RENYI'S ENTROPY FOR DISCRETE MARKOV SOURCES

by

Ziad Rached

A project submitted to the

Department of Mathematics and Statistics

in conformity with the requirements

for the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

September 1998

Acknowledgments

I would like to thank my supervisor, Dr. Fady Alajaji for his support and his helpful suggestions and comments during the completion of this project.

Also, I would like to thank Dr. L. Campbell who luminously pointed out the importance of Perron-Frobenius theory for this project.

Finally, I would like to thank NSERC, the Department of Mathematics and Statistics, and the School of Graduate Studies for their financial support.

Abstract

In this work, a Rényi variable length source coding theorem for memoryless sources [5] for which Shannon's source coding theorem is a particular case, is studied in detail. A natural question to ask is whether this theorem can be extended to more general sources. This question is addressed by solving the formula for the Rényi entropy rate of ergodic Markov sources of arbitrary order. This leads to an extension of the Rényi source coding theorem for ergodic Markov sources. The main tool used to obtain the Rényi entropy rate result is Perron-Frobenius theory.

Contents

Chapter 1

Introduction

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

1.1 Literature review

A detailed analysis of Campbell's variable length source coding theorem for memoryless sources associated with Rényi's entropy is given [5,6]. We also examine several topics from matrix algebra specifically Perron-Frobenius theory [11],[13],[12].

1.2 Contributions

The contributions of this project are as follows:

- A formula for the Rényi entropy rate of ergodic Markov sources of first order.
- A Rényi's variable length source coding theorem for 1^{st} order ergodic Markov sources.
- The extension of these results for ergodic Markov sources of arbitrary order.

1.3 Thesis overview

This project consist mainly of two major parts.

The first part given in Chapter 2 is a detailed analysis of [5] which is a generalization of the source coding theorem to the case of Rényi's entropy where an exponential length function is taken into consideration rather then the usual expected mean length. The main tool in accomplishing this result is Hölder's inequality.

The second part consists of Chapters 3 and 4. Primarily, a general review of Markov chains, entropy rate, determinants, and Perron-Frobenius theory is first provided in Chapter 3 . Then, we calculate the Rényi entropy rate for 1^{st} order ergodic Markov sources when the probability transition matrix is positive. Also we look into the case when the probability transition matrix is non-negative. The last section

illustrates the theory with some examples.

Chapter 4 is an extension of the results of Chapter 3 for ergodic Markov chains of order k. In this case, the probability transition matrix is non-negative. Finally, some numerical examples are given.

Chapter 2

A Rényi source coding theorem for memoryless sources

2.1 Preliminaries [8]

We will first introduce the concept of entropy, which is a measure of uncertainty of a random variable. Let X be a discrete random variable with finite alphabet $\mathcal{X}$ and probability mass function $p(x) \stackrel{\triangle}{=} Pr\{X = x\}$, $\forall x \in \mathcal{X}$.

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

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

The log is usually in base 2 and entropy is expressed in bits. If the base of the logarithm is e, then the entropy is measured in nats. If the log is in base D, then the

entropy is denoted by $H_D(X)$.

Definition: The *joint entropy* $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

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

Definition: The *conditional entropy* $H(Y|X)$ is defined as

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x).$$

Definition: The *joint entropy* $H(X_1,X_2,\ldots,X_n)$ of a sequence of random variables $X_1,X_2,\ldots,X_n$ with a joint distribution $p(x_1,x_2,\ldots,x_n)$ is defined as

$$H(X_1,X_2,\ldots,X_n) = -\sum_{x_1,x_2,\ldots,x_n} p(x_1,x_2,\ldots,x_n)\log p(x_1,x_2,\ldots,x_n)$$

Now, we will introduce some definitions and theorems about source coding.

Definition: A variable length source code C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of finite length strings of symbols from a D-ary alphabet. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$.

Definition: The *expected length* $L(C)$ of a source code $C(x)$ for a random variable X with probability mass function $p(x)$ is given by

$$L(C) = \sum_{x\in\mathcal{X}} p(x)l(x),$$

where $l(x)$ is the length of the codeword associated with x.

Without loss of generality, we can assume that the D-ary alphabet is $\mathcal{D} = \{0, 1, \ldots, D-1\}$.

Definition: A variable length code is said to be *non-singular* if every element of $\mathcal{X}$ maps into a different string in $\mathcal{D}^*$, i.e.,

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j).$$

Definition: The *extension* C^* of a code C is the mapping from finite length strings of $\mathcal{X}$ to finite length strings of $\mathcal{D}$, defined by

$$C(x_1 x_2 \cdots x_n) = C(x_1)C(x_2)\cdots C(x_n),$$

where $C(x_1)C(x_2)\cdots C(x_n)$ indicates concatenation of the corresponding codewords.

Definition: A code is called *uniquely decodable* if its extension is non-singular.

Theorem 2.1.1 *(Kraft inequality) The codeword lengths $l_1, l_2, \ldots, l_m$ of any uniquely decodable code must satisfy the Kraft inequality*

$$\sum_i D^{-l_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, it is possible to construct a uniquely decodable code with these codeword lengths.

Theorem 2.1.2 *The expected length L of any uniquely decodable D-ary code for a random variable X is greater than or equal to the entropy $H_D(X)$, i.e.,*

$$L \geq H_D(X)$$

with equality iff $D^{-l_i} = p_i$ *for each* i.

Theorem 2.1.3 *If* L^* *is the minimum expected length, then*

$$H_D(X) \leq L^* < H_D(X) + 1.$$

The following inequalities are useful in order to prove the lemma in the next section and some lemmas in Section 2.3.

Definition: A function $f(x)$ is said to be *convex* over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

Definition: A function f is *concave* if $-f$ is convex.

Theorem 2.1.4 *If the function* f *has a second derivative which is non-negative everywhere, then the function is convex.*

Theorem 2.1.5 *(Jensen's inequality): If* f *is a convex function and* X *is a random variable , then*

$$E[f(X)] \geq f(E[X]),$$

where E *denote expectation.*

Theorem 2.1.6 *(Log sum inequality): For non-negative numbers,* $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $\frac{a_i}{b_i} = constant$.

Theorem 2.1.7 $H(X) \leq \log |\mathcal{X}|$, *where* $|\mathcal{X}|$ *denotes the number of elements in the range of* $\mathcal{X}$, *with equality iff* X *has a uniform distribution over* X.

Definition: A discrete memoryless source (DMS) is a source for which the symbols are independently generated and identically distributed.

2.2 A measure of length [5]

Consider a DMS with alphabet $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ and distribution $p = (p_1, p_2, \ldots, p_N)$ where we assume that the probability of x_i is $p_i > 0 \ \forall i$ from now until the end of this chapter. Suppose that we wish to represent the letters in $\mathcal{X}$ by finite sequences of symbols from the set $\{0, 1, \ldots, D-1\}$ where $D > 1$. It is known that there exists a uniquely decodable code which represents each x_i by a sequence of l_i D-ary symbols iff the lengths l_i satisfy the Kraft inequality

$$\sum_{i=1}^{N} D^{-l_i} \leq 1. \tag{2.1}$$

An interesting problem is to minimize the average code length subject to (2.1). This is a good procedure if the cost of using a sequence of length l_i is directly proportional to l_i. But, this is not always the case. In some occasions, the cost can be a non-linear function. For example, an exponential cost occurs frequently in many interesting

applications. This could be the case for example if the cost of encoding and decoding equipment were an important factor, or, if buffer overflow caused by long codewords is important. Therefore, for these kinds of applications, a better procedure is to minimize the quantity

$$C = \sum_{i=1}^{N} p_i D^{tl_i},$$

where $t \neq 0$ is some parameter related to the cost.

For arbitrary cost functions refer to [6].

Definition: A code length of order t is defined by

$$L(t) = \frac{1}{t} \log_D (\sum_{i=1}^{N} p_i D^{tl_i}) \quad (0 < t < \infty). \tag{2.2}$$

Since $L(t)$ is clearly a monotonic function of C, minimizing C is equivalent to minimizing $L(t)$.

The code length of order t has several properties.

By l'Hospital's rule

$$L(0) \stackrel{\triangle}{=} \lim_{t \to 0} L(t) = \sum_{i=1}^{N} l_i p_i, \tag{2.3}$$

which is the expected length of the source X.

When t is large the sum $\sum_{i=1}^{N} p_i D^{tl_i}$ is dominated by the term $p_j D^{tl_j}$, where l_j is the largest of the numbers $l_1, l_2, \ldots, l_N$. Therefore

$$L(\infty) \stackrel{\triangle}{=} \lim_{t \to \infty} L(t) = max_{1 \leq i \leq N} \quad l_i. \tag{2.4}$$

Lemma 2.2.1 *([4], p. 16) $L(t)$ is monotonic nondecreasing function of t.*

Proof: Since the logarithm is a monotonic nondecreasing function, it is equivalent to prove that the function $f(t) = (\sum_{i=1}^{N} p_i D^{tl_i})^{\frac{1}{t}}$. Let $f'(t)$ denotes the derivative of $f(t)$ with respect to t. Then

$$f'(t) = \frac{f(t)}{t^2 \sum_i p_i D^{tl_i}} (\sum_i p_i D^{tl_i} \ln D^{tl_i} - \sum_i p_i D^{tl_i} \ln(\sum_i p_i D^{tl_i})).$$

By Theorem 2.1.6 (Log sum inequality), if $a_i = p_i D^{tl_i}$ and $b_i = p_i$ then

$$\sum_i p_i D^{tl_i} \ln D^{tl_i} - \sum_i p_i D^{tl_i} \ln(\sum_i p_i D^{tl_i}) \geq 0.$$

Since $f(t) > 0$, then $f'(t) \geq 0$. Hence, $f(t)$ is monotonic nondecreasing function of t which yields that L(t) is also monotonic nondecreasing function of t. □

Note that when the maximum length is an important factor, $L(\infty)$ is a good measure of the cost. $L(0)$ is used when the cost is linear. Intermediate values of t provide a measure of length which lies between these limits.

Note also that when $l_i = l$ for all $i = 1, 2, \ldots, N$, then $L(t) = l$. This is a reasonable property for any measure of length to possess.

2.3 Rényi's entropy

In this section we introduce Rényi's entropy and examine its properties [3],[5],[7],[9].

Definition: Rényi's entropy of order α for a random variable X with distribution $(p_1, \ldots, p_N)$ is defined by

$$H_\alpha = \frac{1}{1-\alpha} \log_D(\sum_{i=1}^{N} p_i^\alpha), \tag{2.5}$$

where $\alpha \geq 0$ and $\alpha \neq 1$.

Rényi's entropy has several important properties. Some of these properties are clear, so we just declare them without proof.

L'Hospital's rule shows that

$$H_1 \stackrel{\triangle}{=} \lim_{\alpha \to 1} H_\alpha = -\sum_{i=1}^{N} p_i \log_D p_i. \tag{2.6}$$

Thus H_1 is the ordinary Shannon entropy. The entropy of order α behaves in much the same way as H_1. For example, H_α is a continuous and symmetric function of $p_1, \ldots, p_N$. If $p_i = N^{-1}$ for each i, $H_\alpha = \log_D N$.

Lemma 2.3.1 *If $X_1, X_2, \ldots, X_M$ is a sequence of independent and identically random variables with alphabet $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, then*

$$H_\alpha(X_1, X_2, \ldots, X_M) \stackrel{\triangle}{=} H_\alpha(M) = MH_\alpha. \tag{2.7}$$

Proof: Consider a typical sequence of length M, say $s = (i_1, i_2, \ldots, i_M)$. The probability of s is

$$P(s) = p_{i_1} p_{i_2} \cdots p_{i_M}. \tag{2.8}$$

Then

$$H_\alpha(M) = \frac{1}{1-\alpha} \log_D Q,$$

where

$$Q = \sum_{s \in \mathcal{X}^M} P(s)^\alpha.$$

It follows directly from (2.8) that

$$Q = (\sum_{i=1}^{N} p_i^{\alpha})^M$$

and hence that

$$H_\alpha(M) = MH_\alpha. \tag{2.9}$$

□

Lemma 2.3.2 *The Rényi's entropy of order α is a decreasing function of α.*

Proof: The derivative of H_α with respect to α is given by

$$H'_\alpha = \frac{(1-\alpha)\sum_i p_i^\alpha \log p_i + (\sum_i p_i^\alpha)\log(\sum_i p_i^\alpha)}{\sum_i p_i^\alpha (1-\alpha)^2}.$$

The denominator is clearly positive.

