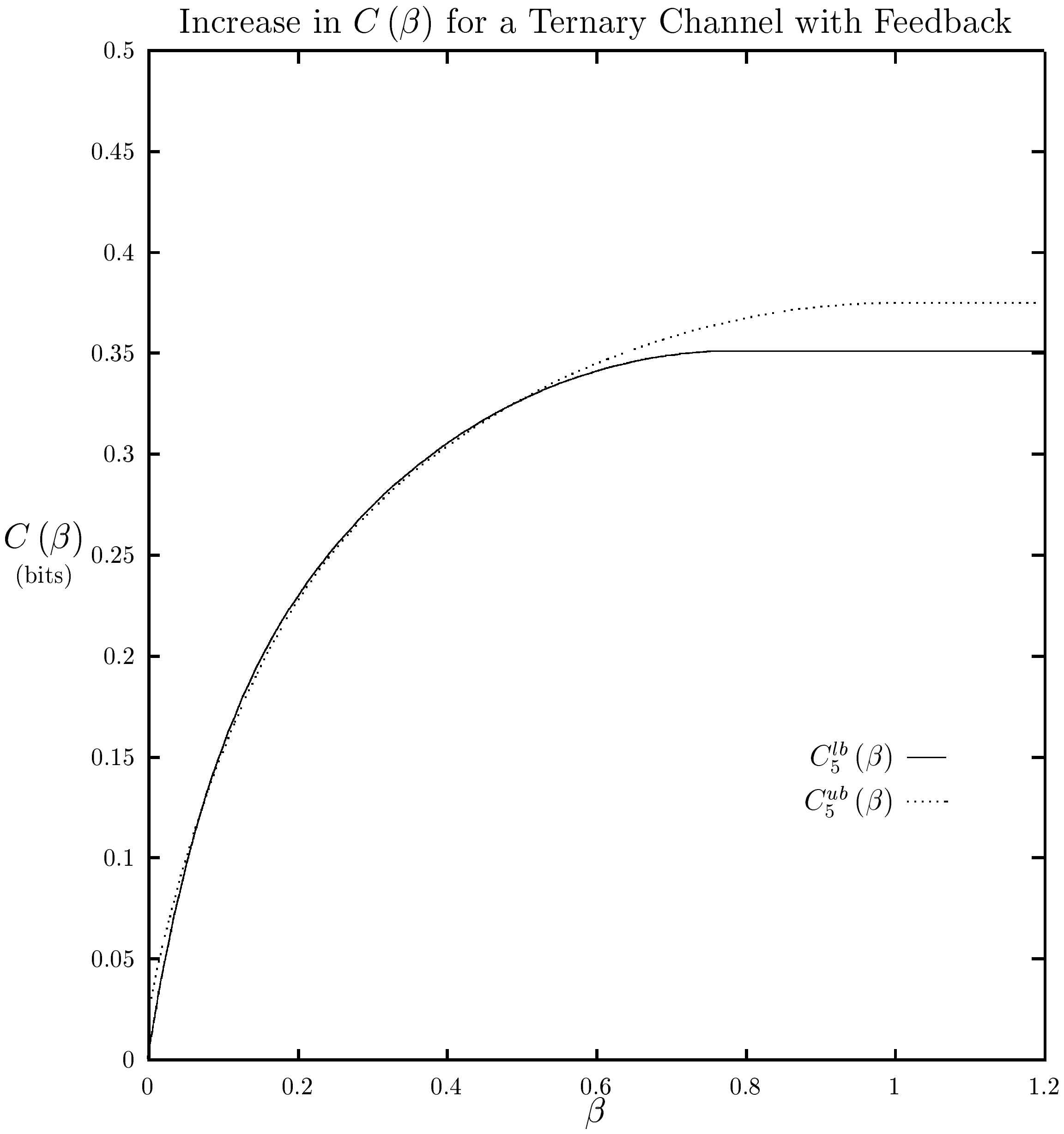


Figure 4.4: Increase of $C_5^{lb}\left(\beta\right)$ over $C_5^{ub}\left(\beta\right)$ for a 3-ary channel with 1st order Markov noise defined in Example 4.2.

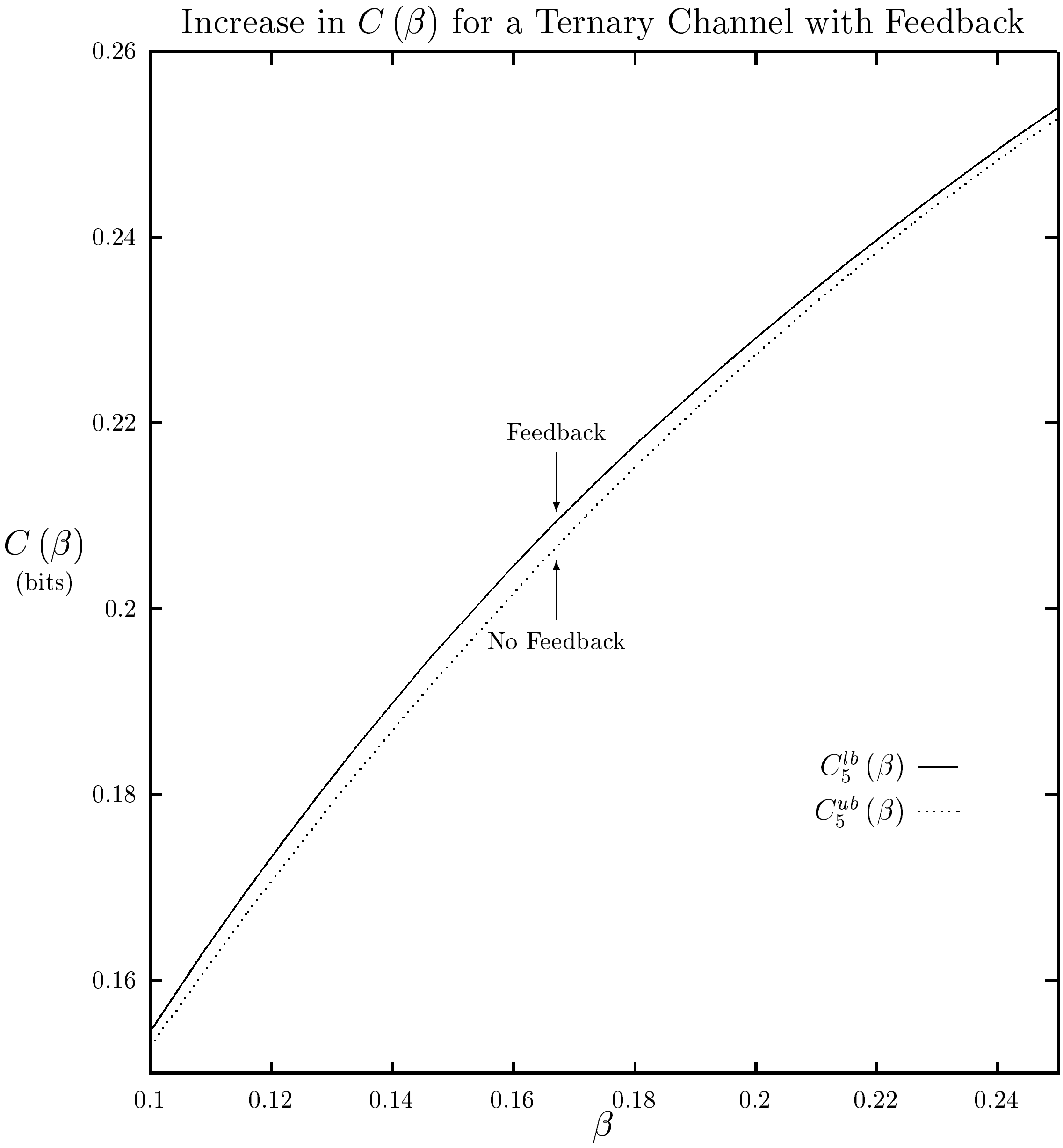


Figure 4.5: Zoom view of Figure 4.4: $C_5^{lb}\left(\beta\right) > C_5^{ub}\left(\beta\right)$ for $0.10 < \beta < 0.25$ for the channel defined in Example 4.2.

