AN INTRODUCTION TO MATHEMATICAL ENTROPY

NISHKARSH SINGH

Abstract. Entropy is a concept with wide applications in all forms of sciences. In this article, we will be exploring the subject in a mathematical context, with inclination towards ergodic theory and measure theory. As a motivation, we first delve into various aspects of mathematical entropy and its definitions in terms of information theory. Subsequently, attention is directed towards entropy in the context of measure and probability spaces, we will also see how we can calculate the entropy measure preserving transformations on those measure spaces. We will then establish some interesting results and theorems such as the Kolmogorov-Sinai Theorem and Lochs' Theorem.

> . . . no one knows what entropy really is, so in a debate you will always have the advantage.
>
> *John von Neumann*

1. Introduction

Entropy is arguably the most misunderstood yet the most elegant topic in all of science. There are various versions of entropy, although they all follow the same theme. This implies that the core idea behind entropy is extremely versatile and general, which will be our motivation throughout this paper. There are many types of entropies, thermodynamic entropy, information theoretic entropy, metric entropy, topological entropy, etc.

Mathematics and Science were developed in order to understand and to some limit, quantify the workings of the universe. There were some things though, which were just not calculable; Some measurements were hindered by our lack of detailed information about the events, which were termed random or stochastic. Your true chances of winning the lottery or the angle at which a particle would move when I heat it, are mathematically random variables. Some information in the universe, is unattainable, due to varying factors, which gives rise to disorder, and uncertainty. The essence of entropy lies in trying to quantify this uncertainty at a larger scale, so it averages out or normalizes.

In this article, we will be majorly focusing on measure-theoretic entropy and Kolgomorov-Sinai entropy (KS entropy) which can be understood as generalizations of Information entropy (Shannon entropy). So we will be defining a few common terms in Shannon entropy, which will serve as a conceptual base and motivation for our main topic.

2. Information entropy

Information theoretic entropy was the first significant appearance of entropy in mathematics, Claude Shannon introduced this in his 1948 paper "A Mathematical Theory of Communication". Information entropy is vaguely defined as the average amount of 'information' or 'surprise' you get from the result of a probabilistic event for all possibilities.

Date: March 16, 2024.

Definition 2.1. *If there is a random variable X, which takes values from χ, and is distributed by the probability distribution $p : \chi \to [0,1]$, the* Shannon entropy $H[X]$ *is defined as :-*

$$H[X] \coloneqq -\sum_{x\in\chi} p[X=x]\log(p[X=x])$$

This formula might seem random for now, but let's see why this formula works as a metric for our definition. We need the average amount of information or surprise for all possibilities. So we know our formula should look something like

$$H[X] = \frac{1}{|\chi|}\sum_{x\in\chi}(\text{Information or Surprise if } X = x)$$

Where $|\chi|$ is the number of possibilities in the distribution. So let's try to define the information we would obtain for each outcome x. Let's call this $I(x)$, the information or surprisal of x. We would want some properties from $I(x)$,

(1) when the probability of something is higher, we want the surprisal to be less, and vice versa.
(2) It would make sense to have $I(x) = 0$, when $p[x] = 1$, and $I(x)$ to be undefined or ∞ when $p[x] = 0$.

$I(x) = \log(\frac{1}{p[x]})$ seems like the right choice for this as it satisfies all the properties we wanted. Now we want to average this out for all x; We know that we would choose x, $p[x]|\chi|$ times. So plugging this in we get :-

$$\begin{aligned} H[X] &= \frac{1}{|\chi|}\sum_{x\in\chi} p[X=x]|\chi|\log\left(\frac{1}{p[X=x]}\right) \\ &= \sum_{x\in\chi} p[X=x]|\log\left(\frac{1}{p[X=x]}\right) \\ &= -\sum_{x\in\chi} p[X=x]\log(p[X=x]) \end{aligned}$$

Note that we use $0\log(0) = 0$ as a convention, which makes sense as $\lim_{x\to 0} x\log(x) = 0$.