Using Theorem 2.1.4, it can be verified that the function $f(x) = x\log x$ is a convex function $\forall x > 0$. If we denote by $E[X]$ the expected value of the random variable X then

$$\sum_i p_i p_i^{\alpha-1} \log p_i^{\alpha-1} = E[p_i^{\alpha-1} \log p_i^{\alpha-1}].$$

By Jensen's inequality (Theorem 2.1.5) we obtain that

$$E[p_i^{\alpha-1} \log p_i^{\alpha-1}] \geq E[p_i^{\alpha-1}] \log E[p_i^{\alpha-1}] = \sum_i p_i^\alpha \log \sum_i p_i^\alpha.$$

Therefore

$$\sum_i p_i^\alpha \log p_i^{\alpha-1} \geq \sum_i p_i^\alpha \log \sum_i p_i^\alpha,$$

and thus

$$(1-\alpha)\sum_i p_i^\alpha \log p_i + \sum_i p_i^\alpha \log \sum_i p_i^\alpha \leq 0.$$

We conclude that $H'_\alpha \leq 0$, hence, H_α is a decreasing function of α.

Note that $H'_\alpha = 0$ iff $p_i = N^{-1}\ \forall i$ by direct calculation. Hence H'_α is strictly decreasing unless p is the uniform distribution. □

Lemma 2.3.3 *If $\alpha \to \infty$, then, $H_\infty \stackrel{\triangle}{=} \lim_{\alpha\to\infty} H_\alpha = -\log_D \bar{p}$, where $\bar{p} = max(p_1, \ldots, p_N)$.*

Proof: Since $0 < p_i < 1$, as $\alpha \to \infty$, the sum $\sum_i p_i^\alpha$ is clearly dominated by $\bar{p}^\alpha$. Therefore

$$H_\infty = \lim_{\alpha\to\infty} \frac{1}{1-\alpha} \log_D \bar{p}^\alpha = -log_D \bar{p}.$$

□

Lemma 2.3.4 *The Rényi entropy H_α is non-negative.*

Proof: If $0 < \alpha < 1$, then $p_i^\alpha \geq p_i\ \forall i$. Hence, $\sum_i p_i^\alpha \geq \sum_i p_i = 1$. Therefore, $\log(\sum_i p_i^\alpha) \geq 0$. Since $1-\alpha > 0$, we get that $H_\alpha \geq 0$.

If $\alpha > 1$, then $p_i^\alpha \leq p_i\ \forall i$. Hence, $\sum_i p_i^\alpha \leq \sum_i p_i = 1$. Therefore, $\log(\sum_i p_i^\alpha) \leq 0$. Since $1-\alpha < 0$, we get that $H_\alpha \geq 0$.

Note that $H_\alpha = 0$ iff the distribution is a point mass. □

Lemma 2.3.5 *$H_\alpha \leq \log|\mathcal{X}|$ with equality iff $(p_1, \ldots, p_N)$ is uniform.*

Proof: Consider first the case $\alpha > 1$. By Lemma 2.3.2, $H_\alpha \leq H_1$. But $H_1 = H$ by (2.6). Also by Theorem 2.1.7, $H \leq \log |\mathcal{X}|$. Therefore, $H_\alpha \leq \log |\mathcal{X}|$.

If $0 < \alpha < 1$ we need the following observation.

Let p_i, q_i be non-negative numbers defined over a finite set of i with $\sum_i q_i = \sum_i p_i = 1$. Then

$$\sum_i p_i^\alpha q_i^{1-\alpha} \leq 1.$$

The function $f(x) = x^\alpha$ is concave by Theorem 2.1.4 since its second derivative is negative.

Note that

$$\sum_i p_i^\alpha q_i^{1-\alpha} = \sum_i \left(\frac{p_i}{q_i}\right)^\alpha q_i = E\left[\left(\frac{p_i}{q_i}\right)^\alpha\right],$$

where E denote the expectation with respect to the probability distribution q_i.

Applying Jensen's inequality to the function X^α where X is a random variable taking on the values p_i/q_i, we get

$$E\left[\left(\frac{p_i}{q_i}\right)^\alpha\right] \leq \left(E\left[\frac{p_i}{q_i}\right]\right)^\alpha = \left(\sum_i \frac{p_i}{q_i} q_i\right)^\alpha = (\sum_i p_i)^\alpha = 1.$$

Therefore

$$\sum_i p_i^\alpha q_i^{1-\alpha} \leq 1,$$

and the observation is proved.

Let $q_i = 1/|\mathcal{X}|$. This substitution is valid since

$$\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} = 1.$$

Therefore

$$\sum_i p_i^\alpha \left(\frac{1}{|\mathcal{X}|}\right)^{1-\alpha} \leq 1,$$

or equivalently

$$(\sum_i p_i^\alpha) \leq |\mathcal{X}|^{1-\alpha}.$$

Taking the logarithms of both sides of the last inequality, and then dividing by $1-\alpha$ yield the desired result.

Note that by direct calculation $H_\alpha = \log|\mathcal{X}|$ iff p is uniform on $\mathcal{X}$. □

2.4 A source coding theorem [5]

Lemma 2.4.1 *Let $l_1, l_2, \ldots, l_N$ satisfy Kraft's inequality. Then*

$$L(t) \geq H_\alpha, \tag{2.10}$$

where $\alpha = 1/(t+1)$.

Proof: If $t = 0$, the result is given in Theorem 2.1.2.

If $t = \infty$, we have $L(\infty) = max(l_i)$ by (2.4). Also, by simple calculation, $H_0 = \log_D N$. If the $l_i's$ satisfy Kraft's inequality we must have

$$D^{-l_i} \leq N^{-1}$$

for at least one value of i and hence for the maximum l_i. Otherwise, if $D^{-l_i} > N^{-1}$ for all i, then $\sum_i D^{-l_i} > \sum_i N^{-1} = 1$. This yields that $\sum_i D^{-l_i} > 1$ which contradicts

Kraft's inequality. Taking the log on both sides of the inequality $D^{-max(l_i)} \leq N^{-1}$ yields $max(l_i) \geq \log_D N$, and hence $L(\infty) \geq H_0$.

It remains to prove the lemma for $0 < t < \infty$. By Hölder's inequality ([4] p. 19),

$$(\sum_{i=1}^{N} x_i^p)^{1/p} (\sum_{i=1}^{N} y_i^q)^{1/q} \leq \sum_{i=1}^{N} x_i y_i \tag{2.11}$$

where $p^{-1} + q^{-1} = 1$ and $p < 1$. In (2.11), let $p = -t$, $q = 1 - \alpha$, $x_i = p_i^{-1/t} D^{-l_i}$, and $y_i = p_i^{1/t}$. Substituting p and q by their values in the equation $p^{-1} + q^{-1} = 1$, yields $\alpha = (t+1)^{-1}$. With these substitutions (2.11) becomes

$$(\sum_i p_i D^{tl_i})^{-1/t} (\sum_i p_i^{\alpha})^{1/(1-\alpha)} \leq \sum_i D^{-l_i}.$$

Therefore

$$(\sum_i p_i D^{tl_i})^{1/t} \geq \frac{(\sum_i p_i^{\alpha})^{1/(1-\alpha)}}{\sum_i D^{-l_i}} \geq (\sum_i p_i^{\alpha})^{1/(1-\alpha)}, \tag{2.12}$$

where the last inequality follows from the assumption that Kraft's inequality is satisfied. Taking logarithms of the first and last member of (2.12) proves the statement of the lemma. □

Lemma 2.4.2 *Under the same assumptions of the previous lemma, there exists some l_i's, $i = 1, \ldots, N$ such that*

$$H_\alpha \leq L(t) < H_\alpha + 1. \tag{2.13}$$

Proof: We observe first that we have an equality in (2.10) and (2.12) if and only if we have an equality in (2.1) and (2.11). By ([4] p. 19), equality in Hölder's inequality

occurs when $x_i^p = ay_i^q$, $i = 1, \ldots, N$, for some real number a. If we replace x_i and y_i by their values from the previous lemma we get

$$\begin{aligned} D^{-l_i} &= a^{\frac{1}{p}} p_i^{\frac{q}{tp}} p_i^{\frac{1}{t}} \\ &= a^{\frac{1}{p}} p_i^{\frac{q}{t}} \quad (1/p + 1/q = 1) \\ &= a^{\frac{1}{p}} p_i^{\frac{1-\alpha}{t}} \quad (\alpha = 1/(1+t)) \\ &= a^{\frac{1}{p}} p_i^{\alpha}. \end{aligned}$$

Equality in (2.1) occurs when $\sum_i D^{-l_i} = 1$. This yields

$$a^{\frac{1}{p}} = \frac{1}{\sum_i p_i^{\alpha}}.$$

Therefore, we conclude that

$$D^{-l_i} = \frac{p_i^{\alpha}}{\sum_j p_j^{\alpha}}.$$

Thus,

$$\log_D D^{-l_i} = \log_D p_i^{\alpha} - \log_D(\sum_j p_j^{\alpha}),$$

yielding

$$l_i = \lceil -\alpha \log_D p_i + \log_D(\sum_{j=1}^{N} p_j^{\alpha}) \rceil,$$

since l_i the length of the i^{th} codeword must be an integer. If we choose the l_i's to satisfy the above equality, letting

$$W = \sum_{j=1}^{N} p_j^{\alpha},$$

yields

$$-\alpha \log_D p_i + log_D W \leq l_i < 1 - \alpha \log_D p_i + \log_D W,$$

or equivalently

$$p_i^{-\alpha t} W^t \leq D^{tl_i} < D^t p_i^{-\alpha t} W^t.$$

Now, if we multiply each member by p_i, sum over all i, and use the fact that $\alpha t = 1-\alpha$, we get

$$W^{1+t} \leq \sum_i p_i D^{tl_i} < D^t W^{1+t}.$$

By taking logarithms, dividing by t, and using the relations $1+t = \alpha^{-1}$ and $\alpha t = 1-\alpha$, we get

$$H_\alpha \leq L(t) < H_\alpha + 1. \tag{2.14}$$

□

We can now prove a coding theorem for a DMS.

Theorem 2.4.1 *Let $\alpha = (1+t)^{-1}$. By encoding sufficiently long sequences of input symbols of a DMS it is possible to make the average code length of order t per input symbol as close to H_α as desired. Also, it is not possible to find a uniquely decodable code whose average length of order t is less than H_α.*

Proof: Let a sequence s of input symbols of length M be generated independently, where each symbol is governed by the probability distribution $(p_1, \ldots, p_N)$. We

can consider these sequences as supersymbols from the alphabet $\mathcal{X}^M$. Hence by Lemma 2.4.2

$$H_\alpha(M) \leq L_M(t) < H_\alpha(M) + 1. \tag{2.15}$$

Let $L_M(t)$ denote the length of order t for the M-sequences given by

$$L_M(t) = \frac{1}{t} \log_D \sum P(s) D^{tl(s)},$$

where the summation extends over the N^M sequences s. Let $l(s)$ denote the length of the codeword associated with the sequence s.

Now, by Lemma 2.3.1, if we divide (2.15) by M, we get

$$H_\alpha \leq \frac{L_M(t)}{M} < H_\alpha + \frac{1}{M}. \tag{2.16}$$

By (2.16), if we choose M sufficiently large the average length can be made as close to H_α as desired. □

Remark 1: Note that when $t = 0$, this theorem is just the extension of Theorem 2.1.3 to supersymbols from $\mathcal{X}^M$.

Remark 2: Note also that the theorem holds when $t = \infty$. By Kraft's inequality,

$$\sum_s D^{-l(s)} \leq 1,$$

where the summation extends over the N^M sequences s.

Clearly, $D^{-l(s)} \leq N^{-M}$ for at least one sequence s. Otherwise, if $D^{-l(s)} > N^{-M}$ for all s then $\sum_s D^{-l(s)} > \sum_s N^{-M} = 1$ which contradicts Kraft's inequality. Therefore

$$l(s) \geq M \log_D N,$$

and hence

$$max(l(s)) \geq M \log_D N.$$

Taking into consideration the integer restriction of $l(s)$ we must have,

$$M \log_D N \leq l(s) < M \log_D N + 1.$$

Since $H_0 = \log_D N$ and $L_M(\infty) = max(l(s))$, dividing by M we get

$$H_0 \leq \frac{L_M(\infty)}{M} < H_0 + \frac{1}{M}.$$

Thus the theorem follows as before.