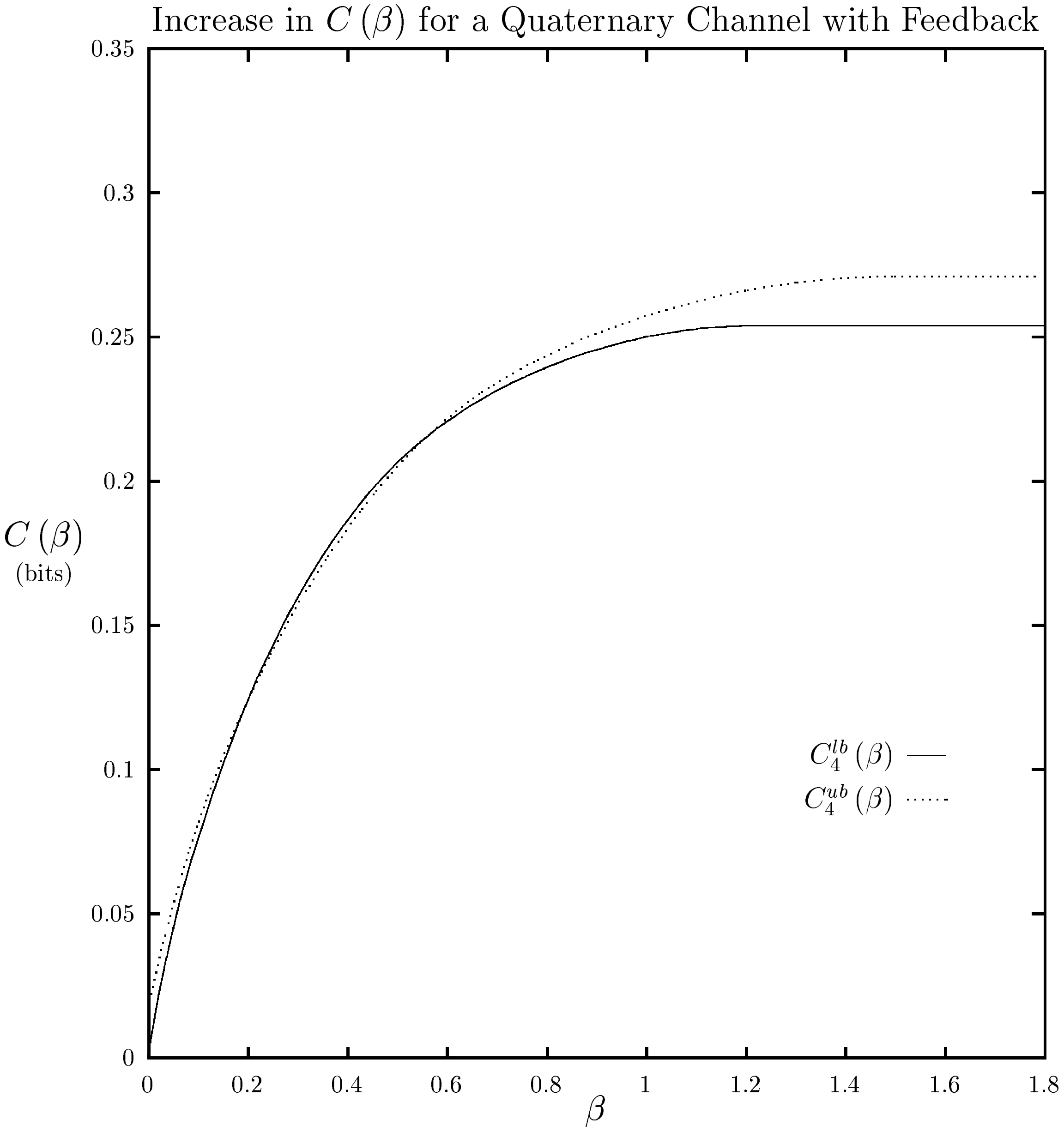


Figure 4.6: Increase of $C_4^{lb}(\beta)$ over $C_4^{ub}(\beta)$ for a 4-ary channel with 1st order Markov noise defined in Example 4.3.

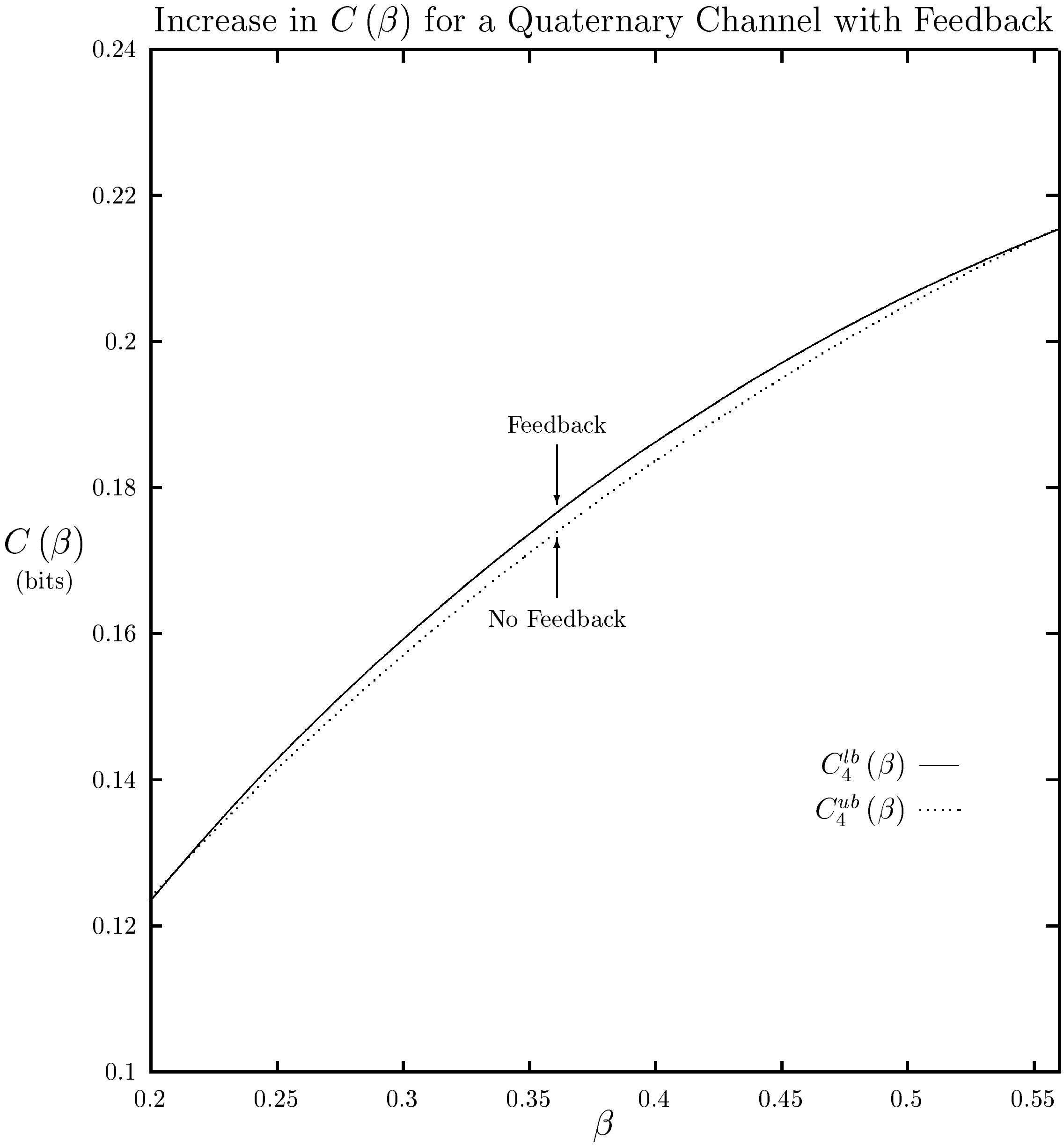


Figure 4.7: Zoom view of Figure 4.6: $C_4^{lb}\,(\beta) > C_4^{ub}\,(\beta)$ for $0.21 < \beta < 0.55$ for the channel with defined in Example 4.3.

