9-2016

Probabilistic Methods In Information Theory

Erik W. Pachas
Cal State University-San Bernardino, realp73@hotmail.com

Probabilistic Methods In Information Theory

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Mathematics

---

by

Erik W. Pachas

September 2016

Probabilistic Methods In Information Theory

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Erik W. Pachas

September 2016

Approved by:

Dr. Yûichirô Kakihara, Committee Chair	Date
Dr. Hajrudin Fejzic, Committee Member	
Dr. Shawn McMurran, Committee Member	
Dr. Charles Stanton, Chair, Department of Mathematics	Dr. Corey Dunn Graduate Coordinator, Department of Mathematics

Abstract

Given a probability space, we will analyze the uncertainty, that is, the amount of information of a finite system, by studying the entropy of the system. We also extend the concept of entropy to a dynamical system by introducing a measure preserving transformation on a probability space. After showing some theorems and applications of entropy theory, we study the concept of ergodicity, which helps us to further analyze the information of the system.

Acknowledgements

First of all, I would like to express my great gratitude to my advisor Dr. Yûichirô Kakihara. His patience, dedication and understanding have contributed greatly in order to achieve my goal. I would like to thank the members of my committee, Dr. Hajrudin Fejzic and Dr. Shawn McMurran, for taking the time to read and make any comments or suggestions to improve this paper.

I would also like to thank the stuff from the math department for their support and positive attitude, which has made this graduate school experience a pleasant journey. Especially, I would like to express my appreciation to Dr. Corey Dunn. His words of encouragement and enthusiastic personality has influenced me to do my best during all these years.

Finally, I would like to thank my family for their great support and contribution to reach my goals. All this work and effort has been inspired and dedicated to my mother, Julia Flores.

Table of Contents

Abstract **iii**

Acknowledgements **iv**

1 Shannon Entropy **1**
1.1 Introduction 1
1.2 Properties and Axioms 1
1.3 Deriving The Entropy Function 8
1.4 Additional Properties of Entropy and Coding Theory 12

2 The Kolmogorov-Sinai Entropy **16**
2.1 Introduction 16
2.2 The Kolmogorov-Sinai Theorem 16
2.3 Bernoulli and Markov Shifts 22

3 Relative Entropy and Kullback-Leibler Information **26**
3.1 Introduction 26
3.2 Discrete Relative Entropy and Its Properties 26
3.3 Continuous Entropy and Relative Entropy 32
3.4 Birkhoff Pointwise Ergodic Theorem 37

4 Conclusion **41**

Bibliography **42**

Chapter 1

Shannon Entropy

1.1 Introduction

The concept of entropy is used in different fields of study such as thermodynamics, statistical mechanics, and communication theory, just to name a few. In thermodynamics, entropy is an indicator of reversibility. That is, when there is no change of entropy, the process is reversible. The unpredictability based on a lack of knowledge of positions and velocities of molecules is given by the entropy in statistical mechanics. Now, a different perspective of entropy is given in communication theory. Here we consider a message source, such as a writer or speaker. The amount of information conveyed by the message increases as the amount of uncertainty as to what message actually will be produced becomes greater [Pie80]. Thus, in general, we can state that entropy measures the amount of information given by a source, and a way to describe that source is using ergodicity. These two concepts are part of a bigger spectrum called information theory and this theory will be developed using concepts of probability theory, which will help us to generalize and understand it mathematically.

1.2 Properties and Axioms

Definition 1.2.1. *Let $n \in \mathbb{N}$ and $X = \{x_1, ..., x_n\}$ be a finite set with probability distribution $p = (p_1, ..., p_n)$. That is, $0 \leq p_j = p(x_j) \leq 1$, and these probabilities also satisfy the condition that $\sum_{j=1}^{n} p_j = 1$. We usually denote this as (X, p) and call it a complete system of events or finite scheme.*

The entropy or the Shannon entropy $H(X)$ of a finite scheme (X, p) is defined by

$$H(X) = -\sum_{j=1}^{n} p_j \log p_j.$$

We say that $H(X)$ is the measure of uncertainty or information of the system (X, p).

We will use the convention that $0 \log 0 = 0$, which is easily justified by continuity since $x \log x \to 0$ as $x \to 0$. Also, notice that adding terms of zero probability does not change the entropy. If the base of the logarithm is b, we denote the entropy as $H_b(X)$. A common units of entropy measure are base 2 and e. If the base of the logarithm is e, the entropy is measured in nats. And, if the base of the logarithm is 2, the entropy is measured in bits. Furthermore, note that entropy is a functional of the distribution of X. Consequently, it does not depend on the actual values taken by the random variable X, but only on the probabilities [CT06].

Example 1.2.1. ***Bernoulli Entropy*** *Let $X = \{0, 1\}$ be a random variable with a probability distribution $p(x_1 = 0) = 1 - p$ and $p(x_2 = 1) = p$. Then its entropy is given by*

$$H_2(X) = -p \log p - (1 - p) \log(1 - p).$$

If we differentiate the entropy function $H_2(X)$ with respect to p, we find that $H_2'(X) = H_2'(p) = -\frac{1}{\log_e 2}(\log_e p - \log_e(1 - p)) = 0$ when $p = 1/2$. That is, the entropy $H_2(X)$ attains its maximum value of 1 bit at $p = 1/2$.

Example 1.2.2. ***Geometric Entropy*** *Assume that we perform a number of independent trials until a success happens with probability p. We define the random variable X to be the number of trials required until the first success. Then X is known as a geometric random variable with parameter p and probability distribution*

$$p(X = n) = (1 - p)^{n-1} p, \qquad n = 1, 2, ...$$

Then we find the entropy of X,

$$H_2(X) = -\sum_{n=1}^{\infty} (1 - p)^{n-1} p \log(1 - p)^{n-1} p$$

$$
\begin{aligned}
&= -\left[p\log(1-p)\sum_{n=1}^{\infty}(n-1)(1-p)^{n-1} + p\log p\sum_{n=1}^{\infty}(1-p)^{n-1}\right] \\
&= -\left[p(1-p)\log(1-p)\sum_{n=0}^{\infty}n(1-p)^{n-1} + p\log p\sum_{n=0}^{\infty}(1-p)^{n}\right] \\
&= -\left[-p\log(1-p)\sum_{n=1}^{\infty}\frac{d}{dp}(1-p)^{n} + p\log p\frac{1}{1-(1-p)}\right] \\
&= -\left[-p\log(1-p)\frac{d}{dp}\sum_{n=1}^{\infty}(1-p)^{n} + \frac{p\log p}{p}\right] \\
&= -p(1-p)\log(1-p)\frac{1}{p^2} - \frac{p\log p}{p} \\
&= \frac{-(1-p)\log(1-p) - p\log p}{p}.
\end{aligned}
$$

Example 1.2.3. ***Poisson Entropy*** *A random variable $X = \{0, 1, 2, ...\}$ is said to be Poisson with parameter λ if for some $\lambda > 0$,*

$$p(x_i = i) = e^{-\lambda}\frac{\lambda^i}{i!} \qquad\qquad i = 0, 1, ...$$

If we calculate the entropy over all the possible values of a Poisson random variable, then we have

$$
\begin{aligned}
H_e(X) &= -\sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\log e^{-\lambda}\frac{\lambda^i}{i!} \\
&= -e^{-\lambda}\left[\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}(\log e^{-\lambda} + \log\lambda^i - \log i!)\right] \\
&= -e^{-\lambda}\left[\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}(-\lambda) + \sum_{i=0}^{\infty}\frac{\lambda^i}{i!}(i\log\lambda) - \sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\log i!\right] \\
&= -e^{-\lambda}\left[-\lambda e^{\lambda} + \lambda\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!}(\log\lambda) - \sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\log i!\right] \\
&= -e^{-\lambda}\left[-\lambda e^{\lambda} + \lambda e^{\lambda}\log\lambda - \sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\log i!\right] \\
&= \lambda(1-\log\lambda) + e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\log i!.
\end{aligned}
$$

Thus, the entropy of a random variable with *Poisson* distribution is given by

$$H_e(X) = \lambda(1-\log\lambda) + e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\log i!.$$

Because the entropy is calculated using the probabilities of each value of the random variable, we can easily show that $H(X) \geq 0$. We get this property since $0 \leq p_i \leq 1$, which implies that $-\log p_i \geq 0$. Given this fact, we can state that the minimum value of $H(X)$ is 0, and that this minimum value is achieved whenever we have $p_i = 0$ or 1. Can we also talk about a maximum value of $H(X)$ given this finite scheme? We will first claim that $H(X) \leq \log n$ assuming that X is taking over n distinct finite values. Before we prove our claim, we will define the set of probability distribution, and then prove a lemma in [Ash67].

Definition 1.2.2. *For $n \in \mathbb{N}$, Δ_n denotes the set of all probability distributions $p = (p_1, ..., p_n)$, i.e.,*

$$\Delta_n = \left\{ p = (p_1, ..., p_n) : \sum_{j=1}^{n} p_j = 1, p_j \geq 0, 1 \leq j \leq n \right\}.$$

Lemma 1.2.1. *Let $p, q \in \Delta_n$. Then*

$$\sum_{j=1}^{n} p_j \log q_j \ \leq \sum_{j=1}^{n} p_j \log p_j,$$

where the equality is true when $p_i = q_i$, $1 \leq i \leq n$.

Proof. Because of the convexity of $\log x$ function, we know that $\log x \leq x - 1$. Then using this inequality, we have

$$\log \frac{q_j}{p_j} \leq \frac{q_j}{p_j} - 1 \quad \text{or} \quad p_j \log \frac{q_j}{p_j} \leq p_j \frac{q_j}{p_j} - p_j \quad \text{for} \quad j = 1, ..., n.$$

Given that $\sum_{j=1}^{n} p_j = \sum_{j=1}^{n} q_j = 1$, we obtain

$$\sum_{j=1}^{n} p_j \log \frac{q_j}{p_j} \leq \sum_{j=1}^{n} (q_j - p_j) = 0.$$

Thus

$$\sum_{j=1}^{n} p_j \log \frac{q_j}{p_j} = \sum_{j=1}^{n} p_j \log q_j - \sum_{j=1}^{n} p_j \log p_j \leq 0.$$

And, the equality follows from the fact that $\log x = x - 1$ if and only if $x = 1$. □

Theorem 1.2.1. *Let $X = \{x_1, ..., x_n\}$ be a random variable with probability distribution $p = (p_1, ..., p_n)$. Then*

$$H(X) \leq \log n,$$

where the maximum value is attained if we have equally likely events, that is, $p_i = \frac{1}{n}$, $1 \leq i \leq n$.

Proof. Applying the previous lemma, we have

$$\begin{aligned} H(X) - \log n &= -\sum_{j=1}^{n} p_i \log p_i - \sum_{j=1}^{n} p_i \log n \\ &= -\sum_{j=1}^{n} p_i \log p_i + \sum_{j=1}^{n} p_i \log \frac{1}{n} \\ &\leq 0, \end{aligned}$$

which shows that $H(X) \leq \log n$. Note that if the random variable X has probabilities $p_j = \frac{1}{n}$ for $j = 1, ..., n$,

$$H(X) = H(\frac{1}{n}, ..., \frac{1}{n}) = -\sum_{j=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n,$$

which is the maximum value for $H(X)$. □

Based on the theorem, we can answer with certainty that there is a maximum value for $H(X) = \log n$ given a finite scheme of n outcomes.