Chapter 3

Rényi's entropy for 1^{st} order ergodic Markov sources

3.1 Markov chains [8]

A stochastic process is an indexed sequence of random variables. In general, there can be an arbitrary dependence among the random variables. The process is characterized by the joint probability mass functions $Pr\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\} \stackrel{\triangle}{=} p(x_1, x_2, \ldots, x_n), (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ for $n = 1, 2, \ldots$.

Definition: 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.,

$$Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\}$$

for every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Definition: A discrete stochastic process $X_1, X_2, \ldots$ is said to be a *Markov chain* or a *Markov process* if, for $n = 1, 2, \ldots$,

$$Pr(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = Pr(X_{n+1} = x_{n+1}|X_n = x_n)$$

for all $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}$.

In this case, the joint probability mass function of the random variables can be written as

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}).$$

Definition: The Markov chain is said to be *time invariant* or homogeneous if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n; i.e., for $n = 1, 2, \ldots$

$$Pr\{X_{n+1} = b|X_n = a\} = Pr\{X_2 = b|X_1 = a\}, \quad \text{for all} \quad a, b \in \mathcal{X}.$$

From now on all Markov chains are time invariant. If $\{X_i\}$ is a Markov chain, then X_n is called the *state* at time n. A time invariant Markov chain is characterized by its initial state and a *probability transition matrix* $P = [p_{ij}]$, $i, j \in \{1, 2, \ldots, m\}$, where $p_{ij} = Pr\{X_{n+1} = j|X_n = i\}$.

Definition: If it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps, then the Markov chain is said to be *irreducible.*

Definition: A distribution on the states such that the distribution at time $n+1$ is the same as the distribution at time n is called a *stationary distribution.*

The stationary distribution draws its name from the fact that if the initial state of a Markov chain is drawn according to the stationary distribution, then the Markov chain forms a stationary process.

Definition: The *period* of a state i is defined as the greatest common divisor of those values of n for which $p_{ii}^n > 0$ where p_{ij}^n denotes the ij^{th} element of the n^{th} power of the transition matrix P. If the period is 1, the state is said to be *aperiodic.* If the period is 2 or more, the state is said to be *periodic.* An irreducible Markov chain is *aperiodic* if the period of any of its states is 1.

Definition: An irreducible and aperiodic Markov chain is called *ergodic.*

Theorem 3.1.1 *If the finite state Markov chain is ergodic, then the stationary distribution is unique, and from any starting distribution, the distribution of X_n tends to the stationary distribution as $n \to \infty$.*

Theorem 3.1.2 *([11], page 108) If a finite state Markov chain is ergodic and has N states, then $p_{ij}^m > 0$ for all i, j, and all $m \geq N(N-1)$.*

3.2 Entropy rate [8]

Definition: The *entropy rate* of a stochastic process $\{X_i\}$ is defined by

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

when the limit exists.

Example: If $X_1, X_2, \ldots$ are *i.i.d. random variables*, i.e., independent and identical, then

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim_{n\to} \frac{nH(X_1)}{n} = H(X_1),$$

which is what one would expect for the entropy rate per symbol.

We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1),$$

when the limit exists.

The two quantities $H(\mathcal{X})$ and $H'(\mathcal{X})$ correspond to two different notions of entropy rate. The first is the per symbol entropy of the n random variables, and the second is the conditional entropy of the last random variable given the past. An important result is that for a stationary process both limits exists and are equal.

Theorem 3.2.1 *For a stationary stochastic process,* $H(\mathcal{X}) = H'(\mathcal{X})$.

Corollary 3.2.1 *For a stationary Markov chain, the entropy rate is given by*

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1),$$

where the conditional entropy is calculated using the stationary distribution.

Corollary 3.2.2 *Let $\{X_i\}$ be a stationary Markov chain with stationary distribution q and transition matrix P. Then the entropy rate is*

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_{ij} q_i p_{ij} \log p_{ij}.$$

Remark: If the Markov chain is ergodic, then it has a unique stationary distribution on the states, and any initial distribution tends to the stationary distribution as $n \to \infty$. In this case, even though the initial distribution is not the stationary distribution, the entropy rate, which is defined in terms of long term behavior, is $H(\mathcal{X})$ as remarked in the two previous corollaries.

3.3 Perron–Frobenius theory [11]

Definition: A real vector x is defined to be *positive*, denoted $x > 0$ if $x_i > 0$ for each component i.

Definition: A real matrix P is *positive*, denoted $P > 0$, if $p_{ij} > 0$ for each i, j.

Definition: x is *non-negative*, denoted $x \geq 0$, if $x_i \geq 0$ for all i.

Definition: P is *non-negative*, denoted $P \geq 0$, if $p_{ij} \geq 0$ for all i, j.

Remark: Note that it is possible to have $x \geq 0$ and $x \neq 0$ without having $x > 0$, since $x > 0$ means that all components of x are positive and $x \geq 0$, $x \neq 0$ means that

at least one component of x is positive and all are non-negative.

Definition: The row vector a is a *left eigenvector* of P of eigenvalue λ if $a \neq 0$ and $aP = \lambda a$.

Definition: The column vector b is a *right eigenvector* of eigenvalue λ if $b \neq 0$ and $Pb = \lambda b$.

Theorem 3.3.1 (Perron) *Let $P > 0$ be a square matrix. Then P has a positive eigenvalue λ that exceeds the magnitude of each other eigenvalue. There is a positive right eigenvector, $b > 0$, corresponding to λ, and the following properties hold for λ and b:*

1. *If $\lambda x \leq Px$ for $x \geq 0$, then $\lambda x = Px$.*

2. *If $\lambda x = Px$, then $x = \alpha b$ for some scalar α.*

Definition: Let P be an $N \times N$ non-negative square matrix. A directed graph is associated with P by drawing a directed edge that goes from i to j if $p_{ij} > 0$, $i, j = 1, 2, \ldots, N$. P is *irreducible* if for every pair of nodes i, j in this graph, there is a walk from i to j.

Denote a typical element of P^m by p_{ij}^m. If P is irreducible, a walk exists from any i to any $j \neq i$ with length at most $N - 1$, since the walk needs to go through at most each of the other nodes. Thus $p_{ij}^m > 0$ for some m, $1 \leq m \leq N - 1$, and $\sum_{m=1}^{N-1} p_{ij}^m > 0$. The key to analyzing irreducible matrices is the fact that the matrix

$\sum_{m=0}^{N-1} P^m$ is positive. The $m = 0$ term, P^0 is just the identity matrix, which covers the case $i = j$.

Theorem 3.3.2 (Frobenius) *Let $P \geq 0$ be an irreducible square matrix. Then P has a positive eigenvalue λ that is greater than or equal to the magnitude of each other eigenvalue. There is a positive right eigenvector, $b > 0$ corresponding to λ, and the following properties hold for λ and b:*

1. *For any non-zero $x \geq 0$, if $\lambda x \leq Px$, then $\lambda x = Px$.*

2. *If $\lambda x = Px$, then $x = \alpha b$ for some scalar α.*

Corollary 3.3.1 *The largest real eigenvalue λ of an irreducible matrix $P \geq 0$ has a positive left eigenvector a. a is unique (within a scale factor) and is the only non-negative non-zero vector (within a scale factor) that satisfies $\lambda a \leq aP$.*

Corollary 3.3.2 *Let λ be the largest real eigenvalue of an irreducible matrix and let the right and left eigenvectors of λ be $b > 0$ and $a > 0$. Then, within a scale factor, b is the only non-negative right eigenvector of P (i.e., no other eigenvalues have non-negative eigenvectors). Similarly, a is the only non-negative left eigenvector of P.*

Corollary 3.3.3 *Let P be the transition matrix of an irreducible Markov chain. Then $\lambda = 1$ is the largest real eigenvalue of P, $e = (1, 1, \ldots, 1)^T$ is the right eigenvector of*

$\lambda = 1$ unique within a scale factor, and there is a unique probability vector $a > 0$ that is a left eigenvector of $\lambda = 1$.

Corollary 3.3.4 *The largest real eigenvalue λ of an irreducible matrix $P \geq 0$ is a strictly increasing function of each component of P.*

Corollary 3.3.5 *Let λ be the largest eigenvalue of $P > 0$ and let $a(b)$ be the positive left (right) eigenvector of λ normalized so that $ab = 1$. Then*

$$\lim_{n\to\infty} \frac{P^n}{\lambda^n} = ba$$

Theorem 3.3.3 *Let P be the transition matrix of an ergodic finite state Markov chain. Then $\lambda = 1$ is the largest real eigenvalue of P, and $\lambda > |\lambda'|$ for every other eigenvalue λ'. Furthermore, $lim_{m\to\infty} P^m = ea$, where $a > 0$ is the unique probability vector satisfying $aP = a$ and $e = (1, 1, \ldots, 1)^T$ is the unique b (within a scale factor) satisfying $Pe = e$.*

3.4 Some determinant properties [12]

Some important properties about determinants are useful for the last section. Let A be an $n \times n$ square matrix. We start by defining a *diagonal* of A.

Definition: A *diagonal* of A is a sequence of n elements of the matrix containing one and only one element from each row of A and one and only one element from each

column of A. A diagonal of A is always assumed to be ordered according to the row indices; therefore it can be written in the form

$$a_{1j_1}, a_{2j_2}, \ldots, a_{nj_n},$$

where $(j_1, j_2, \ldots, j_n)$ is a permutation of the numbers $1, 2, \ldots, n$. In particular, if $(j_1, j_2, \ldots, j_n) = (1, 2, \ldots, n)$, we obtain the main diagonal of A. Clearly, A has exactly $n!$ distinct diagonals.

Definition: We say that a pair of numbers j_k and j_p in the permutation $(j_1, j_2, \ldots, j_n)$ form an *inversion* if $j_k > j_p$ while $k < p$, that is, if a larger number in the permutation precedes a smaller one. Each permutation $j = (j_1, j_2, \ldots, j_n)$ has a certain number of inversions associated with it, denoted briefly by $t(j)$.

Definition: The permutation is called *odd* or *even* according to whether the number $t(j)$ is odd or even. This property is known as the *parity* of the permutation.

Definition: The *determinant* of A, denoted det A or $|A|$, is defined by

$$|A| = \sum_j (-1)^{t(j)} a_{1j_1} a_{2j_2} \cdots a_{nj_n}, \qquad (*)$$

where j varies over all $n!$ permutations of $1, 2, \ldots, n$.

In other words, $|A|$ is a sum of $n!$ products. Each product involves n elements of A belonging to the same diagonal. The product is multiplied by $+1$ or -1 according to whether the permutation $(j_1, j_2, \ldots, j_n)$ that defines the diagonal is even or odd, respectively.

Lemma 3.4.1 *If B denotes a matrix obtained from A by multiplying one of its rows (or columns) by a scalar k, then $|B| = k|A|$.*

Lemma 3.4.2 *If the matrix B is obtained by interchanging two rows (or columns) of A, then $|B| = -|A|$.*

Lemma 3.4.3 *Let B be the matrix obtained from A by adding the elements of its i^{th} row (or column) to the corresponding elements of its j^{th} row (or column) multiplied by a scalar α $(j \neq i)$. Then $|B| = |A|$.*

Lemma 3.4.4 *Suppose that the entries of A are functions of some parameter α. Let $|A|_i$ be the determinant obtained from $|A|$ by replacing the elements in the i^{th} row by their derivatives with respect to α and leaving the other rows unchanged. Then*

$$|A|' = \sum_{i=1}^{n} |A|_i.$$

Proof: If we differentiate $(*)$, then by the sum rule of derivatives

$$|A|' = \sum_{i} (-1)^{t(j)} (a_{1j_1} a_{2j_2} \ldots a_{nj_n})',$$

where j varies over all $n!$ permutations of $1, 2, \ldots, n$. By the product rule of derivatives

$$(a_{1j_1} a_{2j_2} \ldots a_{nj_n})' = a'_{1j_1} a_{2j_2} \ldots a_{nj_n} + a_{1j_1} a'_{2j_2} \ldots a_{nj_n} + \cdots + a_{1j_1} a_{2j_2} \ldots a'_{nj_n}.$$

Therefore

$$\begin{aligned} |A|' &= \sum_{j} (-1)^{t(j)} a'_{1j_1} a_{2j_2} \cdots a_{nj_n} + \sum_{j} (-1)^{t(j)} a_{1j_1} a'_{2j_2} \cdots a_{nj_n} \\ &\quad + \cdots + \sum_{j} (-1)^{t(j)} a_{1j_1} a_{2j_2} \cdots a'_{nj_n}. \end{aligned}$$

Hence, we conclude that

$$|A|' = \sum_i |A|_i.$$

□

Definition: A *Minor* of order $n-1$ of A is defined to be the determinant of a submatrix of A obtained by striking out one row and one column from A. The minor obtained by striking out the i^{th} row and j^{th} column is written $M_{ij} \quad (1 \leq i, j \leq n)$. The *cofactor* A_{ij} of an element a_{ij} is given by: $A_{ij} = (-1)^{i+j} M_{ij}$.