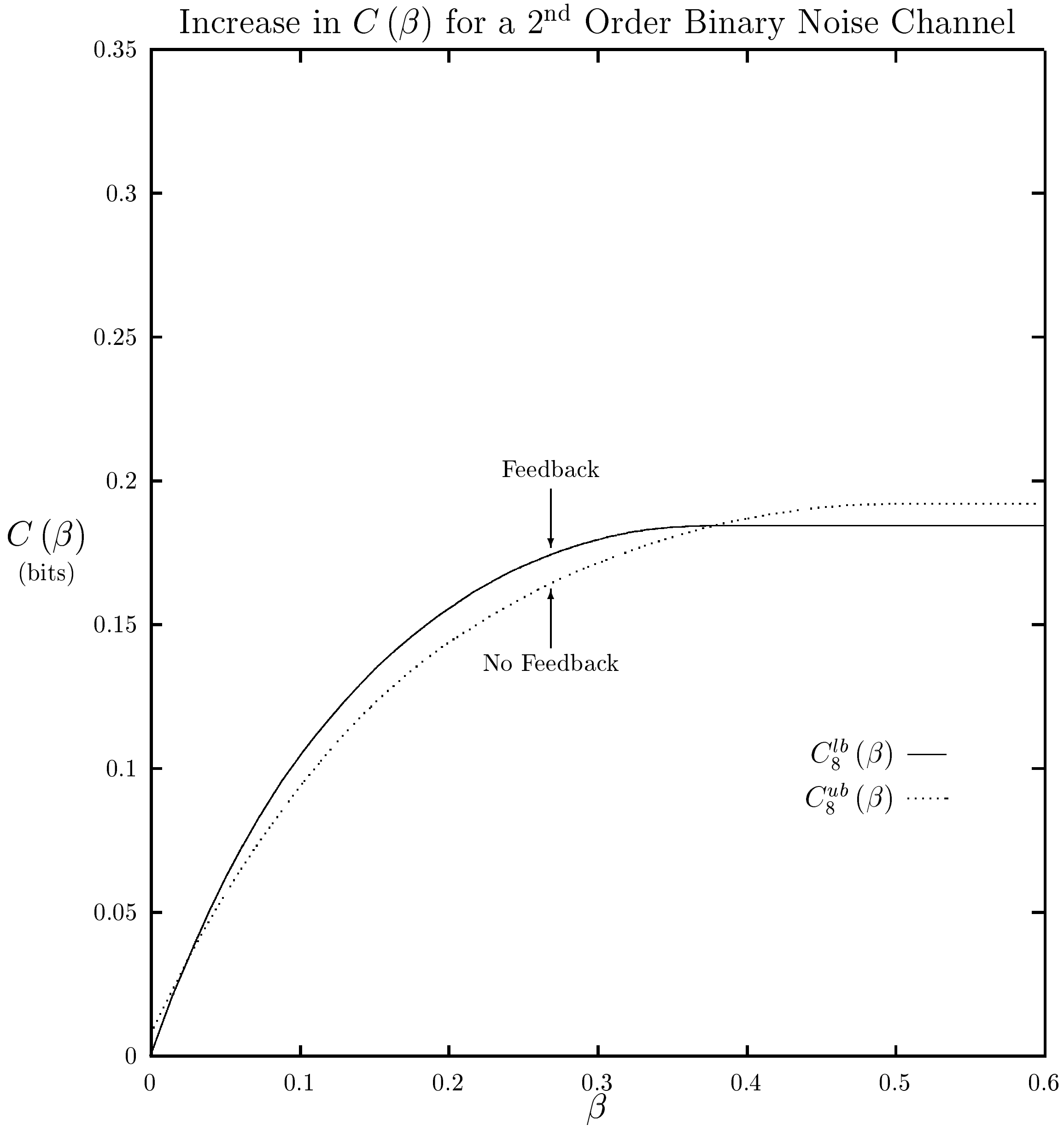


Figure 4.8: Increase in $C_8^{lb}(\beta)$ over $C_8^{ub}(\beta)$ for the channel with Markov noise $\mathbf{\Pi}$ given in Example 4.4.

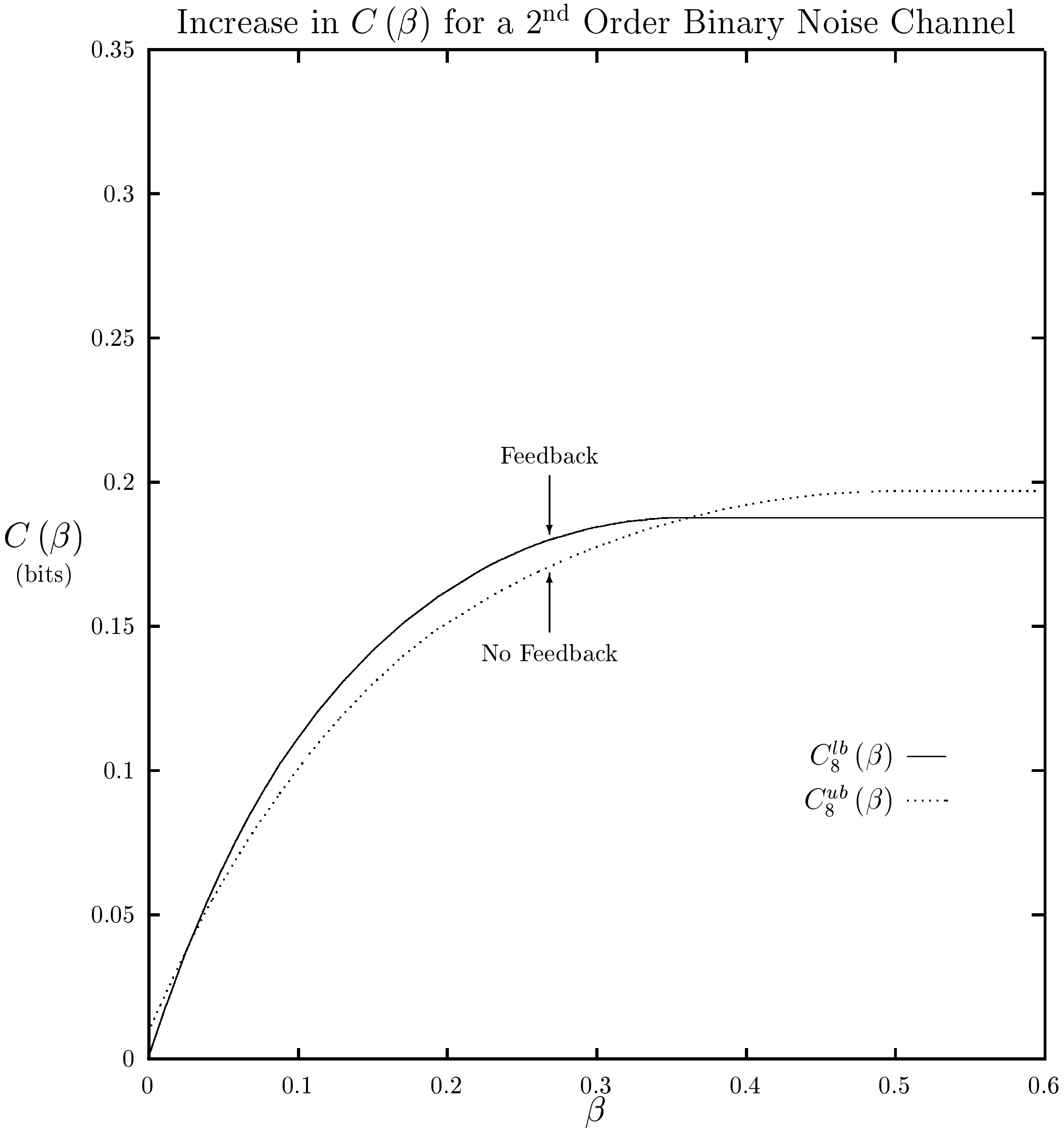


Figure 4.9: Increase in $C_8^{lb}(\beta)$ over $C_8^{ub}(\beta)$ for the channel with Markov noise $\mathbf{\Pi}^*$ given in Example 4.4.

Chapter 5

Conclusions and Future Work

5.1 Summary

In this thesis we have implemented the bounding techniques of Alajaji and Blahut for modulo q channels with stationary ergodic finite alphabet Markov noise. Numerical results verify that both bounds are tight, yet converge slowly in block length. An additional block length independent bound, Mrs. Gerber's Lemma, was also implemented on mod 2 channels. Computation of the bounds for block length n provided an envelope on the capacity-cost function.