Since entropy measures that amount of uncertainty, it is important to define the entropy involving two random variables. If we let $Y = \{y_1, ..., y_m\}$ be another finite set, then we define the following:

Definition 1.2.3. *Let* X *and* Y *be two random variables. The pair* (X, Y) *with joint distribution* $p(x, y)$ *has a joint entropy defined as*

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

Definition 1.2.4. *The conditional entropy* $H(X|Y)$ *of* X *given* Y *is defined by*

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

and we define conditional entropy of X *given an observed value of* $Y = y$*, by*

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

Definition 1.2.5. *For two random variables X and Y with a joint distribution $p(x, y)$, the mutual information $I(X, Y)$ between them is defined by*

$$I(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

Notice if X and Y are independent random variables then the *mutual information* between them $I(X, Y) = 0$. Now, if we want to find the amount of mutual information between the random variable X and itself, we see that $I(X, X) = H(X)$, i.e., the *self-mutual information* is the entropy of X. Using the above definition, we can easily prove the following:

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y). \end{aligned}$$

As we mentioned before, the measure of dependence between X and Y is relevant to the computation of the *mutual information* $I(X, Y)$ as well as the entropy. As another example, suppose that we know that Y gives all the information about X. Then we have that the measure of the entropy of X given Y is zero, $H(X|Y) = 0$, and it follows that there is no change of uncertainty of X given Y. The next theorem will give some important inequalities and illustrate the importance of dependence in order to arrive to equality.

Theorem 1.2.2. *Let $p, q \in \Delta_n$ be the probability distributions of X and Y respectively. Then*

i. $H(X, Y) \leq H(X) + H(Y)$

ii. $H(X|Y) \leq H(X)$

In both cases, equality holds true if and only if the random variables X and Y are independent.

Proof. (i) We have that

$$
\begin{aligned}
H(X)+H(Y) &= -\left(\sum_x p(x)\log p(x) + \sum_y p(y)\log p(y)\right) \\
&= -\left(\sum_x\sum_y p(x,y)\log p(x) + \sum_x\sum_y p(x,y)\log p(y)\right) \\
&= -\sum_x\sum_y p(x,y)\log p(x)p(y) \\
&\geq -\sum_x\sum_y p(x,y)\log p(x,y), \qquad \text{by lemma 1.4.1,} \\
&= H(X,Y).
\end{aligned}
$$

The equality holds if and only if $p(x,y) = p(x)p(y)$ for all x, y if and only if X and Y are independent.

(ii) First we claim that the compound entropy can be written as $H(X,Y) = H(Y) - H(X|Y)$. By definition we know that

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

This shows that the equation above is true. Now suppose that X and Y are independent random variables. Then

$$
\begin{aligned}
H(X|Y) &= H(X,Y) - H(Y) \\
&= H(X) + H(Y) - H(Y), \qquad \text{by (i),} \\
&= H(X).
\end{aligned}
$$

If the random variables X and Y are not independent, then

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y) \\ &< H(X) + H(Y) - H(Y), \qquad \text{also by (i)}, \\ &= H(X). \end{aligned}$$

Thus, we also proved the inequality in (ii). □

1.3 Deriving The Entropy Function

After stating and proving several properties of the entropy function $H(X)$, we want to show that the definition for such function makes sense and it is well-defined. In order to do this, we list three more properties that uniquely define the entropy function [Rom92].

i. $H(p_1, ..., p_n)$ is defined and continuous for all $p_1, ..., p_n$, where $0 \leq p_i \leq 1$ for $i = 1, ..., n$ and $\sum_{i=1}^{n} p_i = 1$. We want this function to be continuous so that small change in probabilities will result in a small change in uncertainty.

ii. $H\left(\frac{1}{n}, ..., \frac{1}{n}\right) < H\left(\frac{1}{n+1}, ..., \frac{1}{n+1}\right)$. This property tells us that the uncertainty increases as the number of outcomes increases, outcomes that are equally likely to occur. In fact, this entropy of equal likelihood is a monotonically increasing function.

iii. For $c_i \in \mathbb{N}$ and $\sum_{i}^{k} c_i = n$, we have

$$H\left(\frac{1}{n}, ..., \frac{1}{n}\right) = H\left(\frac{c_1}{n}, ..., \frac{c_k}{n}\right) + \sum_{i=1}^{k} \frac{c_i}{n} H\left(\frac{1}{c_i}, ..., \frac{1}{c_i}\right).$$

To construct this equation, let the set $X = \{x_1, ..., x_n\}$ be partitioned into nonempty disjoint subsets $C_1, ..., C_k$. Let the size of each subset be $|C_i| = c_i$ for $i = 1, ..., k$, and $\sum_{i=1}^{k} c_i = n$. Now, let us choose a subset C_i with probability proportional to its size. That is to say, $P(C_i) = \frac{c_i}{n}$. After that, we choose an element from the

subset C_i with equal probability. If the element x_j is in the subset C_u, then because

$$P(x_j|C_i) = \begin{cases} 0 & \text{if } i \neq u \\ \dfrac{1}{c_u} & \text{if } i = u \end{cases}$$

we have

$$P(x_j) = \sum_{i=1}^{n} P(x_j|C_i)P(C_i) = \frac{1}{c_u}\frac{c_u}{n} = \frac{1}{n}.$$

This shows that if we choose x_j this way, the probability will be the same as if we were to choose directly from the whole set X with equal probability. Consequently, the uncertainty of the outcomes remains the same.

If we choose directly from X with equal probability, the uncertainty will be

$$H\left(\frac{1}{n}, ..., \frac{1}{n}\right).$$

But now if we choose one of the subsets $C_1, ..., C_k$, the uncertainty is

$$H\left(\frac{c_1}{n}, ..., \frac{c_k}{n}\right).$$

Now, once we have chosen the subset, we still have the uncertainty of choosing an element from that subset. Then the average uncertainty in choosing an element is

$$\sum_{i=1}^{k} P(C_i)\dot{H}\left(\frac{1}{c_i}, ..., \frac{1}{c_i}\right) = \sum_{i=1}^{k} \frac{c_i}{n} H\left(\frac{1}{c_i}, ..., \frac{1}{c_i}\right).$$

Therefore, we have

$$H\left(\frac{1}{n}, ..., \frac{1}{n}\right) = H\left(\frac{c_1}{n}, ..., \frac{c_k}{n}\right) + \sum_{i=1}^{k} \frac{c_i}{n} H\left(\frac{1}{c_i}, ..., \frac{1}{c_i}\right).$$

Theorem 1.3.1. *A function H satisfies properties (i)-(iii) if and only if it is of the form*

$$H_b(p_1, ..., p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

where $b > 1$ is the base and $p \log p = 0$ for $p = 0$.

Proof. Suppose that a function H satisfies all three properties mentioned above. Now, pick some positive integers m and n such that m divides n and $c_i = m$ for all $i = 1, ..., k$. Because $mk = \sum_{i=1}^{k} c_i = n$, we get $k = \frac{n}{m}$ and using property (iii) gives

$$\begin{aligned} H\left(\frac{1}{n}, ..., \frac{1}{n}\right) &= H\left(\frac{m}{n}, ..., \frac{m}{n}\right) + \sum_{i=1}^{k} \frac{m}{n} H\left(\frac{1}{m}, ..., \frac{1}{m}\right) \\ &= H\left(\frac{m}{n}, ..., \frac{m}{n}\right) + H\left(\frac{1}{m}, ..., \frac{1}{m}\right) \sum_{i=1}^{k} \frac{m}{n} \\ &= H\left(\frac{m}{n}, ..., \frac{m}{n}\right) + H\left(\frac{1}{m}, ..., \frac{1}{m}\right). \end{aligned}$$

Now, let $n = m^s$ where s is also a positive integer. Then the above equation becomes

$$H\left(\frac{1}{m^s}, ..., \frac{1}{m^s}\right) = H\left(\frac{1}{m^{s-1}}, ..., \frac{1}{m^{s-1}}\right) + H\left(\frac{1}{m}, ..., \frac{1}{m}\right).$$

Define the function $f(n) = H\left(\frac{1}{n}, ..., \frac{1}{n}\right)$. Now, we have

$$\begin{aligned} f(m^s) &= f(m^{s-1}) + f(m) \\ &= f(m^{s-2}) + f(m) + f(m) \\ &= f(m) + \cdots + f(m) \\ &= sf(m). \end{aligned}$$

And, this is true for all positive integers m and s. Because of property (ii), we now get

$$\begin{aligned} f(m^s) &< f(m^{s+1}), \\ sf(m) &< (s+1)f(m). \end{aligned}$$

It follows that $f(m)$ must be a positive function. Let us choose some positive numbers r, t and s so that

$$m^s \leq r^t < m^{s+1}$$

Then because f is a monotonically increasing function,

$$f(m^s) \leq f(r^t) < f(m^{s+1})$$

$$sf(m) \leq tf(r) < (s+1)f(m)$$
$$\frac{s}{t} \leq \frac{f(r)}{f(m)} < \frac{s+1}{t}.$$

Also, we have

$$s \log m \leq t \log r < (s+1) \log m,$$
$$\frac{s}{t} \leq \frac{\log r}{\log m} < \frac{s+1}{t}.$$

From the last two inequalities, we will get

$$-\frac{1}{t} \leq \frac{f(r)}{f(m)} - \frac{\log r}{\log m} < \frac{1}{t}.$$

Now since t was arbitrarily chosen, we must have

$$\frac{f(r)}{f(m)} = \frac{\log r}{\log m}$$

or

$$\frac{f(r)}{\log r} = \frac{f(m)}{\log m}$$

Since this is true for all positive integers r, we have

$$f(r) = C \log r \quad \text{for some} \quad C > 0$$

since we also know that $f(r) > 0$. Now suppose that $C = 1$ by choosing the base b of the logarithm appropriately. Then

$$f(r) = \log_b r \quad \text{for all} \quad r > 0.$$

By property (iii),

$$\begin{aligned} H(\frac{c_1}{n}, ..., \frac{c_k}{n}) &= f(n) - \sum_{i=1}^{k} \frac{c_i}{n} f(c_i) \\ &= \log_b n - \sum_{i=1}^{k} \frac{c_i}{n} \log_b c_i \\ &= \sum_{i=1}^{k} \frac{c_i}{n} \log_b n - \sum_{i=1}^{k} \frac{c_i}{n} \log_b c_i \\ &= -\sum_{i=1}^{k} \frac{c_i}{n} (\log_b c_i - \log_b n) \end{aligned}$$

$$= -\sum_{i=1}^{k} \frac{c_i}{n} \log_b \frac{c_i}{n}.$$

Since any rational $p_1, ..., p_k \in (0, 1)$ can be expressed in the form $\frac{c_1}{n}, ..., \frac{c_k}{n}$, we have

$$H_b(p_1, ..., p_k) = -\sum_{i=1}^{k} p_i \log_b p_i.$$

But we also know that H is a continuous function so this must also hold for all positive real numbers $p_1, ..., p_k$. Next, we will show that $p \log p = 0$ if $p = 0$. For simplicity, let log be the natural logarithmic function of base e and notice that

$$\begin{aligned} \lim_{p \to 0^+} p \log p &= \lim_{p \to 0^+} \frac{\log p}{1/p} \\ &= \lim_{p \to 0^+} \frac{1/p}{-1/p^2} \\ &= 0. \end{aligned}$$

Therefore,

$$H_b(p_1, ..., p_k) = -\sum_{i=1}^{k} p_i \log_b p_i$$

holds for all nonnegative real numbers $p_1, ..., p_k$ where $0 \leq p_i \leq 1$ and $\sum_{i=1}^{k} p_i = 1$ for $i = 1, ..., k$. For the converse, it is straight forward to show that the entropy function H satisfies the three properties mentioned above. □

1.4 Additional Properties of Entropy and Coding Theory

To finish this section, we will present an inequality involving binomial coefficients, which plays an important role not only in information but also in coding theory. In fact, this inequality is very useful in order to prove *The Noisy Coding Theorem* [Rom92].