Theorem 3.4.1 *(Cofactor expansion). The determinant of A can be computed as follows:*

$$|A| = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in},$$

or similarly,

$$|A| = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj}$$

For the following two lemmas refer to ([13], page 10).

Lemma 3.4.5 *Let $A(\lambda) = |\lambda I - A|$. Denote by $A_{ij}(\lambda)$ the cofactor of the ij^{th} element of the matrix $\lambda I - A$. I is the $n \times n$ identity matrix. Then*

$$\frac{dA(\lambda)}{d\lambda} = \sum_{i=1}^{n} A_{ii}(\lambda).$$

Proof: By applying Lemma 3.4.4 to the determinant $A(\lambda)$, the i^{th} row of $A_i(\lambda)$ consists of zeroes except the i^{th} position which is 1. Then expanding each $A_i(\lambda)$ along this row by the previous theorem yields the desired result. □

Lemma 3.4.6 *Suppose in addition to the previous lemma that* $\lambda = 1$ *and each row of* A *sums to* 1. *Then*

$$A_{i1}(1) = A_{i2}(1) = \cdots = A_{in}(1),$$

for all $i = 1, 2, \ldots, n$.

Proof: This statement follows by using the properties of determinants in Lemma 3.4.1, Lemma 3.4.2, and Lemma 3.4.3. □

3.5 Perron's formula and some applications

Let A denote an $n \times n$ square matrix. Perron's formula permits to express an arbitrary element a_{ij}^k of the matrix A^k in terms of the eigenvalues of A and the cofactors of the matrix $\lambda I - A$.

Theorem 3.5.1 (Perron's formula) *Let* $\lambda_0, \lambda_1, \ldots, \lambda_r$ *be the eigenvalues of* A, *with algebraic multiplicities* $m_0, m_1, \ldots, m_r$, *respectively. Define* $\psi_t(\lambda)$ *by*

$$A(\lambda) = |\lambda I - A| = (\lambda - \lambda_t)^{m_t} \psi_t(\lambda), \quad t = 0, \ldots, r,$$

such that $\psi_t(\lambda)$ *are polynomials of degree* $n - m_t$ *which differ from zero for* $\lambda = \lambda_t$. *Then, we have identically for all* $i, j = 1, \ldots, n$ *and* $k = 1, 2, 3, \ldots$

$$a_{ij}^k = \sum_{t=0}^{r} \frac{1}{(m_t - 1)!} D_\lambda^{m_t - 1} \left[\frac{\lambda^k A_{ij}(\lambda)}{\psi_t(\lambda)} \right]_{\lambda = \lambda_t},$$

where $A_{ij}(\lambda)$ is the cofactor of the ij'th element of $\lambda I - A$. In this equation, $D_\lambda^{m_t-1}$ denotes the derivative of order $m_t - 1$ with respect to λ, evaluated at $\lambda = \lambda_t$

For a proof of this result refer to ([13], Section 5.). The proof is not included because it is not directly relevant to this work. What is important for this project is the applications of Perron's formula to the probability transition matrix P for an ergodic finite state Markov source. For the remaining of this section refer to ([13], Section 6.).

Note first that the largest eigenvalue of P is equal to 1 by Theorem 3.3.3. Applying Perron's formula to P yields

$$\begin{aligned} p_{ij}^k &= \frac{1}{(m_0-1)!} D_\lambda^{m_0-1}\left[\frac{\lambda^k P_{ij}(\lambda)}{p_0(\lambda)}\right]_{\lambda=1} \\ &\quad + \sum_{t=1}^{r} \frac{1}{(m_t-1)!} D_\lambda^{m_t-1}\left[\frac{\lambda^k P_{ij}(\lambda)}{p_t(\lambda)}\right]_{\lambda=\lambda_t}, \quad (*) \end{aligned}$$

in which $\lambda_0 = 1, \lambda_1, \ldots, \lambda_r$ are the eigenvalues of P and $m_0, m_1, \ldots, m_r$ their respective multiplicities, so that $m_0+m_1+\cdots+m_r = n$. The polynomials $p_0(\lambda), p_1(\lambda), \ldots, p_r(\lambda)$ are defined by

$$P(\lambda) = (\lambda-1)^{m_0} p_0(\lambda) = (\lambda-\lambda_t)^{m_t} p_t(\lambda), \quad t = 1, \ldots, r,$$

where

$$p_0(1) \neq 0, \quad p_t(\lambda_t) \neq 0, \qquad t = 1, \ldots, r.$$

This relationship has a particular importance for the ergodic Markov chain associated with P when $\lambda_0 = 1$ is a simple eigenvalue, i.e., $m_0 = 1$. But this follows directly

from Theorem 3.3.3 since $\lambda_0 > |\lambda'|$ for every other eigenvalue λ'. Indeed, if $\lambda_0 = \lambda'$ for some λ' then $|\lambda_0| = |\lambda'|$. But, $\lambda_0 > |\lambda'|$ clearly implies that $|\lambda_0| > |\lambda'|$ which yields a contradiction. In this case, the formula $(*)$ assumes the form

$$p_{ij}^k = \frac{P_{ij}(1)}{p_0(1)} + \sum_{t=1}^{r} \frac{1}{(m_t-1)!} D_\lambda^{m_t-1} \left[\frac{\lambda^k P_{ij}(\lambda)}{p_t(\lambda)} \right]_{\lambda=\lambda_t}. \quad (**)$$

By Lemma 3.4.6, $P_{ij}(1) = P_{ii}(1)$. Also, since $P(\lambda) = (\lambda - 1)p_0(\lambda)$, then, $P'(\lambda) = p_0(\lambda) + (\lambda - 1)p_0'(\lambda)$, and, $P'(1) = p_0(1) \neq 0$.

But by Lemma 3.4.5 $P'(\lambda) = \sum_i P_{ii}(\lambda)$. Therefore, $P'(1) = \sum_i P_{ii}(1) \neq 0$.

For simplicity let

$$\frac{1}{(m_t-1)!} D_\lambda^{m_t-1} \left[\frac{\lambda^k P_{ij}(\lambda)}{\lambda_t^k p_t(\lambda)} \right]_{\lambda=\lambda_t} \triangleq Q_{ijt}(k);$$

Clearly, $Q_{ijt}(k)$ represents a polynomial in k of degree not greater than $(m_t - 1)$, and we can therefore write

$$Q_{ijt}(k) = \sum_{h=0}^{m_t-1} Q_{ijt}^{(h)} k^h,$$

where the $Q_{ijt}^{(h)}$ represent some specific numbers which do not depend on k. We conclude that $(**)$ can be written as

$$p_{ij}^k = p_i + \sum_{t=1}^{r} Q_{ijt}(k) \lambda_t^k,$$

where

$$p_i = \frac{P_{ii}(1)}{P'(1)} = \frac{P_{ii}(1)}{\sum_j P_{jj}(1)}.$$

Note that by Theorem 3.3.3 the magnitude of all the remaining eigenvalues of P are less than unity. Since $Q_{ijt}(k)$ are polynomials of finite degree in k, it follows that

$$\lim_{k\to\infty} p_{ij}^{(k)} = p_i, \quad i = 1, 2, \ldots, n,$$

since

$$\lim_{k\to\infty} k^h \lambda^k = 0,$$

which is equivalent to

$$\lim_{k\to\infty} k^h |\lambda|^k = 0.$$

This argument follows by taking the ratio of two successive terms of the sequence $\{k^h|\lambda|^k\}$. It can be shown easily that this ratio is equal to $|\lambda|$ asymptotically. Since $|\lambda| < 1$, the sequence of positive numbers $\{k^h|\lambda|^k\}$ is asymptotically decreasing, and hence converges to 0. Finally, we have the following theorem.

Theorem 3.5.2 *Let P be the $n \times n$ probability transition matrix for an ergodic Markov chain. The stationary distribution is given by*

$$p_i = \frac{P_{ii}(1)}{\sum_j P_{jj}(1)}, \quad i = 1, \ldots, n.$$

3.6 Rényi's entropy rate

3.6.1 Assumptions

Let $X_1, X_2, \ldots$ be an ergodic Markov chain with transition matrix $P = (p_{ij})$ where

$$p_{ij} \stackrel{\triangle}{=} Pr\{X_{k+1} = j | X_k = i\}, \quad i, j = 1, 2, \ldots, N.$$

Suppose X_1 has distribution $\mathbf{q} = (q_1, \ldots, q_N)$. Then

$$Pr\{X_1 = i_1, \ldots, X_M = i_M\} = q_{i_1} p_{i_1 i_2} \cdots p_{i_{M-1} i_M}.$$

Let

$$V(M, \alpha) \stackrel{\triangle}{=} \sum_{i_1, i_2, \ldots, i_M} (q_{i_1} p_{i_1 i_2} \cdots p_{i_{M-1} i_M})^\alpha,$$

where $\alpha > 0$, $\alpha \neq 1$.

The Rényi entropy of $(X_1, \ldots, X_M)$ is

$$H_\alpha(M) = \frac{1}{1-\alpha} \log V(M, \alpha).$$

The base of the logarithm is arbitrary. For coding purposes, as seen in Theorem 2.4.1 from the previous chapter, we need the Rényi entropy rate defined as

$$\lim_{M \to \infty} \frac{H_\alpha(M)}{M}.$$

3.6.2 The limit

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = (p_{ij})^\alpha, \quad i, j = 1, 2, \ldots, N,$$

and define new vectors $\mathbf{s} = (s_1, \ldots, s_N)$ and $\mathbf{1}$ by

$$s_i = (q_i)^\alpha, \quad \mathbf{1}^T = (1, 1, \ldots, 1).$$

Then, clearly $V(M, \alpha)$ can be written as

$$V(M, \alpha) = \mathbf{s}R^{M-1}\mathbf{1}.$$

Theorem 3.6.1 *If $P > 0$, then*

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log \lambda(\alpha, P)}{1-\alpha},$$

where $\lambda(\alpha, P)$ is the largest positive eigenvalue of R.