A model for channels with time invariant feedback was then developed. We derived a lower bound to the capacity-cost function of this channel by showing achievability. This bound is also a computable by Blahut's algorithm. We also developed one feedback scheme and a class of Markov noise sources which result in an increase in the capacity-cost function. We show analytically that this feedback scheme gives

$$C_{FB}\left(\beta\right) > C\left(\beta\right) \tag{5.1}$$

for $0 < \beta < \beta_{max}$. Numerically, we are able to show

$$C_n^{lb}\left(\beta\right) > C_n^{ub}\left(\beta\right) \tag{5.2}$$

for $0 < \beta_1 < \beta < \beta_2 < \beta_{max}$, where $C_n^{lb}(\beta)$ is the lower bound on the n^{th} capacity-cost function of the feedback channel and $C_n^{ub}(\beta)$ is Alajaji's upper bound on the n^{th} capacity-cost function of the non-feedback channel. Therefore, we can now state that in some instances, feedback increases the capacity-cost function of discrete additive Markov channels.

All the numerical results were obtained with a C++ program capable not only of computing Blahut's algorithm for arbitrary input and output alphabets, but also of computing the stationary distribution of a k^{th} order Markov noise process with q^k states.

5.2 Future Work

This thesis leads naturally into some other areas of information theory research.

- A generalization of this result to all stationary Markov noise processes is still necessary.
- Treating the case of real addition channels with feedback may show an increase in the channel capacity.
- Expanding the results on the effect of feedback to the channel reliability function of channels with memory.

In preparing this thesis, a large amount of effort went into examining other types of Markov noise with few results. By searching over all possible feedback strategies, we may be able to find an increase in the capacity-cost function in general.

The questions surrounding real addition channels were investigated. The feedback strategy employed for the mod q channel is not effective for the real adder channel since it is not symmetric. It is postulated, however, that feedback should increase the actual capacity of these channels. One new direction with potential would be to examine variable length feedback codes for the real adder channel.

Computation of the channel reliability function is given in [3]. Using bounding techniques discussed in [7], it could be possible to prove that feedback increases the reliability function of channels with memory.

Appendix A

Basic Information Theory Concepts

In this Appendix, we review some key information theory definitions and theorems.

A.1 Information Theory Definitions [11]

Definition A.1 The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x). \tag{A.1}$$

Definition A.2 The *joint entropy* of a pair of discrete random variables (X, Y) with a joint distribution $P_{X,Y}(x, y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) \log P_{X,Y}(x,y). \tag{A.2}$$

Definition A.3 If (X, Y) is a pair of discrete random variables with joint distribution $P_{X,Y}(x, y)$, then the *conditional entropy* $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) && \text{(A.3)} \\ &= -\sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) \log P_{Y|X}(y|x). && \text{(A.4)} \end{aligned}$$

Definition A.4 The *relative entropy* between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{A.5}$$

Definition A.5 Consider two random variables X and Y with a joint probability mass function $P_{X,Y}(X,Y)$. The *mutual information* $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $P_X(x)P_Y(y)$, i.e.,

$$\begin{aligned} I(X;Y) &= D(P_{X,Y}(x,y)||P_X(x)P_Y(y)) && \text{(A.6)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}. && \text{(A.7)} \end{aligned}$$

These definitions can be conveniently described in the following Venn diagram.

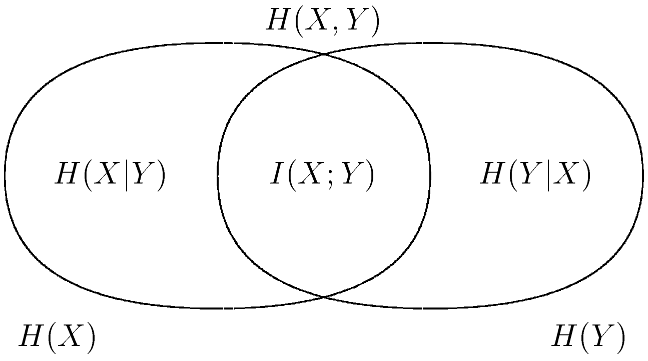


Figure A.1: The relationship between entropy and mutual information.

A.2 Theorems

Theorem A.1 ([11]) $H(X) \leq \log |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution of $\mathcal{X}$.

Theorem A.2 ([11]) For any two random variable X and Y,

$$I(X;Y) \geq 0 \tag{A.8}$$

with equality if and only if X and Y are statistically independent (i.e., $P_{X,Y}(x,y) = P_X(x)P_Y(y)$).

We now wish to present a final result concerning the definition of the capacity-cost function given in Equation (2.42). It is not obvious that the limit as $n \to \infty$ of $C_n(\beta)$ is equivalent to the supremum over all n, and so we shall now prove it. We begin by stating the following lemma.

Lemma A.1 ([12]) Let $\{a_N\}$, $N \in \{0, 1, \ldots\}$ be a bounded sequence of numbers and let

$$\bar{a} = \sup_N a_N \tag{A.9}$$

and

$$\underline{a} = \inf_N a_N. \tag{A.10}$$

(By a bounded sequence we mean that $\bar{a} < \infty$ and $\underline{a} > -\infty$.) Assume that for all $n \geq 1$, and $N > n$,

$$a_N \geq \frac{n}{N} a_n + \frac{N-n}{N} a_{N-n}, \tag{A.11}$$

then

$$\lim_{N \to \infty} a_N = \bar{a}. \tag{A.12}$$

Conversely, if for all $n \geq 1$, and $N > n$,

$$a_N \leq \frac{n}{N} a_n + \frac{N-n}{N} a_{N-n}, \tag{A.13}$$

we have

$$\lim_{N \to \infty} a_N = \underline{a}. \tag{A.14}$$

The following theorem and proof are exactly dual to the results concerning $R_L(d^*)$ in Chapter 9 of [12]. It asserts that the limit in Equation (2.42) exists and also that, for any N, $C_N(\beta)$ is a lower bound to $C(\beta)$.

Theorem A.3 For a discrete (finite alphabet) stationary channel

$$\sup_N C_N(\beta) = \lim_{N\to\infty} C_N(\beta). \tag{A.15}$$

Proof of Theorem A.3 First observe that $\{C_N(\beta)\}$ is a bounded sequence since

$$0 \le C_N(\beta) \le \log|\mathcal{X}|, \quad \forall\ N. \tag{A.16}$$

Let l and n be arbitrary positive integers and, for a given $\beta \ge \beta_{min}$, let $P^*_{X^l}$ and $P^*_{X^n}$ be β-admissible input distributions that achieve $C_l(\beta)$ and $C_n(\beta)$ respectively. Furthermore, let $N = n + l$, and choose

$$P_{X^N}(x^N) = P^*_{X^n}(x^n)P^*_{X^l}(x^l), \tag{A.17}$$

where $x^N = (x_1, \ldots, x_N)$, $x^n = (x_1, \ldots, x_n)$ and $x^l = (x_{n+1}, \ldots, x_N)$, and where X^N, X^n and X^l are the respective random vectors of these sequences. Let Y^N, Y^n and Y^l be the corresponding output random vectors. Since $P_{X^N}(x^N)$ is not necessarily the input distribution that achieves $C_N(\beta)$, we have

$$\begin{aligned} NC_N(\beta) &\ge I(X^nX^l;Y^nY^l) \\ &= I(X^n;Y^nY^l) + I(X^l;Y^nY^l|X^n) \end{aligned} \tag{A.18}$$

The first term in (A.18) is lower bounded by

$$I(X^n;Y^nY^l) \ge I(X^n;Y^n) = nC_n(\beta) \tag{A.19}$$

since $P^*_{X^n}(x^n)$ achieves the n^{th} capacity-cost. The second term in (A.18) can be rearranged as

$$I(X^l;Y^nY^l|X^n) = I(X^l;Y^nY^lX^n) - I(X^l;X^n) \tag{A.20}$$

$$= \quad I(X^l; Y^n Y^l X^n) \tag{A.21}$$

$$\geq \quad I(X^l; Y^l). \tag{A.22}$$

The second equality uses the statistical independence of X^n and X^l from (A.17). Since the channel is understood to be stationary, the joint probability mass function on X^l and Y^l is invariant to time shifts, and $P^*_{X^l}(x^l)$ achieves the l^{th} capacity-cost. Therefore

$$I(X^l; Y^l) = lC_l\left(\beta\right). \tag{A.23}$$

Using (A.19) and (A.23) in (A.18), we have

$$NC_N\left(\beta\right) \geq nC_n\left(\beta\right) + lC_l\left(\beta\right) \tag{A.24}$$

or

$$C_N\left(\beta\right) \geq \frac{n}{N} C_n\left(\beta\right) + \frac{N-n}{N} C_{N-n}\left(\beta\right). \tag{A.25}$$

Applying the results of Lemma A.1 proves the theorem. □

Appendix B

Justification of Blahut's Algorithm

In Blahut's published paper [8], much of the justification for the capacity cost function algorithm was cited either as a trivial modification on proofs found elsewhere [12] or as trivial modification on proofs found within different sections of the paper itself. Since we are interested in applying the algorithm to channels with memory, a block approach is required. The n^{th} capacity-cost function of a block length n channel with input alphabet of size r and output alphabet of size t is equal to the capacity-cost function of a corresponding memoryless channel with input alphabet of size r^n and output alphabet of size t^n. It is both useful and necessary that we incorporate into this appendix a complete bottom up proof for the legitimacy of the n^{th} capacity-cost function algorithm, which bounds $C_n(\beta)$ from above by $C_n^U(\beta)$ and from below by $C_n^L(\beta)$. Continuing with the notation introduced in Section 3.1.1, we index the input vectors using $j \in \{0, 1, \ldots, N-1\}$ where $N = r^n$, and we index the output n-tuples using $k \in \{0, 1, \ldots, M-1\}$ where $M = t^n$. Throughout this appendix, the units for mutual information and capacity-cost functions are in nats.