Lemma 1.4.1. *Define the entropy function*

$$H_q(\lambda) = \lambda \log_q \frac{1}{\lambda} + \mu \log_q \frac{1}{\mu}$$

for $0 \leq \lambda \leq 1$ *and* $\mu = 1 - \lambda$. *Then for any integer* $q \geq 2$, *we have*

$$q^{H_q(\lambda)} = 2^{H(\lambda)}.$$

Proof. If $q \geq 2$, then

$$\begin{aligned} q^{H_q(\lambda)} &= q^{\lambda \log_q \frac{1}{\lambda} + \mu \log_q \frac{1}{\mu}} \\ &= q^{\lambda \log_q \frac{1}{\lambda}} q^{\mu \log_q \frac{1}{\mu}} \\ &= 2^{\lambda \log_2 \frac{1}{\lambda}} 2^{\mu \log_2 \frac{1}{\mu}} \\ &= 2^{\lambda \log_2 \frac{1}{\lambda} + \mu \log_2 \frac{1}{\mu}} \\ &= 2^{H(\lambda)}. \quad \square \end{aligned}$$

Theorem 1.4.1. *Let* $H(\lambda) = \lambda \log \frac{1}{\lambda} + (1-\lambda) \log \frac{1}{(1-\lambda)}$, *where* $0 \leq \lambda \leq \frac{1}{2}$. *Then*

$$\sum_{k=0}^{\lfloor \lambda n \rfloor} \binom{n}{k} \leq 2^{nH(\lambda)}$$

where $\binom{n}{k}$ *is the binomial coefficient and the upper limit* $\lfloor \lambda n \rfloor$ *of the summation is largest integer smaller or equal to* $n\lambda$ *if* $n\lambda$ *is not an integer.*

Proof. We first observe that inequality holds trivially on the endpoints of the values of λ. Specifically, if $\lambda = 0$, then $H(0) = 0$ and both sides of the inequality equal to 1. Now, if $\lambda = 1/2$, we have that $H(1/2) = 1$ and that inequality becomes $2 \leq 2^n$, which is true for $n \geq 1$. Now, suppose that $0 < \lambda < 1/2$.

From the Markov's Inequality, we know if X is a random variable that takes only non-negative values, then for any value $a > 0$

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Let us assume that X has the form $X = e^{tY}$, where Y is a random variable, and t is a real number. If we set $a = e^{tb}$, then it becomes

$$P(e^{tY} \geq e^{tb}) \leq \frac{E(e^{tY})}{e^{tb}} \quad \text{for all} \quad b \in \mathbb{R}.$$

Now, if $t < 0$, then we have $e^{tY} < e^{tb}$ if and only if $tY \geq tb$ if and only if $Y \leq b$, and so this is equivalent to

$$P(Y \leq b) \leq \frac{E(e^{tY})}{e^{tb}} \quad \text{for all} \quad b \in \mathbb{R} \quad \text{and} \quad t < 0.$$

If Y is a binomial random variable, with parameters (n, p), then

$$P(Y \leq b) = \sum_{k=0}^{b} \binom{n}{k} p^k q^{n-k}$$

where $q = 1-p$. Furthermore, $E(e^{tY})$ is the binomial moment generating function, which is well known to be

$$E(e^{tY}) = (q + pe^t)^n.$$

Thus,

$$\sum_{k=0}^{b} \binom{n}{k} p^k q^{n-k} \leq e^{-tb}(q + pe^t)^n.$$

Setting $b = \lambda n$, where $0 < \lambda < 1$, we get

$$\sum_{k=0}^{b} \binom{n}{k} p^k q^{n-k} \leq e^{-\lambda nt}(q + pe^t)^n \tag{1.1}$$

valid for $t < 0$. Let $x = e^t$ and $f(x) = x^{-\lambda n}(q + px)^n$. Since $t < 0$, we will minimize f over $0 < x < 1$. By differentiating f with respect to x, we get

$$f'(x) = nx^{-\lambda n-1}(q + px)^{(n-1)}[-\lambda(q + px) + px].$$

Thus, the value of x that will minimize f is

$$x = \frac{\lambda q}{\mu p}$$

where $\mu = 1 - \lambda$, and $\lambda < p$. Substituting this value of x into f gives

$$\begin{aligned}
\left(\frac{\lambda q}{\mu p}\right)^{-\lambda n} \left(q + p\frac{\lambda q}{\mu p}\right)^n &= \left(\frac{\lambda q}{\mu p}\right)^{-\lambda n} q^n \left(1 + \frac{\lambda}{\mu}\right)^n \\
&= \left(\frac{\lambda q}{\mu p}\right)^{-\lambda n} \left(\frac{q}{\mu}\right)^n \\
&= \lambda^{-\lambda n}\mu^{-\mu n}p^{\lambda n}q^{\mu n}
\end{aligned}$$

and (1.1) becomes

$$\sum_{k=0}^{\lambda n} \binom{n}{k} p^k q^{n-k} \leq \lambda^{-\lambda n}\mu^{-\mu n}p^{\lambda n}q^{\mu n}$$

for $\lambda < p$. Setting $p = q = \frac{1}{2}$ gives

$$\sum_{k=0}^{\lambda n} \binom{n}{k} \leq \lambda^{-\lambda n} \mu^{-\mu n}$$

for $\lambda < \frac{1}{2}$. From 1.4.1, we know $\lambda^{-\lambda n}\mu^{-\mu n} = 2^{n[-\lambda \log \lambda - \mu \log \mu]} = 2^{nH(\lambda)}$, and the above inequality becomes

$$\sum_{k=0}^{\lambda n} \binom{n}{k} \leq 2^{nH(\lambda)}. \quad \square$$

Chapter 2

The Kolmogorov-Sinai Entropy

2.1 Introduction

In the previous chapter, we developed *Shannon's* way of measuring the information of a system. This notion of measuring the amount of uncertainty of source, represented as a random variable along with its distribution, provided us with a probabilistic way of quantifying the amount of uncertainty, and we called this entropy. Now, in this chapter, we extend the concept of the entropy to a dynamical system, which is a description of a physical system and its evolution over time. Therefore, we introduce the concept of measure preserving dynamical systems and measure its unpredictability. We will be able to state how unpredictable is a dynamical system depending on its entropy. The higher the unpredictability, the higher the entropy. Furthermore, we will be able to determine how the structure of two dynamical systems relates, i.e., whether or not two dynamical system are isomorphic. To start, we will define some basic concepts and a probability measure, found in [Shr04], so that we can define a dynamical system.

2.2 The Kolmogorov-Sinai Theorem

Definition 2.2.1. *Let X be an arbitrary set. A collection $\mathfrak{X}$ of subsets of X is a σ-algebra of X if*

i. $X \in \mathfrak{X}$;

ii. If $A \in \mathfrak{X}$, *then* $A^c \in \mathfrak{X}$;

iii. If $(A_n : n \in \mathbb{N}) \subset \mathfrak{X}$, *then* $\bigcup_{n\in\mathbb{N}} A_n \in \mathfrak{X}$.

Definition 2.2.2. *Let* X *be an arbitrary set, and* $\mathfrak{X}$ *be a* σ*-algebra of* X. *A function* $\mu : \mathfrak{X} \mapsto [0, 1]$ *is a probability measure if it satisfies the following properties:*

i. $\mu(\emptyset) = 0$;

ii. $\mu(X) = 1$;

iii. For every disjoint sequence $(A_n : n \in \mathbb{N})$ *in* $\mathfrak{X}$, *we have*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Definition 2.2.3. *Let* $(X, \mathfrak{X}, \mu)$ *be a probability measure space, and* $S : X \to X$. *The transformation* S *is said to be measurable if* $S^{-1}\mathfrak{X} \subseteq \mathfrak{X}$. *That is,* $S^{-1}A \in \mathfrak{X}$ *for every* $A \in \mathfrak{X}$. *Let* S *be measurable. Then* S *is called a measure preserving transformation with respect to* μ *if*

$$\mu(S^{-1}(A)) = \mu(A) \quad \text{for every} \quad A \in \mathfrak{X}.$$

Definition 2.2.4. *Let* $(X, \mathfrak{X}, \mu)$ *be a probability measure space, and let* $S : X \to X$ *be a one-to-one measure preserving transformation. If* S^{-1} *is measurable, that is,* S *is invertible, then*

$$S^{-1}\mathfrak{X} = \mathfrak{X} = S\mathfrak{X}$$

and S^{-1} *is also a measure preserving transformation. Now, the space* $(X, \mathfrak{X}, \mu, S)$ *is called a dynamical system, where* S *is measure preserving and not necessarily invertible.*

Definition 2.2.5. *Let* $1 \leq p < \infty$. *We say that the space* $L^p(X, \mathfrak{X}, \mu)$ *consists of all complex-valued measurable functions* f *on* X *that satisfy*

$$\int_X |f(x)|^p d\mu(x) < \infty.$$

Then, if $f \in L^p(X, \mathfrak{X}, \mu)$, *we define the* L^p ***norm*** *of* f *by*

$$\|f\|_p = \left(\int_X |f(x)|^p d\mu(x)\right)^{1/p}$$

If we let $p = 1$, *the space* $L^1(X, \mathfrak{X}, \mu)$ *consists of all integrable functions on* X, *and, together with* $\|\cdot\|_1$, *is a complete normed vector space.*

Definition 2.2.6. *For our purpose, we denote L^1-space of $(X, \mathfrak{X}, \mu)$ by $L^1(\mathfrak{X})$. If $\mathfrak{Y}$ is a σ-subalgebra of $\mathfrak{X}$ and $f \in L^1(\mathfrak{X})$, let*

$$\mu_f = \int_A f \, d\mu, \quad A \in \mathfrak{Y}.$$

We notice that μ_f is a countably additive measure on $\mathfrak{Y}$ and is *absolutely continuous* with respect to μ. That is, if $A \in \mathfrak{X}$ and $\mu(A) = 0$, then $\mu_f(A) = 0$. By Radon-Nikodým Theorem, there is a unique $\mathfrak{Y}$-measurable function $g \in L^1(\mathfrak{Y})$ such that

$$\mu_f = \int_A g \, d\mu, \quad A \in \mathfrak{Y}.$$

g is unique in the μ-*a.e.* sense. If we denote $g = E(f|\mathfrak{Y})$, then g is called the conditional expectation of f relative to $\mathfrak{Y}$. If we let $f = 1_A$ be the indicator function of $A \in \mathfrak{X}$, then we denote $E(1_A|\mathfrak{Y}) = P(A|\mathfrak{Y})$, which is the conditional probability of A relative to $\mathfrak{Y}$ [Kak99].

Definition 2.2.7. *Let $\mathfrak{Y}$ be a σ-subalgebra of $\mathfrak{X}$. A $\mathfrak{Y}$-partition is a finite $\mathfrak{Y}$-measurable partition $\mathfrak{A}$ of X. That is,*

1. $\mathfrak{A} = \{A_1, ..., A_n\} \subseteq \mathfrak{Y}$,
2. $A_j \cap A_k = \emptyset$ *if* $j \neq k$,
3. $\bigcup_{j=1}^{n} A_j = X$.