Proof: By definition of R, if $P > 0$ then clearly $R > 0$. By Theorem 3.3.1, R has a positive eigenvalue $\lambda = \lambda(\alpha, P)$ with the property that $\lambda > |\lambda'|$ for any other eigenvalue λ' of R. Also, R has positive left and right eigenvectors $\mathbf{a}$ and $\mathbf{b}$, say, corresponding to the eigenvalue λ. Here, $\mathbf{q}$, $\mathbf{a}$, and $\mathbf{s}$ are row vectors, while $\mathbf{b}$ and $\mathbf{1}$ are column vectors. By Corollary 3.3.5,

$$\lim_{M\to\infty} \frac{R^{M-1}}{\lambda^{M-1}} = \mathbf{ba}.$$

Also, we have

$$\lim_{M\to\infty} \frac{\log V(M, \alpha)}{M} = \lim_{M\to\infty} M^{-1} \log \left[\frac{\mathbf{s}R^{M-1}\mathbf{1}}{\lambda^{M-1}} . \lambda^{M-1} \right].$$

Consider first the limit

$$\lim_{M\to\infty} \log \left[\frac{\mathbf{s}R^{M-1}\mathbf{1}}{\lambda^{M-1}} \right].$$

Since the logarithm is a continuous function and the limit of its argument exists then by definition of the limit of a function we have

$$\lim_{M\to\infty} \log\left[\frac{\mathbf{s}R^{M-1}\mathbf{1}}{\lambda^{M-1}}\right] = \log\left[\mathbf{s}\lim_{M\to\infty}\frac{R^{M-1}}{\lambda^{M-1}}\mathbf{1}\right] = \log[\mathbf{sba1}] = C,$$

where C is some constant. Therefore,

$$\lim_{M\to\infty} M^{-1}\log\left[\frac{\mathbf{s}R^{M-1}\mathbf{1}}{\lambda^{M-1}}\right] = \lim_{M\to\infty} M^{-1}C = 0.$$

Now, clearly

$$\lim_{M\to\infty} M^{-1}\log\left[\lambda^{M-1}\right] = \lim_{M\to\infty} M^{-1}(M-1)\log\lambda = \log\lambda.$$

Since

$$M^{-1}\log\left[\frac{\mathbf{s}R^{M-1}\mathbf{1}}{\lambda^{M-1}}.\lambda^{M-1}\right] = M^{-1}\log\left[\frac{\mathbf{s}R^{M-1}\mathbf{1}}{\lambda^{M-1}}\right] + M^{-1}\log\left[\lambda^{M-1}\right],$$

and the limit of each term of the right hand side of this equality exists, then,

$$\lim_{M\to\infty}\frac{\log V(M,\alpha)}{M} = 0 + \log\lambda = \log\lambda(\alpha,P),$$

and so

$$\lim_{M\to\infty}\frac{H_\alpha(M)}{M} = \frac{\log\lambda(\alpha,P)}{1-\alpha}. \tag{3.1}$$

□

Now, we need two lemmas in order to prove a similar result when $P \geq 0$.

Lemma 3.6.1 *If $P \geq 0$ then there exists some positive number m such that $R^m > 0$.*

Proof: By Theorem 3.1.2, there exists a positive integer m, such that $P^m > 0$. An arbitrary entry of P^m is a linear combination of products of length m of elements of P, so it has the following form:

$$\sum p_{i_1 j_1} p_{i_2 j_2} \cdots p_{i_m j_m},$$

where the sum is over some $i_k, j_k \in \{1, 2, \ldots, N\}$, where $k = 1, 2, \ldots, m$.

Since $P^m > 0$, then each entry is strictly positive; therefore

$$\sum p_{i_1 j_1} p_{i_2 j_2} \cdots p_{i_m j_m} > 0.$$

But, clearly this will imply that

$$\sum p^{\alpha}_{i_1 j_1} p^{\alpha}_{i_2 j_2} \cdots p^{\alpha}_{i_m j_m} > 0,$$

where the sum, as before, is over some $i_k, j_k \in \{1, 2, \ldots, N\}$, where $k = 1, 2, \ldots, m$. But this sum is in fact an arbitrary entry of R^m; therefore $R^m > 0$. □

Lemma 3.6.2 *The largest eigenvalue of R^m is equal to the largest eigenvalue of R raised to the power m.*

Proof: Let $\{\lambda_i\}$, $i = 1, 2, \ldots, N$ be the eigenvalues of R. Clearly $\{\lambda_i^m\}$ are the eigenvalues of R^m. By the previous lemma, $R^m > 0$; therefore, by Theorem 3.3.1, there exists λ such that: $(\lambda^m) > |(\lambda')^m|$ for any other eigenvalue λ' of R, where $(\lambda^m) > 0$. But clearly, $|(\lambda')^m| = |\lambda'|^m$; hence $\lambda^m > |\lambda'|^m$. This implies that $\lambda > |\lambda'|$; therefore λ is the largest eigenvalue of R. We conclude that the largest eigenvalue λ^m of R^m is equal to the largest eigenvalue of R raised to the power m. □

Theorem 3.6.2 *If $P \geq 0$, then*

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log \lambda(\alpha, P)}{1-\alpha},$$

where $\lambda(\alpha, P)$ is the largest positive eigenvalue of R.

Proof: By Lemma 3.6.1 there exists m such that $R^m > 0$. By Theorem 3.3.1, R^m has a positive eigenvalue λ^* with the property that $\lambda^* > |\lambda'|$ for any other eigenvalue λ' of R^m. Also, R^m has positive left and right eigenvectors $\mathbf{a}$ and $\mathbf{b}$, say, corresponding to the eigenvalue λ^*. Here, $\mathbf{q}$, $\mathbf{a}$, and $\mathbf{s}$ are row vectors, while $\mathbf{b}$ and $\mathbf{1}$ are column vectors. By Corollary 3.3.5,

$$\lim_{M\to\infty} \left(\frac{R^m}{\lambda^*}\right)^{M-1} = \mathbf{ba}.$$

Also, we have

$$\lim_{M\to\infty} \frac{\log V(M,\alpha)}{M} = \lim_{M\to\infty} M^{-1} \log \left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-1}{m}} \mathbf{1}.\lambda^{*(\frac{M-1}{m})}\right].$$

Consider first the limit

$$\lim_{M\to\infty} \log \left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-1}{m}} \mathbf{1}\right].$$

Since the logarithm is a continuous function and the limit of its argument exists, then by definition of the limit of a function we have

$$\lim_{M\to\infty} \log \left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-1}{m}} \mathbf{1}\right] = \log \left[\mathbf{s}\left(\lim_{M\to\infty}\left(\frac{R^m}{\lambda^*}\right)^{M-1}\right)^{\frac{1}{m}} \mathbf{1}\right] = \log \left[\mathbf{s}(\mathbf{ab})^{\frac{1}{m}}\mathbf{1}\right] = C,$$

where C is some constant. Therefore,

$$\lim_{M\to\infty} M^{-1} \log \left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-1}{m}} \mathbf{1}\right] = \lim_{M\to\infty} M^{-1} C = 0.$$

Now, clearly

$$\lim_{M\to\infty} M^{-1}\log\left[\lambda^{*(\frac{M-1}{m})}\right] = \lim_{M\to\infty} M^{-1}(\frac{M-1}{m})\log\lambda^* = \frac{\log\lambda^*}{m}.$$

Since

$$M^{-1}\log\left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-1}{m}}\mathbf{1}.\lambda^{*(\frac{M-1}{m})}\right] = M^{-1}\log\left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-1}{m}}\mathbf{1}\right] + M^{-1}\log\left[\lambda^{*(\frac{M-1}{m})}\right],$$

and the limit of each term of the right hand side of this equality exists, then,

$$\lim_{M\to\infty}\frac{\log V(M,\alpha)}{M} = 0 + \frac{\log\lambda^*}{m} = \frac{\log\lambda^*}{m}.$$

But by Lemma 3.6.2, $\lambda^* = \lambda^m$, where λ is the largest eigenvalue of R. Therefore

$$\frac{\log\lambda^*}{m} = \frac{\log\lambda^m}{m} = \log\lambda.$$

Thus

$$\lim_{M\to\infty}\frac{H_\alpha(M)}{M} = \frac{\log\lambda(\alpha,P)}{1-\alpha}, \tag{3.2}$$

which is the same result as when $P > 0$. □

Remark: The function

$$f(\alpha) = \frac{\log\lambda(\alpha,P)}{1-\alpha},$$

is not monotonic in α. For notational convenience set $\lambda(\alpha,P) = \lambda$.

$$f'(\alpha) = \frac{\left[\frac{\lambda'(1-\alpha)}{\lambda} + \log\lambda\right]}{(1-\alpha)^2}.$$

We have two cases.

First case: $0 < \alpha < 1$. By Corollary 3.3.4 λ is a strictly decreasing function of α.

Therefore, $\lambda' < 0$. But $\lambda > 0$ and $1 - \alpha > 0$, therefore

$$\frac{\lambda'(1-\alpha)}{\lambda} < 0.$$

Since $\lambda(\alpha, P)$ is a decreasing function of α and $0 < \alpha < 1$, then $\lambda(\alpha, P) > \lambda(1, P) = 1$. Hence $\log \lambda > 0$. Therefore, if $\log \lambda$ is greater then the absolute value of $\lambda'(1-\alpha)/\lambda$, then $f'(\alpha) > 0$, otherwise, $f'(\alpha) < 0$. Hence, $f(\alpha)$ is not monotonic.

Second case: $\alpha > 1$. In this case $\lambda'(1-\alpha)/\lambda > 0$, and $\log \lambda < 0$. By similar argument as before, $f(\alpha)$ is not monotonic.

Some numerical examples will be given in Chapter 4.

3.7 A source coding theorem for 1^{st} order Markov sources

By (2.15) we have

$$H_\alpha(M) \leq L_M(t) < H_\alpha(M) + 1.$$

Dividing by M yields

$$\frac{H_\alpha(M)}{M} \leq \frac{L_M(t)}{M} < \frac{H_\alpha(M)}{M} + \frac{1}{M}.$$

By Theorem 3.6.1 and Theorem 3.6.2 we have

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log \lambda}{1-\alpha}.$$

Therefore,

$$\lim_{M\to\infty} \frac{L_M(t)}{M} = \frac{\log \lambda}{1-\alpha}.$$

Thus, the following theorem holds for ergodic Markov sources of first order with probability transition matrix $P = (p_{ij})$.

Theorem 3.7.1 *Let $\alpha = (1+t)^{-1}$. By encoding sufficiently long sequences of input symbols from an ergodic Markov source of first order, it is possible to make the average code length of order t per input symbol as close to*

$$\frac{\log \lambda(\alpha, P)}{1-\alpha}$$

as desired where $\lambda(\alpha, P)$ denotes the largest positive eigenvalue of the matrix $R = (p_{ij}^{\alpha})$.

Now, we will illustrate with some examples.

3.8 Special cases

3.8.1 Memoryless sources

If the source is memoryless, $p_{ij} = p_j$ and R consists of N identical rows, each being $(p_1^{\alpha}, \ldots, p_N^{\alpha})$. For this R, $\mathbf{1}$ is a right eigenvector with eigenvalue

$$\sum_{i=1}^{N} (p_i)^{\alpha}.$$

Since the right eigenvector is positive, this is the largest eigenvalue by Corollary 3.3.2. Thus, by (3.2)

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log\left(\sum_{i=1}^{N}(p_i)^\alpha\right)}{1-\alpha} = H_\alpha,$$

which is consistent with Lemma 2.3.1.

3.8.2 Markov sources with symmetry properties

The last result generalizes to any matrix P for which every row is some permutation of the first row. Let every row of P consist of the numbers $p_1, \ldots, p_N$ in some order, where $p_i \geq 0$ and $\sum p_i = 1$. Then $\mathbf{1}$ is a right eigenvector of R, with eigenvalue

$$\sum_{i=1}^{N}(p_i)^\alpha.$$

As before,

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log\left(\sum_{i=1}^{N}(p_i)^\alpha\right)}{1-\alpha}.$$

3.8.3 Binary Markov sources

For a binary Markov source we can calculate the eigenvalues and eigenvectors explicitly and examine the result. Let the transition matrix be

$$P = \begin{bmatrix} x & 1-x \\ 1-y & y \end{bmatrix},$$

where $x > 0$ and $y > 0$. The stationary distribution for this P is the left eigenvector

$$v = \left(\frac{1-y}{2-x-y}, \frac{1-x}{2-x-y} \right). \tag{3.3}$$

The largest eigenvalue of R is found to be

$$\lambda(\alpha, P) = \frac{1}{2} \left(x^\alpha + y^\alpha + [(x^\alpha - y^\alpha)^2 + 4(1-x)^\alpha (1-y)^\alpha]^{1/2} \right). \tag{3.4}$$

$$\begin{aligned}
\lim_{\alpha \to 1} \lambda(\alpha, P) &= \frac{1}{2} \left(x + y + [(x-y)^2 + 4(1-x)(1-y)]^{1/2} \right) \\
&= \frac{1}{2} \left(x + y + (x^2 + y^2 + 2xy + 4 - 4y - 4x)^{1/2} \right) \\
&= \frac{1}{2} \left(x + y + \left[(2-x-y)^2 \right]^{1/2} \right) \\
&= \frac{1}{2}(x + y + 2 - x - y) \\
&= 1.
\end{aligned}$$

Then, by l'Hôpital's rule (natural logarithm is used for convenience), we find that

$$\lim_{\alpha \to 1} \frac{\ln \lambda(\alpha, P)}{1-\alpha} = -\lambda'(1, P). \tag{3.5}$$

From (3.1) and (3.4),

$$\begin{aligned}
\lim_{\alpha \to 1} \lim_{M \to \infty} \frac{H_\alpha(M)}{M} &= -\lambda'(1, P) \\
&= -\left. \frac{(x^\alpha \ln x + y^\alpha \ln y)}{2} \right|_{\alpha=1} \\
&\quad \left. - \frac{[(x^\alpha - y^\alpha)^2 + 4(1-x)^\alpha (1-y)^\alpha]'}{4\,[(x^\alpha - y^\alpha)^2 + 4(1-x)^\alpha (1-y)^\alpha]^{1/2}} \right|_{\alpha=1} \\
&= -\frac{(2-x-y)(x \ln x + y \ln y) + (x-y)(x \ln x - y \ln y)}{2(2-x-y)}
\end{aligned}$$

$$
\begin{aligned}
& -\frac{(1-x)(1-y)(\ln(1-x)+\ln(1-y))}{2-x-y} \\
= & -\frac{x\ln x + y\ln y - xy\ln y - xy\ln x}{2-x-y} \\
& -\frac{(1-x)(1-y)\ln(1-x)+(1-x)(1-y)\ln(1-y)}{2-x-y} \\
= & -\frac{1-y}{2-x-y}[x\ln x+(1-x)\ln(1-x)] \\
& -\frac{1-x}{2-x-y}[y\ln y+(1-y)\ln(1-y)].
\end{aligned}
$$

In view of (3.3), this is the Shannon conditional entropy associated with this Markov chain.