Definition B.1 An *expense schedule* for a channel is a vector e of length N whose j^{th} component

$$e_j \triangleq \frac{1}{n} b\left(x^n = j\right) = \frac{1}{n} \sum_{i=1}^{n} b\left(x_i\right) \tag{B.1}$$

is called the per symbol expense of using the j^{th} channel input word whose block representation is $x^n = (x_1, \ldots, x_n)$.

Definition B.2 The n^{th} *capacity* at cost β is

$$C_n\left(\beta\right) = \frac{1}{n} \max_{p \in \tau_n(\beta)} \sum_{j,k} p_j Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} \tag{B.2}$$

where

$$\tau_n(\beta) = \{p \in \mathbf{P}^N : \sum_j p_j e_j \leq \beta\}, \tag{B.3}$$

and where

$$\mathbf{P}^N \triangleq \left\{ p = [p_1, p_2, \ldots, p_N] \in \mathbf{R}^n : p_j \geq 0, \ \sum_{j=1}^{N} p_j = 1 \right\}. \tag{B.4}$$

A few obvious observations can be made with regards to these definitions using the notation in Chapters 2 and 3. Equation (B.2) is a maximum and not a supremum since $\tau_n(\beta)$ is compact so long as it is nonempty. Therefore, $I(p; Q)$, in Equation (3.5), attains its maximum on $\tau_n(\beta)$.

We can also assume without loss of generality that $\beta_{min} = 0$ and thus $C_n\left(\beta\right)$ is defined for all $\beta \geq 0$. This is equivalent to assuming that $b_{min} = 0$. Were that not the case, an appropriate scalar constant added to all letter costs would shift the graph of $C_n\left(\beta\right)$ along the horizontal axis.

Lemma 2.1 $C_n\left(\beta\right)$ and $C\left(\beta\right)$ are *concave* and *non-decreasing* functions of β, for $\beta \geq \beta_{min}$.

Proof of Lemma 2.1 We begin by showing the results for $C_n(\beta)$. If $\beta' > \beta$ then we have $\tau_n(\beta) \subset \tau_n(\beta')$ and $C_n(\beta) \leq C_n(\beta')$, making $C_n(\beta)$ a monotonic nondecreasing function.

We now demonstrate concavity. $C_n(\beta)$ is concave if

$$C_n(\lambda\beta' + (1-\lambda)\beta'') \geq \lambda C_n(\beta') + (1-\lambda)C_n(\beta''), \tag{B.5}$$

where $\beta', \beta'' \geq \beta_{min}$, and $\lambda \in [0,1]$. Allow p' and p'' to achieve $(\beta', C_n(\beta'))$ and $(\beta'', C_n(\beta''))$ respectively. If $p^* = \lambda p' + (1-\lambda)p''$, then the expected per letter cost of p^* is

$$\sum_j [\lambda p'_j + (1-\lambda)p''_j]e_j \leq \lambda\beta' + (1-\lambda)\beta''. \tag{B.6}$$

This implies that $p^* \in \tau_n(\lambda\beta' + (1-\lambda)\beta'')$ so that $C_n(\lambda\beta' + (1-\lambda)\beta'') \geq \frac{1}{n}I(p^*;Q)$. Hence

$$\begin{aligned}
&C_n(\lambda\beta' + (1-\lambda)\beta'') - \lambda C_n(\beta') - (1-\lambda)C_n(\beta'') && \text{(B.7)}\\
&\geq \frac{1}{n}\left[I(p^*;Q) - \lambda I(p';Q) - (1-\lambda)I(p'';Q)\right] && \text{(B.8)}\\
&= \frac{1}{n}\left[\lambda\sum_{j,k} p'_j Q_{k|j}\ln\frac{\sum_j p'_j Q_{k|j}}{\sum_j p^*_j Q_{k|j}} + (1-\lambda)\sum_{j,k} p''_j Q_{k|j}\ln\frac{\sum_j p''_j Q_{k|j}}{\sum_j p^*_j Q_{k|j}}\right] && \text{(B.9)}\\
&\geq \frac{1}{n}\left[\lambda\left(\sum_{j,k} p'_j Q_{k|j} - \sum_{j,k} p^*_j Q_{k|j}\right) + (1-\lambda)\left(\sum_{j,k} p''_j Q_{k|j} - \sum_{j,k} p^*_j Q_{k|j}\right)\right] && \text{(B.10)}\\
&= 0 && \text{(B.11)}
\end{aligned}$$

where the second inequality follows from the fact that $\ln(x) \geq 1 - \frac{1}{x}$.

The results follow for $C(\beta)$ since the limit of a sequence of non-decreasing concave functions is itself non-decreasing and concave. □

Corollary B.1 $C_n(\beta)$ and $C(\beta)$ are continuous except at β_{min}.

Proof of Corollary B.1 $C_n(\beta)$ and $C(\beta)$ are monotonic and convex. □

Corollary B.2

$$\lim_{\beta \to \beta_{max}^{(n)}} C_n(\beta) = C_n \tag{B.12}$$

where C_n is the n^{th} channel capacity, $\beta_{max}^{(n)} = \sum_j p_j^* e_j$, and p^* achieves capacity.

Proof of Corollary B.2 The result follows using the definition of $\beta_{max}^{(n)}$ and the continuity of $C_n(\beta)$. □

Corollary 2.1 $C_n(\beta)$ is *strictly increasing* in β for $\beta_{min} \leq \beta \leq \beta_{max}^{(n)}$. Therefore, $C(\beta)$ is *strictly increasing* in β for $\beta_{min} \leq \beta \leq \beta_{max}$.

Proof of Corollary 2.1 The results follow by the concavity of $C_n(\beta)$ and $C(\beta)$. □

Corollary B.3 If p^* achieves $(\beta, C_n(\beta))$ and $\beta \leq \beta_{max}^{(n)}$ then

$$\beta(p^*) \triangleq \sum_j p_j^* e_j = \beta. \tag{B.13}$$

Proof of Corollary B.3 $C_n(\beta)$ is strictly increasing if $\beta \leq \beta_{max}^{(n)}$. □

Theorem B.1 If p' and p'' both achieve the point $(\beta, C_n(\beta))$, then so will

$$p = \lambda p' + (1-\lambda)p'', \quad \text{forall } \lambda \in [0,1]. \tag{B.14}$$

Proof of Theorem B.1

$$\beta(p) = \sum_j [\lambda p_j' + (1-\lambda)p_j'']e_j = \lambda\beta + (1-\lambda)\beta = \beta \tag{B.15}$$

hence, $p \in \tau_n(\beta)$ and

$$C_n(\beta) \geq \frac{1}{n} I(p;Q) \geq \frac{1}{n}\left[\lambda I(p';Q) + (1-\lambda) I(p'';Q)\right] = C_n(\beta). \tag{B.16}$$

□

We now develop the transition between non-parameterized and parameterized n^{th} capacity-cost formulas, which allows us to perform our maximization over the entire probability space instead of merely $\tau_n(\beta)$.

Theorem B.2 For a fixed block length n and given $0 \leq \beta \leq \beta_{max}^{(n)}$, $C_n(\beta)$ can be written in terms of a parameter $s \in [0, \infty)$ by

$$C_n(\beta_s) = s\beta_s + V_s \tag{B.17}$$

$$\beta_s = \sum_j p_j^* e_j, \tag{B.18}$$

where

$$V_s = \max_{p \in \mathbf{P}^N} \left\{ \frac{1}{n} \sum_{j,k} p_j Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s \sum_j p_j e_j \right\} \tag{B.19}$$

and p^* achieves this maximum.