Definition 2.2.8. *Consider the dynamical system $(X, \mathfrak{X}, \mu, S)$ and let the set of all $\mathfrak{Y}$-partitions be denoted by $\mathcal{P}(\mathfrak{Y})$. If we let $\mathfrak{A}, \mathfrak{B} \in \mathcal{P}(\mathfrak{Y})$, then we define the following $\mathfrak{Y}$-partitions:*

$$\mathfrak{A} \vee \mathfrak{B} = \{A \cap B : A \in \mathfrak{A}, B \in \mathfrak{B}\}$$

and

$$S^{-1}\mathfrak{A} = \{S^{-1}A : A \in \mathfrak{A}\}$$

Now, let us define the partition

$$\bigvee_{j=0}^{n-1} S^{-j}\mathfrak{A}.$$

Definition 2.2.9. *Let $\mathfrak{A} = \{A_1, ..., A_n\} \in \mathcal{P}(\mathfrak{X})$. The entropy $H(\mathfrak{A})$ of a partition $\mathfrak{A}$ is defined by*

$$\begin{aligned} H(\mathfrak{A}) &= -\sum_{j=1}^{n} \mu(A_i) \log \mu(A_i) \\ &= -\sum_{A \in \mathfrak{A}} \mu(A) \log \mu(A). \end{aligned}$$

If we let the entropy function $I(\mathfrak{A})$ of $\mathfrak{A}$ be defined by

$$I(\mathfrak{A})(\cdot) = -\sum_{A \in \mathfrak{A}} 1_A(\cdot) \log \mu(A),$$

then we have

$$H(\mathfrak{A}) = E(I(\mathfrak{A})) = \int_X I(\mathfrak{A}) d\mu.$$

Definition 2.2.10. *We define the conditional entropy function $I(\mathfrak{A}|\mathfrak{Y})$ as*

$$I(\mathfrak{A}|\mathfrak{Y})(\cdot) = -\sum_{A \in \mathfrak{A}} 1_A(\cdot) \log P(A|\mathfrak{Y})(\cdot).$$

Also, the conditional entropy $H(\mathfrak{A}|\mathfrak{Y})$ is defined by

$$H(\mathfrak{A}|\mathfrak{Y}) = E(I(\mathfrak{A}|\mathfrak{Y})) = \int_X I(\mathfrak{A}|\mathfrak{Y}) d\mu. \tag{2.1}$$

Since the entropy $H(\mathfrak{A})$ of a finite partition $\mathfrak{A} \in \mathcal{P}(\mathfrak{X})$ can be expressed as the *Shannon Entropy*, we also express the conditional entropy $H(\mathfrak{A}|\mathfrak{Y})$ as

$$\begin{aligned} H(\mathfrak{A}|\mathfrak{Y}) &= E(I(\mathfrak{A}|\mathfrak{Y})) \\ &= E(E(I(\mathfrak{A}|\mathfrak{Y})|\mathfrak{Y})) \\ &= -\sum_{A \in \mathfrak{A}} \int_X P(A|\mathfrak{Y}) \log P(A|\mathfrak{Y}) d\mu. \end{aligned}$$

Definition 2.2.11. *Let $\mathfrak{A} \in \mathcal{P}(\mathfrak{X})$. Denote $\tilde{\mathfrak{A}} = \sigma(\mathfrak{A})$, which is the σ-algebra generated by $\mathfrak{A}$. That is, $\tilde{\mathfrak{A}}$ is the smallest σ-subalgebra of $\mathfrak{X}$ that contains $\mathfrak{A}$. Also, if $\mathfrak{Y}_1$,$\mathfrak{Y}_2$ are σ-subalgebras, then let us denote $\mathfrak{Y}_1 \vee \mathfrak{Y}_2 = \sigma(\mathfrak{Y}_1 \cup \mathfrak{Y}_2)$, i.e., the σ-algebra generated by the union of $\mathfrak{Y}_1$ and $\mathfrak{Y}_2$.*

Notice that $P(A|\tilde{\mathfrak{B}}) = \sum_{B \in \mathfrak{B}} \mu(A|B) 1_B$ for $A \in \mathfrak{X}$, where $\mu(A|B)$ is the conditional probability of A given B. Then we define

$$H(\mathfrak{A}|\tilde{\mathfrak{B}}) = \sum_{B \in \mathfrak{B}} \mu(B) \sum_{A \in \mathfrak{A}} \{-\mu(A|B) \log \mu(A|B)\}, \tag{2.2}$$

where $-\sum_{A\in\mathfrak{A}} \mu(A|B)\log\mu(A|B)$ is the conditional entropy of $\mathfrak{A}$ given $B \in \mathfrak{B}$ and the above equation is the average conditional entropy of $\mathfrak{A}$ given $\mathfrak{B}$.

We also say that $\mathfrak{A} \leq \mathfrak{B}$ means that $\mathfrak{B}$ is finer than $\mathfrak{A}$, that is, each $A \in \mathfrak{A}$ can be expressed as a union of some elements in $\mathfrak{B}$.

Next, we state some fundamental theorems and lemmas so that we can prove the Kolmorogov-Sinai Entropy theorem.

Theorem 2.2.1. *Let $\mathfrak{A}, \mathfrak{B} \in \mathcal{P}(\mathfrak{X})$ and $\mathfrak{Y}$, $\mathfrak{Y}_1$, $\mathfrak{Y}_2$ be σ-subalgebras of $\mathfrak{X}$.*

1. $H(\mathfrak{A}|\{\emptyset, X\}) = H(\mathfrak{A})$.
2. $H(\mathfrak{A} \vee \mathfrak{B}|\mathfrak{Y}) = H(\mathfrak{A}|\mathfrak{Y}) + H(\mathfrak{B}|\tilde{\mathfrak{A}} \vee \mathfrak{Y})$.
3. $H(\mathfrak{A} \vee \mathfrak{B}) = H(\mathfrak{A}) + H(\mathfrak{B}|\tilde{\mathfrak{A}})$.
4. $\mathfrak{A} \leq \mathfrak{B} \Rightarrow H(\mathfrak{A}|\mathfrak{Y}) \leq H(\mathfrak{B}|\mathfrak{Y})$.
5. $\mathfrak{A} \leq \mathfrak{B} \Rightarrow H(\mathfrak{A}) \leq H(\mathfrak{B})$.
6. $\mathfrak{Y}_2 \subseteq \mathfrak{Y}_1 \Rightarrow H(\mathfrak{A}|\mathfrak{Y}_1) \leq H(\mathfrak{A}|\mathfrak{Y}_2)$.
7. $H(\mathfrak{A}|\mathfrak{Y}) \leq H(\mathfrak{A})$.
8. $H(\mathfrak{A} \vee \mathfrak{B}|\mathfrak{Y}) \leq H(\mathfrak{A}|\mathfrak{Y}) + H(\mathfrak{B}|\mathfrak{Y})$.
9. $H(\mathfrak{A} \vee \mathfrak{B}) \leq H(\mathfrak{A}) + H(\mathfrak{B})$.
10. $H(S^{-1}\mathfrak{A}|S^{-1}\mathfrak{Y}) = H(\mathfrak{A}|\mathfrak{Y})$.
11. $H(S^{-1}\mathfrak{A}) = H(\mathfrak{A})$.
12. $I(S^{-1}\mathfrak{A}|S^{-1}\mathfrak{Y}) = I(\mathfrak{A}|\mathfrak{Y}) \circ S$.
13. $I(S^{-1}\mathfrak{A}) = I(\mathfrak{A}) \circ S$.

Definition 2.2.12. *Let $\mathfrak{A} \in \mathcal{P}(\mathfrak{X})$. The entropy $H(\mathfrak{A}, S)$ of S relative to $\mathfrak{A}$ is defined by*

$$H(\mathfrak{A}, S) = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{j=0}^{n-1} S^{-j}\mathfrak{A}\right). \tag{2.3}$$

We also define $H(S)$ of S or the Kolmogorov-Sinai entropy of S by

$$H(S) = \sup\{H(\mathfrak{A}, S) : \mathfrak{A} \in \mathcal{P}(\mathfrak{X})\}.$$

Theorem 2.2.2. *Let $\mathfrak{A}, \mathfrak{B} \in \mathcal{P}(\mathfrak{Y})$. Then*

$$H(\mathfrak{A}, S) \leq H(\mathfrak{B}, S) + H(\mathfrak{A}|\tilde{\mathfrak{B}}).$$

Lemma 2.2.1. *If $\mathfrak{Y}_n \uparrow \mathfrak{Y}$ and $\mathfrak{A} \in \mathcal{P}(\mathfrak{X})$, then:*

1. *$I(\mathfrak{A}|\mathfrak{Y}_n) \to I(\mathfrak{A}|\mathfrak{Y})$ μ-a.e. and in L^1.*
2. *$H(\mathfrak{A}|\mathfrak{Y}_n) \downarrow H(\mathfrak{A}|\mathfrak{Y})$.*

Theorem 2.2.3. ***(Kolmogorov-Sinai)**. If S is invertible and $\mathfrak{A} \in \mathcal{P}(\mathfrak{X})$ is such that $\bigvee_{n=-\infty}^{\infty} S^n\tilde{\mathfrak{A}} = \mathfrak{X}$, then $H(S) = H(\mathfrak{A}, S)$.*

Proof. Let $\mathfrak{A}_n = \bigvee_{k=-n}^{n} S^k\mathfrak{A}$ for $n \geq 1$. Then

$$\begin{aligned}
H(\mathfrak{A}_n, S) &= \lim_{p\to\infty} \frac{1}{p} H\left(\bigvee_{j=0}^{p-1} S^{-j}\mathfrak{A}_n\right) \\
&= \lim_{p\to\infty} \frac{1}{p} H\left(\bigvee_{j=0}^{p-1} S^{-j}\left(\bigvee_{k=-n}^{n} S^k\mathfrak{A}\right)\right) \\
&= \lim_{p\to\infty} \frac{1}{p} H\left(S^n\left(\bigvee_{j=0}^{p+2n-1} S^{-j}\mathfrak{A}\right)\right) \\
&= \lim_{p\to\infty} \frac{p+2n-1}{p} \frac{1}{p+2n-1} H\left(\bigvee_{j=0}^{p+2n-1} S^{-j}\mathfrak{A}\right) \\
&= H(\mathfrak{A}, S).
\end{aligned}$$

Now, let $\mathfrak{B} \in \mathcal{P}(\mathfrak{X})$. Then,

$$\begin{aligned}
H(\mathfrak{B}, S) &\leq H(\mathfrak{A}_n, S) + H(\mathfrak{B}|\tilde{\mathfrak{A}}_n), \quad \text{theorem 2.2.2,} \\
&= H(\mathfrak{A}, S) + H(\mathfrak{B}|\tilde{\mathfrak{A}}_n) \\
&\to H(\mathfrak{A}, S) \quad (n \to \infty),
\end{aligned}$$

since $H(\mathfrak{B}|\tilde{\mathfrak{A}}_n) \downarrow H(\mathfrak{B}|\mathfrak{X})$ by lemma 2.2.1 (2) and notice that $H(\mathfrak{B}|\mathfrak{X}) = 0$. This means that $H(\mathfrak{B}, S) \leq H(\mathfrak{A}, S)$ for $\mathfrak{B} \in \mathcal{P}(\mathfrak{X})$, which implies that

$$H(\mathfrak{A}, S) = \sup\{H(\mathfrak{B}, S) : \mathfrak{B} \in \mathcal{P}(\mathfrak{X})\} = H(S). \quad \square$$

The *Kolmogorov-Sinai theorem* provides us with a way to calculate the entropy $H(S)$ of an invertible transformation S by calculating the entropy $H(\mathfrak{A}, S)$ of that invertible transformation S relative to a particular partition $\mathfrak{A}$ of X. Moreover, this theorem will help us to compute the entropy of Bernoulli shifts and Markov shifts [Kak99].