3.8.4 Limiting case for N-ary Markov sources

We now consider an ergodic Markov source $\{X_n\}$ of first order with alphabet size N. Let $P = (p_{ij})$ denotes the probability transition matrix and $R = (p_{ij}^{\alpha})$, $i, j = 1, 2, \ldots, N$. The goal is to find the limit of (3.2) as $\alpha \to 1$. For binary Markov sources, as seen in the previous section, the limiting value of (3.2) is easy to compute since the eigenvalues and eigenvectors can be explicitely determined. However, this calculation for N-ary Markov sources is more complicated, because in general there is no closed form for the eigenvalues and the eigenvectors. The eigenvalues of P are continuous functions of its elements [12]. Note that as $\alpha \to 1$, $R \to P$ and that the largest eigenvalue of the matrix P is 1 by Theorem 3.3.3. Hence

$$\lim_{\alpha \to 1} \lambda(\alpha, P) = 1.$$

From this we see that (3.5) holds for any N. The equation defining the largest positive eigenvalue of R, $\lambda(\alpha, P)$ is

$$\begin{vmatrix} p_{11}^{\alpha} - \lambda & p_{12}^{\alpha} & \cdots & p_{1N}^{\alpha} \\ p_{21}^{\alpha} & p_{22}^{\alpha} - \lambda & \cdots & p_{2N}^{\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1}^{\alpha} & p_{N2}^{\alpha} & \cdots & p_{NN}^{\alpha} - \lambda \end{vmatrix} = 0. \tag{3.6}$$

By differentiating this equation with respect to α, we get by Lemma 3.4.4

$$D_1 + D_2 + \cdots + D_N = 0, \tag{3.7}$$

where D_i is the determinant obtained from (3.6) by replacing the i-th row by

$$(p_{i1}^{\alpha} \ln p_{i1}, p_{i2}^{\alpha} \ln p_{i2}, \ldots, p_{ii}^{\alpha} \ln p_{ii} - \lambda', \ldots, p_{iN}^{\alpha} \ln p_{iN}).$$

and leaving the other $N-1$ rows unchanged. In this equation, λ' denotes the derivative of λ with respect to α.

Note that if $\alpha = 1$ then $\lambda = 1$. Also, by Lemma 3.4.3 if we add in D_i all the other columns to the i-th column, the value of the determinant remains unchanged. Therefore, for $\alpha = 1$ D_i is the determinant with i-th row

$$(p_{i1} \ln p_{i1}, p_{i2} \ln p_{i2}, \ldots, -H(X|i) - \lambda', \ldots, p_{iN} \ln p_{iN}),$$

where

$$H(X|i) = -\sum_{j=1}^{N} p_{ij} \ln p_{ij}.$$

The k-th row of D_i for $k > i$ is

$$(p_{k1}, p_{k2}, \ldots, 0, p_{kk} - 1, \ldots, p_{kN}),$$

and for $k < i$,

$$(p_{k1}, p_{k2}, \ldots, p_{kk} - 1, 0, \ldots, p_{kN}).$$

A 0 occurs in the i-th position because clearly

$$\sum_{j=1}^{N} p_{kj} - 1 = 0.$$

We conclude that

$$D_i = (-H(X|i) - \lambda')c_i, \tag{3.8}$$

where c_i is the $N-1 \times N-1$ cofactor of $p_{ii} - 1$ in the determinant of (3.6) for the case $\alpha = 1$, given by

$$c_i = \begin{vmatrix} p_{11}-1 & p_{12} & \ldots & p_{1,i-1} & p_{1,i+1} & \ldots & p_{1N} \\ p_{21} & p_{22}-1 & \ldots & p_{2,i-1} & p_{2,i+1} & \ldots & p_{2N} \\ \vdots & \vdots & \ldots & \ldots & \ldots & \ldots & \vdots \\ p_{i-1,1} & p_{i-1,2} & \ldots & p_{i-1,i-1}-1 & p_{i-1,i+1} & \ldots & p_{i-1,N} \\ p_{i+1,1} & p_{i+1,2} & \ldots & p_{i+1,i-1} & p_{i+1,i+1}-1 & \ldots & p_{i+1,N} \\ \vdots & \vdots & \ldots & \ldots & \ldots & \ldots & \vdots \\ p_{N1} & p_{N2} & \ldots & p_{N,i-1} & p_{N,i+1} & \ldots & p_{NN}-1 \end{vmatrix}.$$

By Substituting (3.8) in (3.7) we get

$$\lim_{\alpha \to 1} \frac{\ln \lambda(\alpha, P)}{1-\alpha} = -\lambda'(1, P) = \sum_{i=1}^{N} p_i H(X|i), \tag{3.9}$$

where

$$p_i = \frac{c_i}{\sum_j c_j}.$$

But, from Theorem 3.5.2 $(p_1, \ldots, p_N)$ as defined above, is the stationary probability vector of P. Hence the value given in (3.9) is just the Shannon conditional entropy $H(X_2|X_1)$ associated with the Markov source $\{X_n\}$.

Chapter 4

Extension for k^{th} order ergodic Markov sources

We first examine second and third order ergodic Markov sources, and then generalize for ergodic Markov sources of order k. We start by defining a k^{th} order Markov chain.

Definition: A discrete stochastic process $Z_1, Z_2, \ldots$ is said to be a k^{th} *order Markov chain* if, for $n \geq k$

$$Pr(Z_{n+1} = z_{n+1} | Z_n = z_n, Z_{n-1} = z_{n-1}, \ldots, Z_1 = z_1)$$

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

For the sake of simplicity, and without loss of generality, all Markov sources in this chapter are assumed to be binary with state space $\mathcal{Z} = \{0, 1\}$.

4.1 Second order ergodic Markov sources

Let $\{Z_n\}$ be a second order ergodic Markov source. Define the process $\{W_n\}$ such that each random variable W_n is a 2-step blocking of the process $\{Z_n\}$, i.e.

$$W_n \stackrel{\triangle}{=} (Z_n, Z_{n+1}).$$

We have

$$\begin{aligned} P(W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_1 = w_1) &= P(Z_{n+1} = z_{n+1}, Z_n = z_n | Z_n, \ldots, Z_1) \\ &= P(Z_{n+1} = z_{n+1} | Z_n = z_n, Z_{n-1} = z_{n-1}) \\ &= P(W_n = w_n | W_{n-1} = w_{n-1}), \end{aligned}$$

where $z_n \in \{0, 1\}$ and $w_n \in \{(0,0); (0,1); (1,0); (1,1)\}$. Therefore $\{W_n\}$ is a first order Markov source with 4 states. We denote each state by its decimal representation; i.e., state 0 corresponds to state $(0,0)$ or (00); state 1 corresponds to state (01); state 2 corresponds to state (10) and state 3 corresponds to state (11).

Now, we would like to write the joint distribution of $\{Z_n\}$ in terms of the conditional probabilities of $\{W_n\}$, $p(w_n|w_{n-1}) \stackrel{\triangle}{=} P(W_n = w_n | W_{n-1} = w_{n-1})$. Suppose that W_1 has distribution $q(w_1)$. Then

$$\begin{aligned} &P(Z_1 = z_1, \ldots, Z_M = z_M) \\ &= P(Z_1 = z_1, Z_2 = z_2) P(Z_3 = z_3 | Z_1 = z_1, Z_2 = z_2) \\ &\quad P(Z_4 = z_4 | Z_2 = z_2, Z_3 = z_3) \ldots P(Z_M = z_M | Z_{M-1} = z_{M-1}, Z_{M-2} = z_{M-2}) \end{aligned}$$

$$
\begin{aligned}
&= p(z_1, z_2)P(Z_3 = z_3, Z_2 = z_2 | Z_1 = z_1, Z_2 = z_2) \\
&\quad \ldots P(Z_M = z_M, Z_{M-1} = z_{M-1} | Z_{M-1} = z_{M-1}, Z_{M-2} = z_{M-2}) \\
&= q(w_1)P(W_2 = w_2 | W_1 = w_1) \ldots P(W_{M-1} = w_{M-1} | W_{M-2} = w_{M-2}) \\
&= q(w_1)p(w_2|w_1) \ldots p(w_{M-1}|w_{M-2}) \\
&= q_{w_1} p_{w_1,w_2} \ldots p_{w_{M-2},w_{M-1}}.
\end{aligned}
$$

Let

$$V(M, \alpha) = \sum_{w_1, w_2, \ldots, w_{M-1}} (q_{w_1} p_{w_1,w_2} \ldots p_{w_{M-2},w_{M-1}})^\alpha.$$

The Rényi entropy of $(Z_1, \ldots, Z_M)$ is

$$H_\alpha(M) = \frac{1}{1-\alpha} \log V(M, \alpha).$$

The base of the logarithm is arbitrary.

For simplicity of notation denote by p_{ij} the transition probability that W_n goes from state i to state j; $i, j = 0, 1, 2, 3$. Therefore, the probability transition matrix of $\{W_n\}$ is $P = (p_{ij})$.

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = (p_{ij})^\alpha, \quad i, j = 0, 1, 2, 3.$$

Also, define new vectors $\mathbf{s} = (s_0, s_1, s_2, s_3)$ and $\mathbf{1}$ by

$$s_i = (q_i)^\alpha, \quad \mathbf{1}^T = (1, 1, 1, 1),$$

where T denote the transpose operation. Then, clearly $V(M, \alpha)$ can be written as

$$V(M, \alpha) = \mathbf{s} R^{M-2} \mathbf{1}.$$

We next observe that the matrix P is non-negative.

$$
\begin{aligned}
P &= \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ p_{10} & p_{11} & p_{12} & p_{13} \\ p_{20} & p_{21} & p_{22} & p_{23} \\ p_{30} & p_{31} & p_{32} & p_{33} \end{bmatrix} \\
&= \begin{bmatrix} p_{00} & p_{01} & 0 & 0 \\ 0 & 0 & p_{12} & p_{13} \\ p_{20} & p_{21} & 0 & 0 \\ 0 & 0 & p_{32} & p_{33} \end{bmatrix}.
\end{aligned}
$$

The zeros in the above matrix occur because of grouping. For example,

$$
\begin{aligned}
p_{02} &= p(w_2 = 2 | w_1 = 0) \\
&= p(w_2 = (1,0) | w_1 = (0,0)) \\
&= p((z_2, z_3) = (1,0) | (z_1, z_2) = (0,0)) \\
&= p(z_2 = 1, z_3 = 0 | z_1 = 0, z_2 = 0) \\
&= 0.
\end{aligned}
$$

Therefore, $\{W_n\}$ is an ergodic source with probability transition matrix $P \geq 0$.

In the next section we examine the case of third order Markov sources.