Proof of Theorem B.2 Clearly, any point $(\beta_s, C_n(\beta_s))$ satisfying the above equations lies on the graph of $C_n(\beta)$. By showing that every point on $C_n(\beta)$ can be expressed in this fashion we complete the proof.

By Lemma 2.1, $C_n(\beta)$ is concave, and differentiable except possibly at a countable number of points. It also has a left and right derivative everywhere. Given a cost β, let s be the left derivative of $C_n(\beta)$ at β. Then, using the concavity of $C_n(\beta)$, for any β'

$$C_n(\beta') \leq C_n(\beta) + s(\beta' - \beta). \tag{B.20}$$

Suppose the parameter s generates some point on $C_n(\beta)$ whose left derivative is s. Label this point $(\beta_s, C_n(\beta_s))$. Then

$$C_n(\beta_s)$$

$$= \max_{p \in \mathbf{P}^n} \left\{ \frac{1}{n} \sum_{j,k} p_j Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s \sum_j p_j e_j + s\beta_s \right\} \tag{B.21}$$

$$\geq \max_{p \in \{p: \sum_j p_j e_j = \beta\}} \left\{ \frac{1}{n} \sum_{j,k} p_j Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s \sum_j p_j e_j + s\beta_s \right\} \tag{B.22}$$

$$= C_n(\beta) - s\beta + s\beta_s, \tag{B.23}$$

or

$$C_n(\beta_s) \geq C_n(\beta) + s(\beta_s - \beta). \tag{B.24}$$

Therefore, $C_n(\beta_s) = C_n(\beta) + s(\beta_s - \beta)$, which implies that either $\beta_s = \beta$ (in the case of strict concavity) or they are connected by a line segment of slope s. □

Corollary B.4 If $C_n(\beta)$ is strictly concave in the neighbourhood of some point, then the value of s that generates this point generates only this point.

Corollary B.5 If s_1 and s_2 are the left and right derivatives at a point β, then s generates $(\beta, C_n(\beta))$ if and only if $s \in [s_1, s_2]$.

The above parameterization was used by Blahut to restate the problem from a maximization over $p \in \tau_n(\beta)$ to a maximization over all probability vectors. We now derive Blahut's algorithm for a block memoryless channel of length n.

Theorem B.3 Let

$$J(p, Q, P) = \frac{1}{n} \sum_j \sum_k p_j Q_{k|j} \ln \frac{P_{j|k}}{p_j} - s \sum_j p_j e_j. \tag{B.25}$$

Then

a)

$$C_n(\beta) = s\beta + \max_P \max_p J(p, Q, P), \tag{B.26}$$

where

$$\beta = \sum_j p_j^* e_j \tag{B.27}$$

and p^* achieves the above maximum.

b) For fixed p, $J(p, Q, P)$ is maximized by

$$P_{j|k} = \frac{p_j Q_{k|j}}{\sum_j p_j Q_{k|j}}. \tag{B.28}$$

c) For fixed P, $J(p, Q, P)$ is maximized by

$$p_j = \frac{\exp\left\{\sum_k Q_{k|j} \ln P_{j|k} - nse_j\right\}}{\sum_j \exp\left\{\sum_k Q_{k|j} \ln P_{j|k} - nse_j\right\}}. \tag{B.29}$$

Proof of Theorem B.3

a) We need only to demonstrate that

$$I(p; Q) = \max_P \sum_{j,k} p_j Q_{k|j} \ln \frac{P_{j|k}}{p_j}. \tag{B.30}$$

Doing this requires an application of Bayes' rule to find a good guess P^* for P, and also requires the determination of the output distribution q. Let

$$P^*_{j|k} = \frac{p_j Q_{k|j}}{\sum_j p_j Q_{k|j}} \tag{B.31}$$

and

$$q_k = \sum_j p_j Q_{k|j} \tag{B.32}$$

which, if we have choosen P^* correctly, allows us to write

$$I(p; Q) = \sum_{j,k} q_k P^*_{j|k} \ln \frac{P^*_{j|k}}{p_j}. \tag{B.33}$$

To justify our choice of P^*, observe that

$$\begin{aligned} I(p; Q) - \sum_{j,k} p_j Q_{k|j} \ln \frac{P_{j|k}}{p_j} &= \sum_{j,k} q_k P^*_{j|k} \ln \frac{P^*_{j|k}}{P_{j|k}} && \text{(B.34)} \\ &\geq \sum_{j,k} q_k P^*_{j|k} - \sum_{j,k} q_k P_{j|k} && \text{(B.35)} \\ &= 0 && \text{(B.36)} \end{aligned}$$

with equality iff $P_{j|k} = P^*_{j|k}$. The above inequality is an application of $\ln x \geq 1 - \frac{1}{x}$ with equality iff $x = 1$.

b) This is an obvious consequence of the equality condition in a) that was just proved.

c) If for some k, $P_{j|k} = 0$, then p_j should be set equal to 0 in order to maximize J as it is. Such a letter j can be omitted from further consideration. $J(p, Q, P)$ can now be maximized over p by temporarily ignoring the constraint $p_j \geq 0$, and using a Lagrange multiplier to constrain

$$\sum_j p_j = 1. \tag{B.37}$$

Setting all partial derivatives equal to zero gives

$$\frac{\partial}{\partial p_j} \left\{ \frac{1}{n} \sum_{j,k} p_j Q_{k|j} \ln \frac{P_{j|k}}{p_j} - s \sum_j p_j e_j + \lambda (\sum_j p_j - 1) \right\} = 0, \tag{B.38}$$

or

$$-\frac{1}{n} \ln p_j - \frac{1}{n} + \frac{1}{n} \sum_k Q_{k|j} \ln P_{j|k} - s e_j + \lambda = 0. \tag{B.39}$$

Hence

$$p_j = \frac{\exp\left(\sum_k Q_{k|j} \ln P_{j|k} - n s e_j\right)}{\sum_j \exp\left(\sum_k Q_{k|j} \ln P_{j|k} - n s e_j\right)}, \tag{B.40}$$

where λ is selected so that Equation (B.37) holds. Conveniently, p_j is always positive, so we need not concern ourselves with our previous simplification on the constraint $p_j \geq 0$.

□

Theorem B.3 expresses the computation of $C_n(\beta)$ as a maximization problem over p and P. The following corollary is stated here because it is an immediate consequence of Theorem B.3 and because its convenient form motivates the remainder of this appendix

Corollary B.6 If p is a probability distribution acheiving the n^{th} capacity at cost β, then for some $s \in [0, \infty)$

$$p_j = \frac{p_j \exp\left(\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - nse_j\right)}{\sum_j p_j \exp\left(\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - nse_j\right)}. \tag{B.41}$$

Proof of Corollary B.6 This result follows from the conditions guaranteed by b) and c) above. □

The equation in Corollary B.6 suggests, under appropriate conditions, that any p can be used on the right hand side to generate a *better* p on the left hand side.

Corollary B.7 As a function of the slope s, a parametric solution for the n^{th} capacity-cost function is

$$C_n\left(\beta_s\right) = s\beta_s + \max_P \frac{1}{n}\left[\ln \sum_j \exp\left(\sum_k Q_{k|j} \ln P_{j|k} - nse_j\right)\right] \tag{B.42}$$

$$\beta_s = \sum_j e_j \frac{\exp\left(\sum_k Q_{k|j} \ln P^*_{j|k} - nse_j\right)}{\sum_j \exp\left(\sum_k Q_{k|j} \ln P^*_{j|k} - nse_j\right)}, \tag{B.43}$$

where $P^*_{j|k}$ achieves this maximum.