$H[X]$ can also be viewed as the expected value of the information by an event. It would be helpful to consider another type of entropy, namely conditional entropy. This is analogous to conditional probability, where we are calculating the entropy of event X, assuming that Y has happened.

Definition 2.2. *If $(X,Y) \sim p[X,Y]$, then the* conditional entropy $H[Y|X]$ *is defined as*

$$\begin{aligned} H[Y|X] &\coloneqq \sum_{x\in\chi} p[X=x]H[Y|X=x] \\ &= \sum_{x\in\chi} p[X=x]\sum_{y\in\chi} p[Y=y|X=x]\log(p[Y=y|X=x]) \\ &= \sum_{x\in\chi}\sum_{y\in\chi} p[(X,Y)=(x,y)]\log(p[Y=y|X=x]) \end{aligned}$$

Now that we are comfortable with the notion of mathematical entropy and information, we shall move closer to our real topic.

3. Partitions and measure-theoretic entropy

Now we will be seeing what entropy means in the context of measure spaces. First, we will define a partition.

Definition 3.1. *In a probability space $(X, \mathcal{A}, \mu)$, a* partition *is a pairwise disjoint collection of sets $\{A_i\}$ such that*

$$\bigcup_i A_i = X$$

Definition 3.2. *For a probability space $(X, \mathcal{A}, \mu)$ and partition $\mathscr{A}$, we define the* measure-theoretic (or metric) entropy *$H[\mathscr{A}]$ as*

$$H[\mathscr{A}] := -\sum_{A\in\mathscr{A}} \mu(A)\log\mu(A)$$

Definition 3.3. *For a probability space $(X, \mathcal{A}, \mu)$ and two partitions $\mathscr{A}$ and $\mathscr{B}$, the* Relative entropy *$H[\mathscr{A}|\mathscr{B}]$ is defined as*

$$H[\mathscr{A}|\mathscr{B}] := -\sum_{B\in\mathscr{B}} \mu(B) \sum_{A\in\mathscr{A}} \mu(A\cap B)\log\left(\frac{\mu(A\cap B)}{\mu(B)}\right)$$

We will now define an important operation which will be helpful when we will be talking about transformations.

Definition 3.4. *For any partitions $\mathscr{A} = \{A_1, A_2, \dots\}$ and $\mathscr{B} = \{B_1, B_2, \dots\}$ of a set X, the* join *of these partitions is defined as*

$$\mathscr{A} \vee \mathscr{B} = \{A_i \cap B_j\}$$

This can also be extended to any n partitions $\mathscr{A}_1, \mathscr{A}_1, \dots, \mathscr{A}_n$ such as the following,

$$\bigvee_{i=0}^{n} \mathscr{A}_i = \mathscr{A}_1 \vee \mathscr{A}_2 \dots \vee \mathscr{A}_n$$

Following are some easy to observe properties of entropy, whose proof we will skip over. Their proof can be found in almost all of the references.

Proposition 3.5. *Let $\mathscr{A}$,$\mathscr{B}$,$\mathscr{C}$ be partitions, then*

(1) $H[\mathscr{A} \vee \mathscr{B}|\mathscr{C}] = H[\mathscr{A}|\mathscr{C}] + H[\mathscr{B}|\mathscr{A} \vee \mathscr{C}]$

(2) *If we have $\mathscr{A} \leq \mathscr{B}$ ('$\leq$' here means that $\mathscr{B}$ is topologically finer than $\mathscr{A}$)*

$$H[\mathscr{A}|\mathscr{C}] \leq H[\mathscr{B}|\mathscr{C}]$$

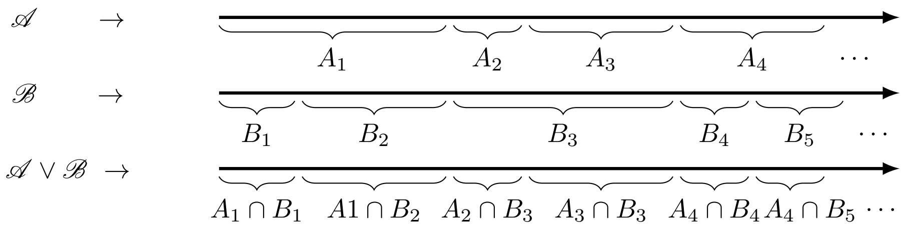
$$H[\mathscr{C}|\mathscr{A}] \leq H[\mathscr{C}|\mathscr{B}]$$