4.2 Third order ergodic Markov sources

Let $\{Z_n\}$ be a third order Markov source. Define the process $\{W_n\}$ such that each random variable W_n is a 3-step blocking of the process $\{Z_n\}$, i.e.

$$W_n \triangleq (Z_n, Z_{n+1}, Z_{n+2}).$$

We have

$$\begin{aligned}
&P(W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_1 = w_1) \\
&= P(Z_{n+2} = z_{n+2}, Z_{n+1} = z_{n+1}, Z_n = z_n | Z_{n+1}, Z_n, \ldots, Z_1) \\
&= P(Z_{n+2} = z_{n+2} | Z_{n+1} = z_{n+1}, Z_n = z_n, Z_{n-1} = z_{n-1}) \\
&= P(Z_{n+2} = z_{n+2}, Z_{n+1} = z_{n+1}, Z_n = z_n | Z_{n+1} = z_{n+1}, Z_n = z_n, Z_{n-1} = z_{n-1}) \\
&= P(W_n = w_n | W_{n-1} = w_{n-1}).
\end{aligned}$$

where $z_n \in \{0, 1\}$ and $w_n \in \{(000), (001), (010), (011), (100), (101), (110), (111)\}$. Therefore $\{W_n\}$ is a first order Markov source with 8 states. Again, We denote each state by its decimal representation. Now, we would like to write the joint distribution of $\{Z_n\}$ in terms of the conditional probabilities of $\{W_n\}$, $p(w_n|w_{n-1}) \triangleq P(W_n = w_n | W_{n-1} = w_{n-1})$. Suppose that W_1 has the distribution $q(w_1)$. Then

$$\begin{aligned}
&P(Z_1 = z_1, \ldots, Z_M = z_M) \\
&= P(Z_1 = z_1, Z_2 = z_2, Z_3 = z_3) P(Z_4 = z_4 | Z_1 = z_1, Z_2 = z_2, Z_3 = z_3) \\
&\quad P(Z_M = z_M | Z_{M-1} = z_{M-1}, Z_{M-2} = z_{M-2}, Z_{M-3} = z_{M-3})
\end{aligned}$$

$$
\begin{aligned}
&= p(z_1, z_2, z_3) P(Z_4 = z_4, Z_3 = z_3, Z_2 = z_2 | Z_1 = z_1, Z_2 = z_2, Z_3 = z_3) \\
&\quad \ldots P(Z_M = z_M, Z_{M-1} = z_{M-1}, Z_{M-2} = z_{M-2} | \\
&\quad Z_{M-1} = z_{M-1}, Z_{M-2} = z_{M-2}, Z_{M-3} = z_{M-3}) \\
&= q(w_1) P(W_2 = w_2 | W_1 = w_1) \ldots P(W_{M-2} = w_{M-2} | W_{M-3} = w_{M-3}) \\
&= q(w_1) p(w_2 | w_1) \ldots p(w_{M-2} | w_{M-3}) \\
&= q_{w_1} p_{w_1, w_2} \ldots p_{w_{M-3}, w_{M-2}}.
\end{aligned}
$$

Let

$$V(M, \alpha) = \sum_{w_1, w_2, \ldots, w_{M-2}} (q_{w_1} p_{w_1, w_2} \ldots p_{w_{M-3}, w_{M-2}})^{\alpha}.$$

The Rényi entropy of $(Z_1, \ldots, Z_M)$ is

$$H_{\alpha}(M) = \frac{1}{1 - \alpha} \log V(M, \alpha).$$

The base of the logarithm is arbitrary.

For simplicity of notation denote by p_{ij} the transition probability that W_n goes from state i to state j; $i, j = 0, 1, \ldots, 7$. Therefore, the probability transition matrix of $\{W_n\}$ is $P = (p_{ij})$.

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = (p_{ij})^{\alpha}, \quad i, j = 0, 1, \ldots, 7.$$

Also, define new vectors $\mathbf{s} = (s_0, s_1, \ldots, s_7)$ and $\mathbf{1}$ by

$$s_i = (q_i)^{\alpha}, \quad \mathbf{1}^T = (1, \ldots, 1),$$

where T denote the transpose of the vector $\mathbf{1}$ which contains 8 components. Therefore, $V(M,\alpha)$ can be written as

$$V(M,\alpha) = \mathbf{s}R^{M-3}\mathbf{1}.$$

Also, because of grouping, some entries of P are zeros. This matrix has the following form

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} & p_{05} & p_{06} & p_{07} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} & p_{27} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} & p_{37} \\ p_{40} & p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} & p_{47} \\ p_{50} & p_{51} & p_{52} & p_{53} & p_{54} & p_{55} & p_{56} & p_{57} \\ p_{60} & p_{61} & p_{62} & p_{63} & p_{64} & p_{65} & p_{66} & p_{67} \\ p_{70} & p_{71} & p_{72} & p_{73} & p_{74} & p_{75} & p_{76} & p_{77} \end{bmatrix}$$

$$= \begin{bmatrix} p_{00} & p_{01} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{12} & p_{13} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{24} & p_{25} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{36} & p_{37} \\ p_{40} & p_{41} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{52} & p_{53} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{64} & p_{65} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{76} & p_{77} \end{bmatrix}.$$

Therefore, the ergodic Markov source $\{W_n\}$ of order 3 has the probability transition matrix $P \geq 0$.

Now, we will look at the general case.

4.3 k^{th} order ergodic Markov sources

Let $\{Z_n\}$ be an ergodic Markov source of order k. Define $\{W_n\}$ as the process obtained by k-step blocking the process $\{Z_n\}$, i.e.,

$$W_n \stackrel{\triangle}{=} (Z_n, Z_{n+1}, \ldots, Z_{n+k-1}).$$

We have that

$$\begin{aligned} & P(W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_1 = w_1) \\ = \;& P(Z_{n+k-1} = z_{n+k-1}, Z_{n+k-2} = z_{n+k-2}, \ldots, Z_n = z_n | Z_{n-2+k}, Z_{n-3+k}, \ldots, Z_1) \end{aligned}$$

$$= P(Z_{n+k-1} = z_{n+k-1} | Z_{n+k-2} = z_{n+k-2}, Z_{n+k-3} = z_{n+k-3}, \ldots, Z_{n-1} = z_{n-1})$$

$$= P(W_n = w_n | W_{n-1} = w_{n-1}).$$

Therefore, $\{W_n\}$ is a first order ergodic Markov source with 2^k states; $w_n \in \{(0\cdots00), (0\cdots01), \ldots, (1\cdots11)\}$ where each string is of length k. As before, we denote each state by its decimal representation. We next write the joint distribution of $\{Z_n\}$ in terms of the conditional probabilities of $\{W_n\}$, $p(w_n|w_{n-1}) \stackrel{\triangle}{=} P(W_n = w_n | W_{n-1} = w_{n-1})$. Suppose that W_1 has the distribution $q(w_1)$. Then

$$P(Z_1 = z_1, \ldots, Z_M = z_M)$$

$$= P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k).$$

$$P(Z_{k+1} = z_{k+1} | Z_k = z_k, \ldots, Z_1 = z_1)$$

$$\ldots P(Z_M = z_M | Z_{M-1} = z_{M-1}, Z_{M-2} = z_{M-2}, \ldots, Z_{M-k} = z_{M-k})$$

$$= p(z_1, \ldots, z_k) P(Z_{k+1} = z_{k+1}, Z_k = z_k, \ldots, Z_2 = z_2 | Z_k = z_k, \ldots, Z_1 = z_1)$$

$$\ldots P(Z_M = z_M, Z_{M-1} = z_{M-1}, \ldots, Z_{M-k+1} = z_{M-k+1} |$$

$$Z_{M-1} = z_{M-1}, \ldots, Z_{M-k} = z_{M-k})$$

$$= q(w_1) P(W_2 = w_2 | W_1 = w_1) \ldots P(W_{M-k+1} = w_{M-k+1} | W_{M-k} = w_{M-k})$$

$$= q_{w_1} p_{w_1,w_2} \cdots p_{w_{M-k}, w_{M-k+1}}.$$

Let

$$V(M, \alpha) = \sum_{w_1, w_2, \ldots, w_{M-k+1}} (q_{w_1} p_{w_1,w_2} \cdots p_{w_{M-k}, w_{M-k+1}})^\alpha.$$

The Rényi entropy of $(Z_1, \ldots, Z_M)$ is

$$H_\alpha(M) = \frac{1}{1-\alpha} \log V(M, \alpha).$$

The base of the logarithm is arbitrary. For simplicity of notation denote by p_{ij} the transition probability that W_n goes from state i to state j; $i, j = 0, 1, \ldots, 2^k - 1$. Therefore, the probability transition matrix of $\{W_n\}$ is $P = (p_{ij})$.

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = (p_{ij})^\alpha, \quad i, j = 0, 1, \ldots, 2^k - 1.$$

Also, define new vectors $\mathbf{s} = (s_0, s_1, \ldots, s_{2^k-1})$ and $\mathbf{1}$ by

$$s_i = (q_i)^\alpha, \quad \mathbf{1}^T = (1, \ldots, 1),$$

where T denotes the transpose of the vector $\mathbf{1}$ which contains 2^k components.

Then, clearly $V(M, \alpha)$ can be written as

$$V(M, \alpha) = \mathbf{s} R^{M-k} \mathbf{1}.$$

Also, because of grouping some entries of P are zeros. This $2^k \times 2^k$ matrix has the following form

$$P = \begin{bmatrix}
p_{00} & p_{01} & 0 & 0 & \dots & \dots & \dots & \dots & 0 & 0 \\
0 & 0 & p_{12} & p_{13} & 0 & \dots & \dots & \dots & 0 & 0 \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
\vdots & \vdots & & & & 0 & & & \vdots & 0 \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 & p_{2^{k-1}-1,2(2^{k-1}-1)} & p_{2^{k-1}-1,2(2^{k-1}-1)+1} \\
p_{2^{k-1},0} & p_{2^{k-1},1} & 0 & 0 & \dots & \dots & \dots & \dots & 0 & 0 \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
\vdots & \vdots & & & & & & & \vdots & \vdots \\
0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 & p_{2^k-1,2^k-2} & p_{2^k-1,2^k-1}
\end{bmatrix}.$$

It is easy to check that all entries of P are zeros except possibly for the positions (i, j) such that $j = 2i \pmod{2^k}$, and $j = (2i+1) \pmod{2^k}$. Therefore, $\{W_n\}$ is an ergodic Markov source of first order with probability transition matrix $P \geq 0$.

4.3.1 Rényi entropy rate

For coding purposes we need

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M}.$$

We obtain the following result.

Theorem 4.3.1 *For an ergodic Markov source of order k*

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log \lambda(\alpha, P)}{1-\alpha},$$

where $P = (p_{ij})$ is the probability transition matrix of the associated first order ergodic Markov source obtained by k-step blocking the original k'th order Markov source, and $\lambda(\alpha, P)$ is the largest positive eigenvalue of the matrix $R = (p_{ij}^\alpha)$.