2.3 Bernoulli and Markov Shifts

We also want to study various dynamical systems and their entropies. Thus, we would like to know if these systems are isomorphic or not.

Definition 2.3.1. *Let $(X_i, \mathfrak{X}_i, \mu_i, S_i)$ $(i = 1, 2)$ be two dynamical systems. These systems are said to be isomorphic, denoted $S_1 \cong S_2$, if there exists some one-to-one and onto mapping $\varphi : X_1 \to X_2$ such that*

i. for any subset $A_1 \subseteq X_1$, $A_1 \in \mathfrak{X}_1$ iff $\varphi(A_1) \in \mathfrak{X}_2$, and $\mu_1(A_1) = \mu_2(\varphi(A_1))$;

ii. $\varphi \circ S_1 = S_2 \circ \varphi$, that is, $\varphi(S_1 x_1) = S_2 \varphi(x_1)$ for $x_1 \in X_1$.

In this case, φ is called an isomorphism.

As a matter of fact, if $S_1 \cong S_2$, then $H(S_1) = H(S_2)$. That is, the *Kolmogorov-Sinai* entropy of measure preserving transformations is invariant under isomorphism. Consequently, if $H(S_1) \neq H(S_2)$, then $S_1 \ncong S_2$. Next, we define and compute the entropy of Bernoulli and Markov shifts.

Example 2.3.1. ***Bernoulli Shifts***. *Let (X_0, p) be a finite scheme, where $X_0 = \{a_1, ..., a_l\}$ and $p = (p_1, ..., p_l) \in \Delta_l$, so that $p(a_j) = p_j$, $1 \leq j \leq l$. Consider the infinite Cartesian product*

$$X = X_0^{\mathbb{Z}} = \{x = (..., x'_{-1}, x'_0, x'_1, ...) : x_k \in X_0, k \in \mathbb{Z}\},$$

where $\mathbb{Z} = \{0, \pm 1, \pm 2, ...\}$, and the shift S on X given by

$$S : (..., x_{-1}, x_0, x_1, ...) \mapsto (..., x'_{-1}, x'_0, x'_1, ...), \text{ where } x'_k = x_{k+1}, k \in \mathbb{Z}.$$

A cylinder set is defined by

$$[x_i^0 \cdots x_j^0] = \{(..., x_{-1}, x_0, x_1, ...) : x_k = x_k^0, i \leq k \leq j\}$$

and let

$$\mu_0([x_i^0 \cdots x_j^0]) = p(x_i^0) \cdots p(x_j^0).$$

Extend μ_0 to the σ-algebra $\mathfrak{X}$ generated by all cylinder sets, denoted by μ. Note that S is measure-preserving w.r.t. μ and hence $(X, \mathfrak{X}, \mu, S)$ is a dynamical system. The shift S is called a $(p_1, ..., p_l)$-Bernoulli shift. Since $\mathfrak{A} = \{[x_0 = a_1], ..., [x_0 = a_l]\}$ is a finite partition of X and $\bigvee_{n=-\infty}^{\infty} S^n \tilde{\mathfrak{A}} = \mathfrak{X}$ by definition, we have

$$H(S) = H(\mathfrak{A}, S) = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{k=0}^{n-1} S^{-k}\mathfrak{A}\right).$$

Now $\bigvee_{k=0}^{n-1} S^{-k}\mathfrak{A} = \{[x_0 \cdots x_{n-1}] : x_j \in X_0, 0 \leq j \leq n-1\}$ and hence

$$\begin{aligned} H\left(\bigvee_{k=0}^{n-1} S^{-k}\mathfrak{A}\right) &= -\sum_{x_0,...,x_{n-1}\in X_0} \mu([x_0 \cdots x_{n-1}]) \log \mu([x_0 \cdots x_{n-1}]) \\ &= -\sum_{x_0,...,x_{n-1}\in X_0} \mu([x_0 \cdots x_{n-1}]) \log \mu([x_0]) \cdots \mu([x_{n-1}]) \\ &= -\sum_{x_0\in X_0} \mu([x_0]) \log \mu([x_0]) - \cdots - \sum_{x_{n-1}\in X_0} \mu([x_{n-1}]) \log \mu([x_{n-1}]) \\ &= nH(\mathfrak{A}) \end{aligned}$$

since $\mu([a_j]) = p(a_j) = p_j$ for $1 \leq j \leq n$. This implies that

$$H(S) = H(\mathfrak{A}) = -\sum_{j=1}^{l} p_j \log p_j.$$

A simple geometric representation of Bernoulli shifts is given by the *Baker's Transformation*, which is an area-preserving transformation of the unit square onto itself. The figure 2.1 will illustrate how to construct $\left(\frac{1}{2}, \frac{1}{2}\right)$-Bernoulli shift.

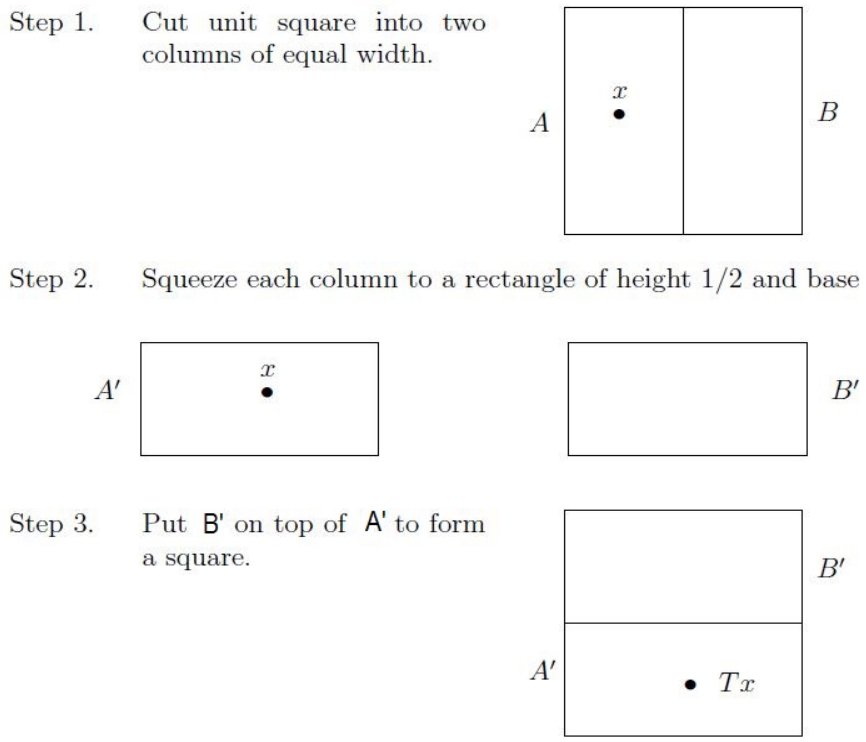

Figure 2.1: $p = (\frac{1}{2}, \frac{1}{2})$-Bernoulli shift

In the same manner, we can construct a $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$-Bernoulli shift using the *Baker's Transformation*. If we compute the entropies of a $\left(\frac{1}{2}, \frac{1}{2}\right)$-Bernoulli shift and a $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$-Bernoulli shift, we will notice that the entropies are not the same; consequently, these Bernoulli shifts are not isomorphic. In fact, we can state that two Bernoulli shifts with the same entropy are isomorphic and this was proved by Ornstein in 1970.

Example 2.3.2. ***Markov Shifts***. *Now, consider a finite scheme* (X_0, p) *and the infinite product space* $X = X_0^{\mathbb{Z}}$ *with a Bernoulli shift* S. *Let* $M = [m_{ij}]$ *be an* $l \times l$ *stochastic matrix, i.e.,* $m_{ij} \geq 0$, $\sum_{j=1}^{l} m_{ij} = 1$ *for* $1 \leq i, j \leq l$, *and* $m = (m_1, ..., m_l)$ *be a probability distribution such that* $\sum_{i=1}^{l} m_i m_{ij} = m_j$ *for* $1 \leq j \leq l$. *Each* m_{ij} *indicates the transition probability from the state* a_i *to the state* a_j *and the row vector of* m *is fixed by* M *in the sense that* $mM = m$. *We always assume that* $m_i > 0$ *for every* $i = 1, ..., l$. *Now we define* μ_0 *on* $\mathfrak{M}$, *the set of all cylinder sets, by*

$$\mu_0([a_{i_0}\cdots a_{i_n}]) = m_{i_0}m_{i_0i_1}\cdots m_{i_{n-1}i_n}.$$

μ_0 is uniquely extended to a measure μ on $\mathfrak{X}$ which is S-invariant. The shift S is called an (M,m)-Markov shift.

To compute the entropy of an (M,m)-Markov shift S consider a partition $\mathfrak{A} = \{[x_0 = a_1], ..., [x_0 = a_l]\} \in \mathcal{P}(\mathfrak{X})$, which satisfies $\bigvee_{n=-\infty}^{\infty} S^n\tilde{\mathfrak{A}} = \mathfrak{X}$. As the example before,

$$\begin{aligned} H\left(\bigvee_{k=0}^{n-1} S^{-k}\mathfrak{A}\right) &= -\sum_{x_0,\ldots,x_{n-1}\in X_0} \mu([x_0\cdots x_{n-1}])\log\mu([x_0\cdots x_{n-1}]) \\ &= -\sum_{i_0,\ldots,i_{n-1}}^{l} m_{i_0}m_{i_0i_1}\cdots m_{i_{n-2}i_{n-1}}\log m_{i_0}m_{i_0i_1}\cdots m_{i_{n-2}i_{n-1}} \\ &= -\sum_{i_0=1}^{l} m_{i_0}\log m_{i_0} - (n-1)\sum_{i,j=1}^{l} m_i m_{ij}\log m_{ij} \end{aligned}$$

since $\sum_{i=1}^{l} m_i m_{ij} = m_j$ and $\sum_{j=1}^{l} m_{ij} = 1$ for $1 \le i,j \le l$. By dividing n and letting $n \to \infty$ we get

$$H(S) = -\sum_{i,j=1}^{l} m_i m_{ij}\log m_{ij}.$$

Chapter 3

Relative Entropy and Kullback-Leibler Information

3.1 Introduction

In chapter 1, we define *mutual information* as a measure of the amount of information one random variable contains about the other one. We also observed the self-information of a random variable becomes the entropy of the same random variable. A more general case of a mutual information is the relative entropy. The *relative entropy* is a measure of the distance between two probability distributions. Here we will define the relative entropy $H(p|q)$ for two finite probability distributions p and q and provide certain properties and an application in the field of statistics. In addition, we define the relative entropy for two distributions p and q of a continuous random variable and extend the concept to an arbitrary pair of probability measures.