(3) $H[\mathscr{A} \vee \mathscr{B}|\mathscr{C}] \leq H[\mathscr{A}|\mathscr{C}] + H[\mathscr{B}|\mathscr{C}]$

(4) $H[\mathscr{A}|\mathscr{B}] = 0$ *if and only if* $\mathscr{A} \leq \mathscr{B}$

Measure-theoretic entropy can be seen as a generalization to the standard information entropy, it replaces probabilities of events with a more general measure. It's important to realize that the entropy here isn't directly uncertainty related to some *event* which can end up certain ways. It might not be trivial to understand what we mean by information of a partition. Suppose we need to *locate* a point $x \in X$, and we are given a partition $\mathscr{A} = \{A_1, A_2, \dots\}$. If we get to know that $x \in A_i$, is the information that we are obtaining, significant or vague? We take the expected value of this information, which is our entropy. This is the basic intuition behind this definition. Which means that the finer the partition, the greater the entropy, and vice versa. Even though this notion of uncertainty may seem very abstract and hard-to-digest for now, it will be justified when we see how transformations behave in this context.

It's easy to see that joining two partitions makes the resulting partition finer, because we contain only the intersection elements of both the partitions. Infact, it is the weakest(or coarsest) partition which is finer than both $\mathscr{A}$ and $\mathscr{B}$.



So as the joined partition is finer, it contains more information. Which means it has more entropy. In this context, we will be mostly joining only transformations of the same measurable subset. If we apply some transformation T on a set $\mathscr{A}$ repetitively and keep joining the sets, it will eventually stop refining the sets, we take the average of entropies at all these stages. This is what essentially the entropy of a transformation is.

Definition 3.6. *We define the* KS entropy $h[T, \mathscr{A}]$ *of a measure preserving transformation* T *and partition* $\mathscr{A}$ *as*

$$h[T, \mathscr{A}] = \lim_{n \to \infty} \frac{1}{n} H \left[\bigvee_{i=0}^{n-1} T^i(\mathscr{A}) \right]$$

We skip over the proof of the existence of this limit, which requires relatively elementary analysis.

Definition 3.7. *We define the* KS entropy $h[T]$ *of a transformation* T *as*

$$h[T] = \sup\left(h[T, \mathscr{A}]\right) \textit{ over all partitions } \mathscr{A}$$

$h[T, \mathscr{A}]$ can be seen as the average amount of information added to $\mathscr{A}$ by the transformation T, which makes perfect sense intuitively as that's what we are trying to quantify. We are now going to establish some basic propositions regarding these before proceeding to our theorems.

Proposition 3.8.
(1) *If* $\mathscr{A} \geq \mathscr{B}$, *then* $h[T, \mathscr{A}] \geq h[T, \mathscr{B}]$
(2) *For any* $n \geq 0$, $h[T, \mathscr{A}] = h[T, \bigvee_{i=0}^{n} T^{-j}(\mathscr{A})]$
(3) $h[T, \mathscr{A}] = \lim_{n \to \infty} h[T, \bigvee_{i=1}^{n} T^{-j}(\mathscr{A})]$

4. Kolgomorov-Sinai theorem

Kolgomorov-Sinai theorem is one of the most important and fundamental results about the entropy of measure-preserving transformations. It's not always trivial or obvious to find the supremum of $h[T, \mathscr{A}]$ every time as it may contain a large amount of possibilities. Kolgomorov-Sinai theorem gives us a solution to this problem and give us a partition, on which if we calculate the entropy of T, it will be equal to $h[T]$ everytime.