Proof of Corollary B.7 This point arises from the substitution of Theorem B.3c) into Theorem B.3a). □

Corollary B.8

$$C_n\left(\beta\right) = \min_{s\in[0,\infty)} \max_P \left[s\beta + \frac{1}{n} \ln \sum_j \exp\left(\sum_k Q_{k|j} \ln P_{j|k} - nse_j\right)\right]. \tag{B.44}$$

Proof of Corollary B.8 Let p^* achieve $C_n(\beta)$ and let β_s be the cost generated by parameter s. Then

$$\begin{aligned}
C_n(\beta) - \max_P \left[s\beta + \frac{1}{n} \ln \sum_j \exp\left(\sum_k Q_{k|j} \ln P_{j|k} - nse_j \right) \right] && \text{(B.45)} \\
= \quad C_n(\beta) - s\beta + s\beta_s && \\
- \max_P \left[s\beta_s + \frac{1}{n} \ln \sum_j \exp\left(\sum_k Q_{k|j} \ln P_{j|k} - nse_j \right) \right] && \text{(B.46)} \\
= \quad C_n(\beta) - s\beta - C_n(\beta_s) + s\beta_s && \text{(B.47)} \\
= \quad \frac{1}{n} \sum_{j,k} p_j^* Q_{k|j} \ln \frac{Q_{k|j} \exp(-nse_j)}{\sum_j p_j^* Q_{k|j}} && \\
- \max_p \frac{1}{n} \sum_{j,k} p_j Q_{k|j} \ln \frac{Q_{k|j} \exp(-nse_j)}{\sum_j p_j Q_{k|j}} && \text{(B.48)} \\
\leq \quad 0. && \text{(B.49)}
\end{aligned}$$

□

The conditions in Theorem B.4, commonly known as the Kuhn-Tucker conditions, are necessary and sufficient for achievability of $C_n(\beta)$ by input word distribution p [12].

Theorem B.4 The n^{th} capacity is achieved at cost β_s by a vector $p \in \mathbf{P}^N$ if and only if there exists a number V such that

$$\begin{aligned}
\frac{1}{n} \sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - se_j &= V, \quad p_j \neq 0 && \text{(B.50)} \\
\frac{1}{n} \sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - se_j &\leq V, \quad p_j = 0, && \text{(B.51)} \\
&&& \text{(B.52)}
\end{aligned}$$

where Q is the channel transition matrix, e is the expense vector and s parametrizes the cost. The constant $V = C_n(\beta) - s\beta$.

Proof of Theorem B.4 We will show that these conditions are the Kuhn-Tucker conditions, and are therefore, necessary and sufficient to prove achievability of the capacity-cost function for a given cost β_s. Since we wish to maximize

$$\frac{1}{n}I(p;Q) - s\sum_j p_j e_j = \frac{1}{n}\sum_{j,k} p_j Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s\sum_j p_j e_j, \tag{B.53}$$

we take partial derivatives with respect to the p_j's to yield

$$\frac{\partial}{\partial p_j}\{\frac{1}{n}I(p;Q) - s\sum_j p_j e_j\} = \frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j. \tag{B.54}$$

Since $I(p;Q)$ is convex $\cap$ in p and since the partial derivatives are continuous, Theorem 4.4.1 in [12] states that the Kuhn-Tucker conditions are valid. Therefore,

$$\begin{aligned}
\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j &= V, \qquad p_j \neq 0 && \text{(B.55)}\\
\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j &\leq V, \qquad p_j = 0, && \text{(B.56)}\\
& && \text{(B.57)}
\end{aligned}$$

where p achieves the capacity-cost. What remains is to find the appropriate constant V. Taking the expected value of both sides with respect to p gives $V = C_n(\beta) - s\beta$. □

Corollary B.9 The original Kuhn-Tucker conditions can be rewritten in a more illustrative form. Bringing terms to the right and taking exponents gives us

$$\begin{aligned}
\exp(-V)\exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j\right) &= 1, \qquad p_j \neq 0 && \text{(B.58)}\\
\exp(-V)\exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - s e_j\right) &\leq 1, \qquad p_j = 0. && \text{(B.59)}
\end{aligned}$$

Theorem 3.1 Let $s \in [0, \infty)$ be given, and for any $p \in \mathbf{P}^N$ let

$$c_j(p) = \exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - se_j\right). \tag{B.60}$$

Then if p^0 is any probability vector in $\mathbf{P}^N$ with all components strictly positive, the sequence of vectors resulting from

$$p_j^{r+1} = p_j^r \frac{c_j^r}{\sum_j p_j^r c_j^r} \tag{B.61}$$

has the properties that

$$\frac{1}{n} I(p^r; Q) \to C_n\left(\beta_s\right), \qquad \text{as } r \to \infty, \tag{B.62}$$

$$e(p^r) \to \beta_s, \qquad \text{as } r \to \infty, \tag{B.63}$$

where β_s is the average per letter cost of the point parametrized by s, and $I(\cdot;\cdot)$ and $C_n(\cdot)$ are measured in nats.

Proof of Theorem 3.1 Letting

$$V(p) = \frac{1}{n} I(p; Q) - se(p) = \sum_j p_j \ln c_j, \tag{B.64}$$

we show that $V(p^r)$ is increasing in r. Let

$$\begin{aligned} W(p^r) &= V(p^{r+1}) - V(p^r) && \text{(B.65)} \\ &= \sum_j p_j^r \frac{c_j^r}{\sum_j p_j^r c_j^r} \ln c_j^{r+1} - \sum_i p_i^r \ln c_i^r && \text{(B.66)} \\ &= \frac{1}{\sum_j p_j^r c_j^r}\left[\sum_i \sum_j p_i^r p_j^r c_j^r \ln c_j^{r+1}\right. \\ &\qquad \left. - \sum_i \sum_j p_i^r p_j^r c_j^r \ln c_i^r\right] && \text{(B.67)} \\ &= \frac{1}{\sum_j p_j^r c_j^r}\left[\sum_i p_i^r \sum_j p_j^r c_j^r \ln \frac{c_j^{r+1}}{c_i^r}\right] && \text{(B.68)} \end{aligned}$$

$$\geq \; 1 - \sum_j p_j^r \frac{c_j^r}{c_j^{r+1}} \tag{B.69}$$

with equality iff

$$\frac{c_j^{r+1}}{c_i^r} = 1 \;\; \forall \; i, \; j \text{ such that } p_i \neq 0 \neq p_j. \tag{B.70}$$

We now substitute the defining equation for c_j to get

$$\begin{aligned}
W(p^r) \;&\geq\; 1 - \sum_j p_j^r \exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{\sum_j p_j^{r+1} Q_{k|j}}{\sum_j p_j^r Q_{k|j}}\right) && \text{(B.71)}\\
&=\; 1 - \sum_j p_j^r \exp\left[\sum_k Q_{k|j} \ln \left(\frac{\sum_j p_j^{r+1} Q_{k|j}}{\sum_j p_j^r Q_{k|j}}\right)^{\frac{1}{n}}\right] && \text{(B.72)}\\
&\geq\; 1 - \sum_j p_j^r \sum_k Q_{k|j} \exp\left[\ln \left(\frac{\sum_j p_j^{r+1} Q_{k|j}}{\sum_j p_j^r Q_{k|j}}\right)^{\frac{1}{n}}\right] && \text{(B.73)}\\
&=\; 1 - \sum_j p_j^r \sum_k Q_{k|j} \left(\frac{\sum_j p_j^{r+1} Q_{k|j}}{\sum_j p_j^r Q_{k|j}}\right)^{\frac{1}{n}} && \text{(B.74)}\\
&\geq\; 1 - \sum_j p_j^r \left[\sum_k Q_{k|j} \frac{\sum_j p_j^{r+1} Q_{k|j}}{\sum_j p_j^r Q_{k|j}}\right]^{\frac{1}{n}} && \text{(B.75)}\\
&\geq\; 1 - \left[\sum_j p_j^r \sum_k Q_{k|j} \frac{\sum_j p_j^{r+1} Q_{k|j}}{\sum_j p_j^r Q_{k|j}}\right]^{\frac{1}{n}} && \text{(B.76)}\\
&=\; 1 - (1)^{\frac{1}{n}} = 0, && \text{(B.77)}
\end{aligned}$$

where the inequality in (B.73) follows from Jensen's inequality, and where (B.75) and (B.76) require the following lemma (which can be proved via Jensen's inequality).

Lemma B.1 ([12], pp. 523) Let p_i and a_i be non-negative numbers defined for $i = 1, 2, \ldots, N$, and let $\sum_i p_i = 1$. Then

$$\sum_i p_i a_i^\alpha \leq \left(\sum_i p_i a_i\right)^\alpha, \tag{B.78}$$

for $\alpha < 1$ with equality if and only if the a_i such that $p_i > 0$ are constant.

Thus $V(p^r)$ is increasing in r; furthermore, it is strictly increasing unless

$$c_j^{r+1} = c_i^r, \quad \forall\ i,\ j \text{ such that } p_i \neq 0 \neq p_j, \tag{B.79}$$

which reduces to the first condition of Theorem B.4.