3.2 Discrete Relative Entropy and Its Properties

Definition 3.2.1. *Let $p, q \in \Delta_n$. The relative entropy $H(p|q)$ of p w.r.t. (with respect to) q is given by*

$$H(p|q) = \sum_{j=1}^{n} p_j \log \frac{p_j}{q_j}.$$

As before, we use the convention that $0 \log \frac{0}{0} = 0$ and $0 \log \frac{0}{q_j} = 0$. But if $p_j > 0$ and $q_j = 0$ for some j, then we define $p_j \log \frac{p_j}{0} = \infty$ and $H(p|q) = \infty$.

The notion of the relative entropy as distance of two probability distributions is not a metric because it does not satisfy the triangle inequality or the symmetry property of a metric. Notwithstanding, the relative entropy $H(p|q)$ is a measure of the inefficiency assumption [Kul97]. The next example illustrates that the relative entropy is not symmetric.

Example 3.2.1. *Let the random variable $X = \{0, 1\}$ and suppose that $\boldsymbol{p}$ and $\boldsymbol{q}$ are two probability distributions of X. Let $\boldsymbol{p} = (1-p, p)$, and let $\boldsymbol{q} = (1-q, q)$. Now, we have*

$$H_2(p|q) = (1-p)\log\frac{1-p}{1-q} + p\log\frac{p}{q}$$

and

$$H_2(q|p) = (1-q)\log\frac{1-q}{1-p} + q\log\frac{q}{p}.$$

Notice if $p = q$, then $H_2(p|q) = H_2(q|p) = 0$. Now suppose that $p = \frac{1}{4}$ and $q = \frac{1}{8}$. Then

$$H_2(p|q) = (1-\frac{1}{4})\log\frac{1-\frac{1}{4}}{1-\frac{1}{8}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{8}} = \frac{3}{4}\log\frac{6}{7} + \frac{1}{4} = 0.0832 \text{ bit.}$$

but

$$H_2(q|p) = (1-\frac{1}{8})\log\frac{1-\frac{1}{8}}{1-\frac{1}{4}} + \frac{1}{8}\log\frac{\frac{1}{8}}{\frac{1}{4}} = \frac{7}{8}\log\frac{7}{6} - \frac{1}{8} = 0.0696 \text{ bit.}$$

Thus, $H_2(p|q) \neq H_2(q|p)$.

In the next definition, we consider $p(x, y)$ and $q(x, y)$ be two joint probability distributions for the pair of random variables (X, Y) and $p(x)$ be the probability distribution for X.

Definition 3.2.2. *Given the joint probabilities $p(x, y)$ and $q(x, y)$, the conditional relative entropy is defined as*

$$H(p(y|x)|q(y|x)) = \sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)}.$$

The conditional relative entropy is the average of the relative entropies between the conditional probability distributions $p(y|x)$ and $q(y|x)$ averaged over the probability distribution $p(x)$. Now, we define the relative entropy between two joint probability distributions on (X, Y) in terms of a sum of relative entropy and a conditional relative entropy.

Theorem 3.2.1. *The relative entropy between two joint probability distributions $p(x,y)$ and $q(x,y)$ on (X,Y) is given by*

$$H(p(x,y)|q(x,y)) = H(p(x)|q(x)) + H(p(y|x)|q(y|x)).$$

Proof. Observe that

$$\begin{aligned}
H(p(x,y)|q(x,y)) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x)\cdot p(y|x)}{q(x)\cdot q(y|x)} \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x)}{q(x)} + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(y|x)}{q(y|x)} \\
&= \sum_{y\in Y,x\in X} p(x,y)\log p(x,y) + \sum_{y\in Y,x\in X} p(x,y)\log p(y) \\
&= H(p(x)|q(x)) + H(p(y|x)|q(y|x)). \quad \square
\end{aligned}$$

To prove some fundamental properties in information theory, we have used the concept of convexity. In this chapter, we will define convexity and state the Jensen's inequality. Then, we will use these concepts to show some properties of the relative entropy.

Definition 3.2.3. *A function $\varphi : (a,b) \to \mathbb{R}$ is convex if*

$$\varphi\left(\sum_{i=1}^{n}\lambda_i x_i\right) \leq \sum_{i=1}^{n}\lambda_i\varphi(x_i)$$

for all $x_i \in (a,b)$ and $\lambda_i \in [0,1]$ with $\sum_{i=1}^{n}\lambda_i = 1$. The equality holds when $x_i = x$ for some $x \in (a,b)$ and all i with $\lambda_i > 0$. We also say that φ is strictly convex if

$$\varphi\left(\sum_{i=1}^{n}\lambda_i x_i\right) < \sum_{i=1}^{n}\lambda_i\varphi(x_i).$$

Definition 3.2.4. *A function f is concave if $-f$ is convex. A function is convex if it always lies bellow any chord. A function is concave if it always lies above any chord.*

We also say a function f is convex if the function f has a second derivative that is nonnegative over an interval [Yeh06].

Lemma 3.2.1. ***Jensen's inequality.*** *Let $\varphi : (a,b) \to \mathbb{R}$ be a convex function and let $f : X \to (a,b)$ be a measurable function in L^1 on a probability space $(X, \mathfrak{X}, \mu)$. Then*

$$\varphi\left(\int f(x)\, d\mu(x)\right) \leq \int \varphi(f(x))\, d\mu(x).$$

and if φ is strictly convex, then

$$\varphi\left(\int f(x)\, d\mu(x)\right) < \int \varphi(f(x))\, d\mu(x)$$

unless $f(x) = t$ for μ-almost every $x \in X$ for some fixed $t \in (a,b)$.

In particular, If f is a convex function and X is a random variable,

$$Ef(X) \geq f(EX).$$

The equality holds true whenever $X = EX$ with probability 1, that is, X is constant.

Lemma 3.2.2. ***(Log-Sum Inequality)*** *Let $p_i, q_i > 0$ for $1 \leq i \leq n$. Then*

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

where equality holds if and only if $\frac{p_i}{q_i} =$ constant.

Proof. First we claim that the function $f(t) = t \log t$ is strictly convex. It is sufficient to show that $f'' > 0$. Then, if we differentiate twice, $f''(t) = \frac{1}{t} > 0$ when $t > 0$. By Jensen's inequality, we also know that $\sum_{i=1}^{n} \alpha_i f(t_i) \geq f(\sum_{i=1}^{n} \alpha_i t_i)$ with $\alpha_i \geq 0$, $\alpha_1 + \cdots + \alpha_n = 1$. Now, let $p_i, q_i > 0$ for $1 \leq i \leq n$. Then we have

$$\begin{aligned}
\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} &= \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) \\
&= \sum_{i=1}^{n} q_i \sum_{i=1}^{n} \frac{q_i}{\sum_{i=1}^{n} q_i} f\left(\frac{p_i}{q_i}\right) \\
&\geq \sum_{i=1}^{n} q_i f\left(\sum_{i=1}^{n} \frac{q_i}{\sum_{i=1}^{n} q_i} \cdot \frac{p_i}{q_i}\right) \\
&= \sum_{i=1}^{n} q_i f\left(\frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} q_i}\right) \\
&= \sum_{i=1}^{n} q_i \cdot \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} q_i} \log\left(\frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} q_i}\right)
\end{aligned}$$

$$= \sum_{i=1}^{n} p_i \log \left(\frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} q_i} \right). \quad \square$$

We next establish some important properties of the relative entropy by using lemma 3.2.2.

Theorem 3.2.2. *Let p and q be two probability distributions of the random variable X. Then we have*

i. ***Nonnegativity.*** *$H(p|q) \geq 0$, and $H(p|q) = 0$ if and only if $p = q$.*

ii. ***Convexity.*** *Let p_1, p_2, q_1, q_2 be probability distributions of the random variable X. Then, for $\alpha \in [0, 1]$, we have*

$$H(\alpha p_1 + (1-\alpha)p_2 | \alpha q_1 + (1-\alpha)q_2) \leq \alpha H(p_1|q_1) + (1-\alpha)H(p_2|q_2).$$

iii. ***Partition Inequality.*** *If $\mathfrak{A} = \{A_1, ..., A_k\}$ is a partition of X. That is, $\cup_{i=1}^{k} A_i = X$, and $A_i \cap A_j = \emptyset$ whenever $i \neq j$. Define*

$$p_{\mathfrak{A}}(i) = \sum_{x \in A_i} p(x), \qquad i = 1, ..., k,$$

$$q_{\mathfrak{A}}(i) = \sum_{x \in A_i} q(x), \qquad i = 1, ..., k,$$

then

$$H(p|q) \geq H(p_{\mathfrak{A}}|q_{\mathfrak{A}}),$$

where equality holds if and only if $p(x) = q(x)$ for each $x \in A_i$.

Proof.

(i.) By definition, we have

$$\begin{aligned} H(p|q) &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\ &\geq \left(\sum_{x \in X} p(x) \right) \log \left(\frac{\sum_{x \in X} p(x)}{\sum_{x \in X} q(x)} \right), \qquad \text{by lemma 3.2.2,} \\ &= 1 \log \frac{1}{1} \\ &= 0 \end{aligned}$$

and it is clear that $H(p|q) = 0$ if and only if $p(x) = q(x)$ for all $x \in X$.

(ii.) Let p_1, p_2, q_1, q_2 be probability distributions of the random variable X and $\alpha \in [0, 1]$. Then

$$
\begin{aligned}
&H(\alpha p_1 + (1-\alpha)p_2 | \alpha q_1 + (1-\alpha)q_2) \\
&\quad = \sum_{x \in X} (\alpha p_1(x) + (1-\alpha)p_2(x)) \log \frac{\alpha p_1(x) + (1-\alpha)p_2(x)}{\alpha q_1(x) + (1-\alpha)q_2(x)} \\
&\quad = \sum_{x \in X} \alpha p_1(x) \log \frac{\alpha p_1(x) + (1-\alpha)p_2(x)}{\alpha q_1(x) + (1-\alpha)q_2(x)} \\
&\quad + \sum_{x \in X} (1-\alpha)p_2(x) \log \frac{\alpha p_1(x) + (1-\alpha)p_2(x)}{\alpha q_1(x) + (1-\alpha)q_2(x)} \\
&\quad \leq \sum_{x \in X} \alpha p_1(x) \log \frac{\alpha p_1(x)}{\alpha q_1(x)} \\
&\quad + \sum_{x \in X} (1-\alpha)p_2(x) \log \frac{(1-\alpha)p_2(x)}{(1-\alpha)q_2(x)}, \quad \text{by lemma 3.2.2,} \\
&\quad = \alpha \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\alpha) \sum_{x \in X} p_2(x) \log \frac{p_2(x)}{q_2(x)} \\
&\quad = \alpha H(p_1|q_1) + (1-\alpha) H(p_2|q_2).
\end{aligned}
$$

(iii.) Let $\mathfrak{A} = \{A_1, ..., A_k\}$ be a partition of X. Then

$$
\begin{aligned}
H(p|q) &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{i=1}^{k} \sum_{x \in A_i} p(x) \log \frac{p(x)}{q(x)} \\
&\geq \sum_{i=1}^{k} \left(\sum_{x \in A_i} p(x) \right) \log \frac{\sum_{x \in A_i} p(x)}{\sum_{x \in A_i} q(x)} \\
&= \sum_{i=1}^{k} p_{\mathfrak{A}}(i) \log \frac{p_{\mathfrak{A}}(i)}{q_{\mathfrak{A}}(i)}, \quad \text{by hypothesis,} \\
&= H(p_{\mathfrak{A}} | q_{\mathfrak{A}}). \quad \square
\end{aligned}
$$