Proof: By Lemma 3.6.1 there exists m such that $R^m > 0$. By Theorem 3.3.1, R^m has a positive eigenvalue λ^* with the property that $\lambda^* > |\lambda'|$ for any other eigenvalue λ' of R^m. Also, R^m has positive left and right eigenvectors $\mathbf{a}$ and $\mathbf{b}$, say, corresponding to the eigenvalue λ^*. Here, $\mathbf{q}$, $\mathbf{a}$, and $\mathbf{s}$ are row vectors, while $\mathbf{b}$ and $\mathbf{1}$ are column vectors. By Corollary 3.3.5,

$$\lim_{M\to\infty} \left(\frac{R^m}{\lambda^*}\right)^{M-k} = \mathbf{ba}.$$

Also, we have

$$\lim_{M\to\infty} \frac{\log V(M,\alpha)}{M} = \lim_{M\to\infty} M^{-1} \log \left[\mathbf{s}\left(\frac{R^m}{\lambda^*}\right)^{\frac{M-k}{m}} \mathbf{1}.\lambda^{*(\frac{M-k}{m})}\right].$$

Consider first the limit

$$\lim_{M\to\infty} \log \left[\mathbf{s} \left(\frac{R^m}{\lambda^*} \right)^{\frac{M-k}{m}} \mathbf{1} \right].$$

Since the logarithm is a continuous function and the limit of its argument exists, then we have

$$\lim_{M\to\infty} \log \left[\mathbf{s} \left(\frac{R^m}{\lambda^*} \right)^{\frac{M-k}{m}} \mathbf{1} \right] = \log \left[\mathbf{s} \left(\lim_{M\to\infty} \left(\frac{R^m}{\lambda^*} \right)^{M-k} \right)^{\frac{1}{m}} \mathbf{1} \right] = \log \left[\mathbf{s} (\mathbf{ab})^{\frac{1}{m}} \mathbf{1} \right] = C,$$

where C is some constant. Therefore,

$$\lim_{M\to\infty} M^{-1} \log \left[\mathbf{s} \left(\frac{R^m}{\lambda^*} \right)^{\frac{M-k}{m}} \mathbf{1} \right] = \lim_{M\to\infty} M^{-1} C = 0.$$

Now, clearly

$$\lim_{M\to\infty} M^{-1} \log \left[\lambda^{*(\frac{M-k}{m})} \right] = \lim_{M\to\infty} M^{-1} (\frac{M-k}{m}) \log \lambda^* = \frac{\log \lambda^*}{m}.$$

Since

$$M^{-1} \log \left[\mathbf{s} \left(\frac{R^m}{\lambda^*} \right)^{\frac{M-k}{m}} \mathbf{1}.\lambda^{*(\frac{M-k}{m})} \right] = M^{-1} \log \left[\mathbf{s} \left(\frac{R^m}{\lambda^*} \right)^{\frac{M-k}{m}} \mathbf{1} \right] + M^{-1} \log \left[\lambda^{*(\frac{M-k}{m})} \right],$$

and the limit of each term of the right hand side of this equality exists, then

$$\lim_{M\to\infty} \frac{\log V(M,\alpha)}{M} = 0 + \frac{\log \lambda^*}{m} = \frac{\log \lambda^*}{m}.$$

But by Lemma 3.6.2 $\lambda^* = \lambda^m$, where λ is the largest eigenvalue of R. Therefore

$$\frac{\log \lambda^*}{m} = \frac{\log \lambda^m}{m} = \log \lambda.$$

Thus

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log \lambda(\alpha, P)}{1-\alpha}. \tag{4.1}$$

□

4.4 A source coding theorem for k^{th} order Markov sources

By (2.15) we have

$$H_\alpha(M) \leq L_M(t) < H_\alpha(M) + 1.$$

Dividing by M yields

$$\frac{H_\alpha(M)}{M} \leq \frac{L_M(t)}{M} < \frac{H_\alpha(M)}{M} + \frac{1}{M}.$$

By Theorem 4.3.1 we have

$$\lim_{M\to\infty} \frac{H_\alpha(M)}{M} = \frac{\log \lambda}{1-\alpha}.$$

Therefore,

$$\lim_{M\to\infty} \frac{L_M(t)}{M} = \frac{\log \lambda}{1-\alpha}.$$

Thus, the following theorem holds for ergodic Markov sources of order k with probability transition matrix $P = (p_{ij})$.

Theorem 4.4.1 *Let $\alpha = (1+t)^{-1}$. By encoding sufficiently long sequences of input symbols from an ergodic Markov source of order k it is possible to make the average code length of order t per input symbol as close to*

$$\frac{\log \lambda(\alpha, P)}{1-\alpha}$$

as desired where $\lambda(\alpha, P)$ denotes the largest positive eigenvalue of the matrix $R = (p_{ij}^{\alpha})$.

4.5 Numerical examples

This section is devoted for some numerical examples which are described in Chapter 2 of [2]. We compute the Rényi entropy rate for different Markov sources. We also verify that as $\alpha \to 1$, the Rényi entropy rate reduces to the Shannon entropy rate. The first example is a second order stationary binary Markov source $\{Z_n\}$ with state space $\{0, 1\}$. The process $\{W_n\}$ such that each random variable W_n is a 2-step blocking of $\{Z_n\}$, i.e.

$$W_n = (Z_n, Z_{n+1}),$$

is a first order stationary Markov source with 4 states. The probability transition matrix P of $\{W_n\}$ is given by

$$\begin{aligned} P &= \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ p_{10} & p_{11} & p_{12} & p_{13} \\ p_{20} & p_{21} & p_{22} & p_{23} \\ p_{30} & p_{31} & p_{32} & p_{33} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma+2\delta}{1+2\delta} & \frac{\rho}{1+2\delta} & 0 & 0 \\ 0 & 0 & \frac{\sigma+\delta}{1+2\delta} & \frac{\rho+\delta}{1+2\delta} \\ \frac{\sigma+\delta}{1+2\delta} & \frac{\rho+\delta}{1+2\delta} & 0 & 0 \\ 0 & 0 & \frac{\sigma}{1+2\delta} & \frac{\rho+2\delta}{1+2\delta} \end{bmatrix}, \end{aligned}$$

where $\rho + \sigma = 1$. The Shannon entropy rate of $\{Z_n\}$ is given by:

$$H(Z_3|Z_2, Z_1) = \frac{\sigma(\sigma+\delta)}{1+\delta} h_b\left(\frac{\rho}{1+2\delta}\right) + \frac{2\rho\sigma}{1+\delta} h_b\left(\frac{\rho+\delta}{1+2\delta}\right) + \frac{\rho(\rho+\delta)}{1+\delta} h_b\left(\frac{\rho+2\delta}{1+2\delta}\right),$$

where $h_b(.)$ is the binary entropy function. Now, we will illustrate equation (3.9) numerically for two different sets of numerical values for the variables ρ, σ and δ.

First set: $\rho = 0.4$, $\sigma = 0.6$, and, $\delta = 0.5$. By direct calculation we get

$$H(Z_3|Z_2, Z_1) = 0.846846 \quad bits.$$

The Rényi entropy rate is calculated for different values of α (close to 1 from above and below) and displayed in the following table.

α	$\frac{\log_2 \lambda(\alpha,P)}{1-\alpha} = \lim_{M\to\infty} \frac{H_\alpha(M)}{M}$
1.001	0.8466999879
0.999	0.8471080727
1.0001	0.8468478218
0.9999	0.8469669159
1.00001	0.8465902779
0.99999	0.8469171509

Observe that as α approaches 1, the Rényi entropy rate converges to the Shannon entropy rate.

The second set is: $\rho = 0.3$, $\sigma = 0.7$, and, $\delta = 0.2$.

In this case, we get:

$$H(Z_3|Z_2, Z_1) = 0.5875376 \quad nats.$$

The Rényi entropy rate is calculated for different values of α (close to 1 from above and below) and displayed in the following table.

α	$\frac{\ln \lambda(\alpha,P)}{1-\alpha} = \lim_{M\to\infty} \frac{H_\alpha(M)}{M}$
1.001	0.5873896795
0.999	0.5876852792
1.0001	0.5875212587
0.9999	0.5875727376
1.00001	0.5873817251
0.99999	0.5875982736

Observe that as α approaches 1, the Rényi entropy rate converges to the Shannon entropy rate.

The last example employs a third order stationary binary Markov source $\{Z_n\}$. The process $\{W_n\}$ obtained by 3-step blocking of $\{Z_n\}$, i.e.

$$W_n = (Z_n, Z_{n+1}, Z_{n+2}),$$

is a first order stationary Markov source with 8 states. The probability transition matrix of $\{W_n\}$ is given by

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} & p_{05} & p_{06} & p_{07} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} & p_{27} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} & p_{37} \\ p_{40} & p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} & p_{47} \\ p_{50} & p_{51} & p_{52} & p_{53} & p_{54} & p_{55} & p_{56} & p_{57} \\ p_{60} & p_{61} & p_{62} & p_{63} & p_{64} & p_{65} & p_{66} & p_{67} \\ p_{70} & p_{71} & p_{72} & p_{73} & p_{74} & p_{75} & p_{76} & p_{77} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sigma+3\delta}{1+3\delta} & \frac{\rho}{1+3\delta} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma+2\delta}{1+3\delta} & \frac{\rho+\delta}{1+3\delta} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\sigma+2\delta}{1+3\delta} & \frac{\rho+\delta}{1+3\delta} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma+\delta}{1+3\delta} & \frac{\rho+2\delta}{1+3\delta} \\ \frac{\sigma+2\delta}{1+3\delta} & \frac{\rho+\delta}{1+3\delta} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma+\delta}{1+3\delta} & \frac{\rho+2\delta}{1+3\delta} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\sigma+\delta}{1+3\delta} & \frac{\rho+2\delta}{1+3\delta} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma}{1+3\delta} & \frac{\rho+3\delta}{1+3\delta} \end{bmatrix},$$

where $\rho + \sigma = 1$. The Shannon entropy rate of $\{Z_n\}$ is given by:

$$H(Z_4|Z_3, Z_2, Z_1) = \frac{\sigma(\sigma+\delta)(\sigma+2\delta)}{(1+\delta)(1+2\delta)} h_b\left(\frac{\sigma+3\delta}{1+3\delta}\right) + \frac{3\sigma\rho(\sigma+\delta)}{(1+\delta)(1+2\delta)} h_b\left(\frac{\sigma+2\delta}{1+3\delta}\right)$$

$$+\frac{3\sigma\rho(\rho+\delta)}{(1+\delta)(1+2\delta)}h_b\left(\frac{\sigma+\delta}{1+3\delta}\right)+\frac{\rho(\rho+\delta)(\rho+2\delta)}{(1+\delta)(1+2\delta)}h_b\left(\frac{\sigma}{1+3\delta}\right),$$

where $h_b(.)$ is the binary entropy function.

Let $\sigma = 0.3$, $\rho = 0.7$, and, $\delta = 0.4$. Then

$$H(Z_4|Z_3, Z_2, Z_1) = 0.533205 \quad nats.$$

The Rényi entropy rate is calculated for different values of α (close to 1 from above and below) and displayed in the following table.

α	$\frac{\ln\lambda(\alpha,P)}{1-\alpha} = \lim_{M\to\infty}\frac{H_\alpha(M)}{M}$
1.001	0.5329080703
0.999	0.5335056605
1.0001	0.5332282164
0.9999	0.5332657811
1.00001	0.5333614224
0.99999	0.5327985806

Clearly, as $\alpha \to 1$,

$$\lim_{M\to\infty}\frac{H_\alpha(M)}{M} = H(Z_4|Z_3, Z_2, Z_1).$$

Chapter 5

Conclusions and future work

5.1 Summary

Primarily, we examine in detail a Rényi variable length source coding theorem for memoryless sources. Then, a formula for the Rényi entropy rate of ergodic Markov sources of arbitrary order is derived using Perron-Frobenius theory. This formula extends the previous theorem for these more general sources.

5.2 Future work

A possible direction is to examine more general sources. One possible source for which the results of this project can be applicable is the non-Markovian stationary ergodic

source.

In the literature several information measures other than the Shannon and the Rényi entropies have been introduced. Some of these entropies are cited in [7] along with several references about their applications. Probably, it will be also useful to obtain formulas for the asymptotic rate of these measures.

Bibliography

[1] J. Aczél and Z. Daróczy, *On measures of Information and their Characterization,* Academic Press, New York, 1975.

[2] F. Alajaji, "New results on the analysis of discrete communication channels with memory," Ph. D. Dissertation, University of Maryland, College Park, MD, August 1994.

[3] M. B. Bassat and J. Raviv, "Rényi's entropy and the probability of error," *IEEE Transactions on Information Theory*, vol. IT-24, No. 3, May 1978.

[4] E. F. Beckenbach and R.Bellman, *Inequalities*, Springer, Berlin, 1961.

[5] L. L. Campbell, "A coding theorem and Rényi's entropy," *Inform. and Control* **8**, 423–429, 1965.

[6] L. L. Campbell, "Definition of entropy by means of a coding problem," *Z. Wahrscheinlichkeitstheorie verw. Geb.* **6**, 113–118, 1996.

[7] I. J. Taneja and R. M. Capocelli, "On some inequalities and generalized entropies: A unified approach," *Cybernetics and systems: An international journal*, 16: 341–376, 1985.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, J. Wiley Inc., 1991.

[9] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, Vol. 41, No. 1, January 1995.

[10] M. Edwards, *Advanced Calculus*, The Houghton Mifflin series in Basic Mathematics, 1969.

[11] R. G. Gallager, *Discrete Stochastic Processes*, Boston, Kluwer, 1996.

[12] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd edition, Toronto, Academic, 1985.

[13] V. I. Romanovsky, *Discrete Markov Chains*, Groningen, Wolters-Noordhoff publishing, 1970.