Since $V(p^r)$ is increasing and bounded by $(C_n(\beta) + s\beta)$, and since $W(p^r) = V(p^{r+1}) - V(p^r) \to 0$, then $V(p^r)$ converges to some number $V(p^\infty) \leq C_n(\beta) + s\beta$. By the Bolzano-Weierstrass Theorem, a subsequence of the probability vectors $\{p^r\}$ must also converge to some p^* which is also the limit point of the sequence.

Now suppose that p^* does not achieve capacity. Then by the sufficiency of the Kuhn-Tucker conditions,

$$\frac{c_j^*}{\sum_j p_j^* c_j^*} > 1 \tag{B.80}$$

for some j, where $c_j^* = c_j(p^*)$.

Since a subsequence $\{p^{r_k}\}$ converges to p^*, continuity requires that $\{c_j^{r_k}\}$ converges to c_j^* for all j. But,

$$p_j^r = p_j^0 \prod_{n=0}^{r} b_j^n \tag{B.81}$$

where

$$b_j^n = \frac{c_j^n}{\sum_j p_j^n c_j^n} \tag{B.82}$$

and $\{b_j^n\}$ has a subsequence converging to a number greater than 1. But this means that the sequence of partial products does not converge which contradicts our earlier result that p_j^r converges.

Therefore, p^* achieves the n^{th} capacity-cost function and $V(p^\infty) = C_n(\beta) + s\beta$. This concludes the justification for the algorithm. □

What remains is to prove that terminating conditions exist for the algorithm. Conditions do exist that allow us to approximate a point $(\beta, C_n(\beta))$ as closely as we desire. The algorithm returns values measured in natural logarithms that can easily be scaled into bits.

Theorem 3.2 Let the left derivative of a point on $C_n(\beta)$ be specified by parameter s. Assuming p is any probability vector, we let

$$c_j = \exp\left(\frac{1}{n}\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - se_j\right). \tag{B.83}$$

Then, for cost $\beta = \sum_j p_j e_j$,

a)

$$C_n(\beta) \geq C_n^L(\beta) \triangleq s\beta + \sum_j p_j \ln c_j, \qquad \text{in nats,} \tag{B.84}$$

b)

$$C_n(\beta) \leq C_n^U(\beta) \triangleq s\beta + \ln \max_j c_j, \qquad \text{in nats.} \tag{B.85}$$

Proof of Theorem 3.2

a) From the theorem statement, p is a probability vector yielding per letter cost β. Therefore,

$$\begin{aligned} C_n(\beta) &\geq \frac{1}{n} I(p;Q) = \frac{1}{n}\sum_j p_j \sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} && \text{(B.86)}\\ &= \sum_j p_j (\ln c_j + se_j) && \text{(B.87)}\\ &= s\beta + \sum_j p_j \ln c_j. && \text{(B.88)} \end{aligned}$$

b) If p^* achieves capacity parametrized by s, then by Corollary B.8

$$C_n(\beta) \leq s\beta + \frac{1}{n}\ln \sum_j p_j^* \exp\left(\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} - nse_j\right). \tag{B.89}$$

Subtracting $s\beta + \ln \max_j c_j$ from both sides gives

$$\begin{aligned} & C_n(\beta) - \left(s\beta + \ln \max_j c_j\right) && \text{(B.90)} \\ \leq\ & \frac{1}{n} \ln \sum_j p_j^* \\ & \cdot \exp\left(\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j^* Q_{k|j}} - nse_j - n \ln \max_j c_j\right) && \text{(B.91)} \\ \leq\ & \frac{1}{n} \ln \sum_j p_j^* \exp\left(\sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j^* Q_{k|j}} - nse_j - n \ln c_j\right). && \text{(B.92)} \end{aligned}$$

We now use

$$n \ln c_j + nse_j = \sum_k Q_{k|j} \ln \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} \tag{B.93}$$

so that

$$\begin{aligned} & C_n(\beta) - \left(s\beta + \ln \max_j c_j\right) \\ \leq\ & \frac{1}{n} \ln \sum_j p_j^* \exp\left(\sum_k Q_{k|j} \ln \frac{\sum_j p_j Q_{k|j}}{\sum_j p_j^* Q_{k|j}}\right) && \text{(B.94)} \\ \leq\ & \frac{1}{n} \ln \sum_j p_j^* \sum_k Q_{k|j} \exp\left(\ln \frac{\sum_j p_j Q_{k|j}}{\sum_j p_j^* Q_{k|j}}\right) && \text{(B.95)} \\ =\ & \ln \sum_{j,k} p_j Q_{k|j} = 0, && \text{(B.96)} \end{aligned}$$

where the final inequality follows from Jensen's inequality.

□

This concludes the derivation of Blahut's algorithm for the computation of the n^{th} capacity-cost function $C_n(\beta)$ in nats for discrete channels.

Bibliography

[1] F. Alajaji, "Feedback Does Not Increase the Capacity of Discrete Channels with Additive Noise," *IEEE Transactions on Information Theory*, Vol. 41, pp. 546-549, March 1995.

[2] F. Alajaji, "New Results on the Analysis of Discrete Communication Channels with Memory," Ph.D. Dissertation, Department of Electrical Engineering, University of Maryland, College Park, MD 20742, USA, August 1994.

[3] S. Arimoto, "Computation of Random Coding Exponent Functions," *IEEE Transactions on Information Theory*, Vol. IT-22, No. 6, November 1976.

[4] S. Arimoto, "An Algorithm for Calculating the Capacity of an Arbitrary Discrete Memoryless Channel." *IEEE Transactions on Information Theory*, Vol. 18, pp. 14-20, 1972.

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

[6] P. Billingsley, *Ergodic Theory and Information*, John Wiley and Sons, Inc., 1965.

[7] R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley Publishing Co., 1987.

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

[9] P-N. Chen and F. Alajaji, "Strong Converse, Feedback Channel Capacity, and Hypothesis Testing," *Journal of the Chinese Institute of Engineers*, Vol. 18, No. 6, pp. 777-785, 1995.

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

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

[12] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, Inc., 1968.

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

[14] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes, Second Edition*, Clarendon Press, Oxford, 1992.

[15] S. Ihara, "Capacity of Discrete Time Gaussian Feedback Channel with and without Feedback, I," *Memoirs of the Faculty of Science Kochi University, Series A Mathematics*, Vol. 9, March, 1988.

[16] S. Ihara, "Capacity of Discrete Time Gaussian Feedback Channel with and without Feedback, II," *Japan Journal of Applied Mathematics*, Vol. 6, No. 2, pp. 245-258, June, 1989.

[17] Serge Lang, *Linear Algebra, Third Edition*, Undergraduate Texts in Mathematics, Springer-Verlag, New York Inc., 1987.

[18] R. J. McEliece, *The Theory of Information and Coding: A Mathematical Framework for Communication*, Encyclopaedia of Mathematics and its Applications, Vol. 3, Addison-Wesley, 1977.

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

[20] 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.

[21] C. E. Shannon, "A Mathematical Theory of Communication," Bell Sys. Tech. Journal, Vol. 27, pp. 379-423, 623-656, 1948.

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

[23] C. E. Shannon, "Channels with Side Information at the Transmitter," *IBM Journal of Research and Development*, Vol. 2, No. 4, pp. 289-293, October 1958.

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

[25] S. Verdú and T. S. Han, "A General Formula for Channel Capacity," *IEEE Transactions on Information Theory*, Vol. 40, pp. 1147-1157, July 1994.

[26] 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.

[27] A. Wyner and J. Ziv, "A Theorem on the Entropy of Certain Binary Sequences and Application (Part 1)," *IEEE Transactions on Information Theory*, Vol. IT-19, pp. 769-777, Nov. 1973.

Vita

Nicholas J. Whalen

Education:

1996 to 1998	M.Sc. (Eng.)	Queen's University	Mathematics and Engineering
1991 to 1996	B.Sc. (Honours)	Queen's University	Mathematics and Engineering

Experience:

1996 to 1998	Research Assistant, Queen's University
1996 to 1998	Teaching Assistant, Queen's University
1997 to 1998	Volunteer, Queen's International Centre English Tutor
1995 to 1998	Summer Lab Assistant, Philips Semiconductor, Gratkorn, Austria
1994 to 1996	Student Member, Alma Mater Society (AMS) Board of Directors
1995 to 1996	Senior Engineering Representative, AMS Assembly