One of the applications that involves the relative entropy is a statistical hypothesis testing problem [Kul97]. In the simplest case, let

$$H_0 : p = (p(a_1), ..., p(a_n))$$

$$H_1 : q = (q(a_1), ..., q(a_n)),$$

and we have to decide which one is true depending on the samples of size k, where $p, q \in \Delta_n$ and $X_0 = \{a_1, ..., a_n\}$. We need to find a set $A \subset X_0^k$, so that, for a sample $(x_1, ..., x_k) \in A$, then H_0 is accepted; otherwise H_1 is accepted. Here, for some $\epsilon \in (0, 1)$, type 1 error probability $P(A)$ satisfies

$$\sum_{(x_1,...,x_k)\in A} \prod_{j=1}^{k} p(x_j) = P(A) \leq \epsilon$$

and type 2 error probability $Q(A)$ is given by

$$\sum_{(x_1,...,x_k)\notin A} \prod_{j=1}^{k} q(x_j) = 1 - Q(A) = Q(A^c),$$

which should be minimized. Thus, for any $\epsilon \in (0, 1)$ and the sample size $k \geq 1$ let

$$\beta(k, \epsilon) = \min\{Q(A^c) : P(A^c) > 1 - \epsilon, A \subset X^k\}.$$

Then we claim that

$$\lim_{k\to\infty} \frac{1}{k}\beta(k, \epsilon) = -\sum_{j=1}^{n} p(a_j) \log \frac{p(a_j)}{q(a_j)} = -H(p|q).$$

It follows from the claim that we can minimize the probability $Q(A^c)$ with respect to $P(A^c)$ by calculating the relative entropy $-H(p|q)$, the negative entropy of p with respect to q. A proof of this claim is found in [Kak99]. Next, we define the relative entropy given a continuous random variable.

3.3 Continuous Entropy and Relative Entropy

Definition 3.3.1. *Let X be a random variable with cumulative distribution function $F(x) = P(X \leq x)$. If $F(x)$ is continuous, the random variable is said to be continuous. Let $f(x) = F'(x)$ when the derivative is defined. If*

$$\int_{\mathbb{R}} f(x)\, dx = 1,$$

$f(x)$ is called the probability density function for X. The set where $f(x) > 0$ is called the support set of X. Now, we define the expected value of X by

$$E(X) = \int_{\mathbb{R}} x f(x)\, dx.$$

Definition 3.3.2. *The entropy $H(X)$ of a continuous random variable X with density $f(x)$ is defined as*

$$H(X) = -\int_S f(x) \log f(x)\, dx,$$

where S is the support set of the random variable, and given that the integral exists.

Now, we define the relative entropy of a continuous random variable X and show some of its properties.

Definition 3.3.3. *The relative entropy $H(f|g)$ between two densities f and g is defined by*

$$H(f|g) = \int f \log \frac{f}{g}$$

Note that $H(f|g)$ is finite only if the support set of f is contained in the support set of g.

In chapter 1, the discrete entropy is maximized when a set of events are equally likely, that is, uniformly distributed. In the continuous case, the result is similar and it is shown in the next example.

Example 3.3.1. ***(Uniform Distribution)*** *Let f be the probability density function on (a, b) given by*

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases}$$

Let us find the entropy of a random variable X with a uniform density function f. Then,

$$\begin{aligned} H(X) &= -\int_{(a,b)} f(x) \log f(x)\, dx \\ &= -\int_{(a,b)} \frac{1}{b-a} \log \frac{1}{(b-a)}\, dx \\ &= \frac{1}{b-a} \log(b-a) \int_{(a,b)} dx \\ &= \log(b-a). \end{aligned}$$

Now, assume that $f(x)$ is a probability density function on (a,b) and $g(x)$ is the uniform density function on (a,b). Then,

$$\begin{aligned} H(f|g) &= \int_{(a,b)} f(x) \log \frac{f(x)}{g(x)}\, dx \\ &= \int_{(a,b)} f(x)(\log f(x) - \log g(x))\, dx \\ &= -H(X) - \log \frac{1}{(b-a)} \int_{(a,b)} f(x)\, dx \\ &= -H(x) + \log(b-a) \geq 0. \end{aligned}$$

First we notice that, as the discrete case, the relative entropy $H(f|g) \geq 0$. Also, this example provides us with an upper bound for a probability density function on the interval (a,b), i.e., $H(X) \leq \log(b-a)$.

We also find the relative entropy between two normal and exponential probability density functions on a random variable X.

Example 3.3.2. ***(Normal Distribution)*** *A random variable X is normally distributed with parameters μ and σ^2 if the probability density function $f(x)$ is defined by*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty.$$

Suppose that f and g are normally distributed with parameters μ_1, σ_1^2 and μ_2 , σ_2^2, respectively. Then we have

$$\begin{aligned} H(f|g) &= \int f(x) \log \frac{f(x)}{g(x)}\, dx \\ &= \int f(x) \log \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(x-\mu_1)^2/2\sigma_1^2}}{\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-(x-\mu_2)^2/2\sigma_2^2}}\, dx \\ &= \int f(x) \log \left(\frac{\sigma_2^2}{\sigma_1^2}\right)^{1/2} dx + \int f(x) \left[-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] dx \\ &= \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} \int f(x)\, dx - \int f(x) \frac{(x-\mu_1)^2}{2\sigma_1^2}\, dx + \int f(x) \frac{(x-\mu_2)^2}{2\sigma_2^2}\, dx \\ &= \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} \cdot 1 - \frac{1}{2\sigma_1^2} Var(X) + \frac{1}{2\sigma_2^2} \int f(x)(x - \mu_1 + \mu_1 - \mu_2)^2\, dx \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} + \frac{1}{2\sigma_2^2}\cdot\int f(x)(x-\mu_1)^2\,dx \\
&+ \frac{1}{2\sigma_2^2}\cdot 2(\mu_1-\mu_2)\int f(x)(x-\mu_1)\,dx + \frac{1}{2\sigma_2^2}\cdot(\mu_1-\mu_2)^2\int f(x)\,dx \\
&= \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} + \frac{1}{2\sigma_2^2}(Var(X) + 2(\mu_1-\mu_2)\cdot(E(X)-\mu_1) + (\mu_1-\mu_2)^2) \\
&= \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{\sigma_2^2} - 1\right),
\end{aligned}
$$

where the variance $Var(X) = E[(X-\mu_1)^2] = \int f(x)(x-\mu_1)^2 = \sigma_1^2$.

Example 3.3.3. ***(Exponential Distribution)*** *A continuous random variable* X *is said to be exponentially distributed for* $\lambda > 0$ *if the density function is defined as*

$$
f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}
$$

The entropy $H(X)$ *with exponentially density function* f *is calculated as follows:*

$$
\begin{aligned}
H(X) &= -\int f(x)\log f(x)\,dx \\
&= -\int \lambda e^{-\lambda x}\log \lambda e^{-\lambda x}\,dx \\
&= -\int \lambda e^{-\lambda x}[\log\lambda + \log e^{-\lambda x}]\,dx \\
&= -\int \lambda e^{-\lambda x}\log\lambda\,dx - \int \lambda e^{-\lambda x}(-\lambda x)\,dx \\
&= -\log\lambda\int \lambda e^{-\lambda x}\,dx + \lambda\int x\lambda e^{-\lambda x}\,dx \\
&= \log\lambda\cdot 1 + \lambda\cdot E(X) \\
&= -\log\lambda + 1.
\end{aligned}
$$

We now find the relative entropy of two exponentially distributed functions f *with parameter* λ_1 *and* g *with parameter* λ_2.

$$
\begin{aligned}
H(f|g) &= \int f(x)\log\frac{f(x)}{g(x)}\,dx \\
&= \int f(x)[\log f(x) - \log g(x)]\,dx
\end{aligned}
$$

$$
\begin{aligned}
&= \int f(x) \log f(x)\, dx - \int f(x) \log g(x)\, dx \\
&= -H(X) - \int \lambda_1 e^{-\lambda_1 x} [\log \lambda_2 + \log e^{-\lambda_2 x}]\, dx \\
&= -(-\log \lambda_1 + 1) - \log \lambda_2 \int \lambda_1 e^{-\lambda_1 x}\, dx + \int \lambda_2 x \lambda_1 e^{-\lambda_1 x}\, dx \\
&= \log \lambda_1 - 1 - \log \lambda_2 \cdot 1 + \lambda_2 \int x \lambda_1 e^{-\lambda_1 x}\, dx \\
&= \log \lambda_1 - 1 - \log \lambda_2 + \lambda_2 \cdot E(X) \\
&= \log \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} - 1,
\end{aligned}
$$

where $E(X) = \int x \lambda_1 e^{-\lambda_1 x}\, dx = \frac{1}{\lambda_1}$.

Now, we extend the definition of the relative entropy between two probability measures.

Definition 3.3.4. *Let* $(X, \mathfrak{X})$ *be a measurable space. Let* $P(X)$ *denote the set of all probability measures on* $\mathfrak{X}$ *and* $\mathcal{P}(\mathfrak{Y})$ *denote set of all finite* $\mathfrak{Y}$*-measurable partitions of* X*, where* $\mathfrak{Y}$ *is a* σ*-subalgebra of* $\mathfrak{X}$*. Let* $\mu, \nu \in P(X)$*. Then, the relative entropy of* μ *with respect to* ν *relative to* $\mathfrak{Y}$ *is defined by*

$$
H_{\mathfrak{Y}}(\mu|\nu) = \sup \left\{ \sum_{A \in \mathfrak{A}} \mu(A) \log \frac{\mu(A)}{\nu(A)} : \mathfrak{A} \in \mathcal{P}(\mathfrak{Y}) \right\}.
$$

If $\mathfrak{Y} = \mathfrak{X}$*, then we write* $H_{\mathfrak{X}}(\mu|\nu) = H(\mu|\nu)$ *and call it the relative entropy of* μ *with respect to* ν*. Note that the relative entropy is defined for any pair of probability measures.*

If $\mu \ll \nu$, μ is absolutely continuous with respect to ν, relative entropy has an integral form. That is,

$$
H(\mu|\nu) = \int_X \left(\frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} \right) d\nu = \int_X \log \frac{d\mu}{d\nu} d\mu,
$$

and if μ is not absolutely continuous with respect to ν, then $H(\mu|\nu) = \infty$.

Definition 3.3.5. ***(Kullback-Leibler information)*** *Let* ξ *and* η *be real random variables on* $(X, \mathfrak{X}, \mu)$*, so that they have probability distributions* μ_ξ *and* μ_η *given by*

$$
\mu_\xi(A) = \mu(\xi^{-1}(A)), \quad \mu_\eta(A) = \mu(\eta^{-1}(A)), \qquad A \in \mathfrak{B},
$$

respectively and $\mathfrak{B}$ is the Borel σ-algebra of $\mathbb{R}$. Suppose that μ_ξ and μ_η are absolutely continuous with respect to the Lebesgue measure dt of $\mathbb{R}$, so that we have the probability density functions f and g, respectively given by

$$f = \frac{d\mu_\xi}{dt}, \quad g = \frac{d\mu_\eta}{dt} \in L^1(\mathbb{R}) = L^1(\mathbb{R}, dt).$$

Then the Kullback-Leibler information between ξ and μ is given by

$$I(\xi|\mu) = \int_{\mathbb{R}} (f(t) \log f(t) - f(t) \log g(t))\, dt.$$

Observe that given the definition above and the integral form of $H(\mu_\xi|\mu_\eta)$, we have

$$\begin{aligned} I(\xi|\mu) &= \int_{\mathbb{R}} (f(t) \log f(t) - f(t) \log g(t))\, dt \\ &= \int_{\mathbb{R}} f(t) \log \frac{f(t)}{g(t)}\, dt \\ &= \int_{\mathbb{R}} \frac{d\mu_\xi}{dt} \log \frac{\frac{d\mu_\xi}{dt}}{\frac{d\mu_\eta}{dt}}\, dt \\ &= \int_{\mathbb{R}} \frac{d\mu_\xi}{d\mu_\eta} \log \frac{d\mu_\xi}{d\mu_\eta} d\mu_\eta \\ &= H(\mu_\xi|\mu_\eta). \end{aligned}$$

Therefore, the relative entropy, in general, is interpreted as the *Kullback-Leibler* information.

3.4 Birkhoff Pointwise Ergodic Theorem

So far, we have introduced a way of measuring the amount of information by calculating the entropy of a system. Now, we want a way of studying the long term average behavior of a system that evolves over time. In this section, we state and show the proof of a main theorem found in *Abstract Methods in Information Theory* by [Kak99]. This theorem will help us to describe that long term average behavior. We say that the collection of all states of the system form a space X, and the evolution is represented by a transformation $S : X \to X$. Since we want S to preserve the basic structure on X, we define S as a measure preserving transformation.

Definition 3.4.1. *Let S be a measure preserving transformation on a probability space $(X, \mathfrak{X}, \mu)$. The map S is said to be* ***ergodic*** *if for every measurable set A satisfying $S^{-1}A = A$, we have $\mu(A) = 0$ or $\mu(A) = 1$.*

The next theorem is a generalization of the Strong Law of Large Numbers. That is, if we have a sequence $X_1, X_2, ...$ of independent and identically distributed random variables with the expectation $E(X_i) = \mu$, then $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu$ almost surely. We say that an event is *almost surely* if the event occurs with probability one even if it does not contain all possible outcomes. The outcome or set of outcomes not contained in the event has probability zero [Dur05].

Theorem 3.4.1. ***Birkhoff Pointwise Ergodic Theorem*** *Let $(X, \mathfrak{X}, \mu)$ be a probability space and $S : X \to X$ a measure preserving transformation and $f \in L^1(X, \mu)$. Then, there exists a unique $f_S \in L^1(X, \mu)$ such that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(S^i x) = f_S$$

exits a.e., is S-invariant and $\int_X f d\mu = \int_X f_S \, d\mu$. If moreover S is ergodic, then f_S is a constant a.e. and $f_S = \int_X f d\mu$.

Proof. Let $f \in L^1(X, \mu)$ and without loss of generality consider $f \geq 0$. Define $f_n(x) = f(x) + \cdots + f(S^{n-1}x)$, $\bar{f} = \lim_{n\to\infty} \sup \frac{f_n}{n}$, and $\underline{f} = \lim_{n\to\infty} \inf \frac{f_n}{n}$. Then $\bar{f}$ and $\underline{f}$ are S-invariant. Observe that

$$\begin{aligned} \underline{f}(Sx) &= \lim_{n\to\infty} \inf \frac{f_n(Sx)}{n} \\ &= \lim_{n\to\infty} \inf \left[\frac{f_{n+1}(x)}{n+1} \cdot \frac{n+1}{n} - \frac{f(x)}{n} \right] \\ &= \lim_{n\to\infty} \inf \frac{f_{n+1}(x)}{n+1}. \end{aligned}$$

We show $\bar{f}$ is S invariant in the same manner. We next show that f_S exists, is integrable and S-invariant. It suffices to show that

$$\int_X \bar{f} \, d\mu \leq \int_X f \, d\mu \leq \int_X \underline{f} \, d\mu.$$

Since $\bar{f} - \underline{f} \geq 0$, this would imply that $\bar{f} = \underline{f} = f_S$ a.e.. Let $M > 0$ and $\epsilon > 0$ be fixed and

$$\bar{f}_M(x) = \min\{\bar{f}(x), M\}, \quad x \in X.$$

Define $n(x)$ to be the least integer $n \geq 1$ such that

$$\bar{f}_M(x) \leq \frac{f_n(x)}{n} + \epsilon = \frac{1}{n}\sum_{j=0}^{n-1} f(S^j x) + \epsilon, \quad x \in X.$$

Note that $n(x)$ is finite for each $x \in X$. Since $\bar{f}$ and $\bar{f}_M(x)$ are S-invariant, we have

$$n(x)\bar{f}_M(x) \leq n(x)[(\frac{f_{n(x)}}{n(x)}\bar{f}_M)(x) + \epsilon] = \sum_{j=0}^{n(x)-1} f(S^j x) + n(x)\epsilon, \quad x \in X. \tag{3.1}$$

Choose a large enough $N \geq 1$ such that

$$\mu(A) < \frac{\epsilon}{M} \quad \text{with} \quad A = \{x \in X : n(x) > N\}.$$

Now we define $\tilde{f}$ and $\tilde{n}$ by

$$\tilde{f}(x) = \begin{cases} f(x), & x \notin A \\ 0, & x \in A \end{cases}, \qquad \tilde{n}(x) = \begin{cases} n(x), & x \notin A \\ 1, & x \in A \end{cases}.$$

Then we see that for all $x \in X$

$$\tilde{n}(x) \leq N, \quad \text{by definition,}$$

$$\sum_{j=0}^{\tilde{n}(x)-1} \bar{f}_M(S^j x) \leq \sum_{j=0}^{\tilde{n}(x)-1} \tilde{f}(S^j x) + \tilde{n}(x)\epsilon, \tag{3.2}$$

by (3.1) and S-invariance of $\bar{f}_M$, and that

$$\begin{aligned} \int_X \tilde{f}\, d\mu &= \int_{A^c} f\, d\mu + \int_A \tilde{f}\, d\mu \\ &\leq \int_{A^c} f\, d\mu + \int_A f\, d\mu + \int_A M\, d\mu \\ &= \int_X f\, d\mu + \int_A M\, d\mu \leq \int_X f\, d\mu + \epsilon. \end{aligned} \tag{3.3}$$

Furthermore, find an integer $L \leq 1$ so that $\frac{NM}{L} < \epsilon$ and define a sequence $\{n_k(x)\}_{k=0}^{\infty}$ for each $x \in X$ by

$$n_0(x) = 0, \qquad n_k(x) = n_{k-1}(x) + \tilde{n}(S^{n_k(x)}x), \qquad k \geq 1.$$

Then it holds that for $x \in X$

$$\sum_{j=0}^{L-1} \bar{f}_M(S^j x) = \sum_{k=1}^{k(x)} \sum_{j=n_{k-1}(x)}^{n_k(x)-1} \bar{f}_M(S^j x) + \sum_{j=n_{k(x)}(x)}^{L-1} \bar{f}_M(S^j x),$$

where $k(x)$ is the largest integer $k \geq 1$ such that $n_k(x) \leq L-1$. Applying (3.2) to each of the $k(x)$ terms and estimating by M the last $L - n_{k(x)}(x)$ terms, we have

$$\begin{aligned}
\sum_{j=0}^{L-1} \bar{f}_M(S^j x) &= \sum_{k=1}^{k(x)} \sum_{j=n_{k-1}(x)}^{n_k(x)-1} \bar{f}_M(S^j x) + \sum_{j=n_{k(x)}(x)}^{L-1} \bar{f}_M(S^j x) \\
&\leq \sum_{k=1}^{k(x)} \left[\sum_{j=n_{k-1}(x)}^{n_k(x)-1} \tilde{f}(S^j x) + (n_k(x) - n_{k-1}(x))\epsilon \right] + (L - n_{k(x)}(x))M \\
&\leq \sum_{j=1}^{L-1} \tilde{f}(S^j x) + L\epsilon + (N-1)M
\end{aligned}$$

since $\tilde{f} \geq 0$, $\bar{f}_M \leq M$ and $L - n_{k(x)}(x) \leq N-1$. If we integrate both sides on X and divide by L, then we get

$$\int_X \bar{f}_M \, d\mu \leq \int_X \tilde{f} \, d\mu + \epsilon + \frac{(N-1)M}{L} \leq \int_X f \, d\mu + 3\epsilon$$

by the S-invariance of μ, (3.3) and $\frac{NM}{L} < \epsilon$. Thus, letting $\epsilon \to 0$ and $M \to \infty$ give the inequality $\int \bar{f} \, d\mu \leq \int_X f \, d\mu$. The other inequality $\int f \, d\mu \leq \int_X \underline{f} \, d\mu$ can be obtained similarly. Hence, $\bar{f} = \underline{f} = f_S$ a.e., and f_S is S-invariant. If S is ergodic, then S-invariance of f_S implies that f_S is a constant $a.e$ and

$$f_S(x) = \int_X f_S(y) \, d\mu(y) = \int_X f(y) \, d\mu(y). \quad \square$$

Chapter 4

Conclusion

In this thesis, we presented some topics in information theory and developed some examples as applications in different fields of study. As we developed this topics, we unraveled the importance of measure theoretic and functional analysis methods in order to define and characterize some of the properties in information theory.

In chapter one, to describe the amount of information of a source, we developed the theory of entropy. Specifically, we defined *Shannon entropy* for finite scheme and some of its properties. We showed some examples by finding the entropy of Bernoulli, Geometric, and Poisson probability distributions. We also defined the entropy function and provided a useful inequality in coding theory. In next chapter, we defined a dynamical system and a measurable partition to define the *Kolmogorov-Sinai* entropy. Next, we stated and proved the *Kolmogorov-Sinai theorem*. We defined and showed that *Bernoulli Shifts* with the same entropy are isomorphic. To finish, we calculated the entropy of *Markov Shifts* by also using *Kolmogorov-Sinai theorem*.

In the last chapter, we do not only define relative entropy for finite probability distributions and probability density functions, but also we extended this definition to an arbitrary pair of probability measures and then defined *Kullback-Leibler* information. We proved some essential properties of the relative entropy and provided an application in statistical hypothesis testing. Furthermore, we calculated the relative entropy for the Uniform, Normal, and Exponential distributions. Finally, we briefly developed the concept of ergodicity and proved one of the main results, *The Birkhoff Ergodic Theorem*.

Bibliography

[Ash67] Robert Ash. *Information Theory.* Interscience, New York, NY, 1967.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements Of Information Theory.* Wiley Interscience, Hoboken, NJ, 2006.

[Dur05] Richard Durrett. *Probability: Theory and Examples.* Curt Hinrichs, Belmont, CA, 2005.

[Kak99] Yuichiro Kakihara. *Abstract Methods In Information Theory.* Word Scientific, River Edge, NJ, 1999.

[Kul97] Solomon Kullback. *Information Theory And Statistics.* Dover, Mineola, New York, 1997.

[Pie80] John R. Pierce. *An Introduction to Information Theory: Symbols, Signals and Noise.* Dover, New York, NY, 1980.

[Rom92] Steven Roman. *Coding and Information Theory.* Springer-Verlag, New York, NY, 1992.

[Ros07] Sheldon M. Ross. *Introduction to Probability Models.* Academic Press, New York, NY, 2007.

[Shr04] Steven E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models.* Springer, New York, NY, 2004.

[Yeh06] James Yeh. *Real Anakysis: Theory of Measure and Integration.* World Scientific, Hackensack, NJ, 2006.