Definition 4.1. *In a invertible, measure-preserving system* $(X, \mathcal{A}, \mu, T)$, *a* T-generator *is a partition* $\mathscr{A}$ *such that* $\mathcal{A}$ *is generated by* $\bigvee_{i=-\infty}^{\infty} T^i(\mathscr{A})$, *i.e. -* $\bigvee_{i=-\infty}^{\infty} T^i(\mathscr{A}) = \mathcal{A}$ *up to a measure 0 with respect to* μ.

Theorem 4.2 (Kolgomorv-Sinai)**.** *If* $\mathscr{A}$ *is a* T*-generator, then we have*

$$h[T, \mathscr{A}] = h[T]$$

In order to prove this, we would need a few other theorems and some more background.

Lemma 4.3. *Let* A *be in the* σ*-algebra* $\mathscr{A} = \lim_{n \to \infty} \bigvee_{i=1}^{n} \mathscr{A}_i$. *Then for all* $\varepsilon > 0$, *there exists an* N, *such that for all* $n \geq N$, *there exists a union of atoms* A_n *(atoms are elements of the the partition) of* $\mathscr{A}_n$, *we have*

$$\mu(A_n \Delta A) \leq \varepsilon$$

Proof. We skip over the formal proof of this lemma, but this should be very intuitive, as $\mathscr{A}$ is infinitely finer than any $\mathscr{A}_n$. We are essentially approximating the element A, and as $\mathscr{A}_{n+1} \geq \mathscr{A}_n$ for all n, there must exist an N, which would be the *threshold* of fineness, after which we can approximate A using $\mathscr{A}_n$ upto ε. □

Theorem 4.4. *In partitions $\mathscr{A}_1 \leq \mathscr{A}_2 \leq \mathscr{A}_3 \leq \dots$, a partition $\mathscr{A}$ would only be contained in the σ-algebra $\bigvee_{i=1}^{\infty} \mathscr{A}_i$ if and only if*

$$\lim_{n \to \infty} H[\mathscr{A} | \mathscr{A}_n] = 0$$

Proof. Assume that $\mathscr{A}$ is contained in the σ-algebra $\vee_{i=1}^{\infty} \mathscr{A}$. Given $\varepsilon > 0$, set an N such that if $n \geq N$, and A_i is an atom of $\mathscr{A}$, then there exists an $A_i^{(n)} \in \mathscr{A}_n$ such that $\mu(A_i^{(n)} \Delta A_i) \leq \varepsilon$. Put

$$P_1^{(n)} := A_1^{(n)}$$

$$P_j^{(n)} := A_j^{(n)} \setminus \bigcup_{i=1}^{j-1} A_i \quad \text{for all } 1 \leq j \leq r-1$$

$$P_r^{(n)} := X \setminus \bigcup_{i=1}^{r-1} A_i$$

Then we have a partition $\mathscr{P}^{(n)} = \{P_1^{(n)}, P_2^{(n)}, \dots, P_r^{(n)}\}$, which for all $1 \leq i \leq r-1$ satisfies,

$$\begin{aligned}
\mu(P_i^{(n)} \Delta A_i) &\leq \mu(A_i^{(n)} \Delta A_i) + \mu(P_i^{(n)} \Delta A_i^{(n)}) \\
&\leq \varepsilon + \mu(P_i^{(n)} \Delta A_i^{(n)}) = \varepsilon + \mu(P_i^{(n)} \cap \bigcup_{j=0}^{i-1} A_j^{(n)}) \\
&\leq \varepsilon + \mu(P_i^{(n)} \cap A_i^{(n)}) = \varepsilon + \sum_{i=1}^{j-1} \mu(P_i^{(n)} \cap P_j^{(n)}) \\
&\leq \varepsilon + \sum_{i=1}^{j-1} \mu((P_i^{(n)} \setminus P_i) \cap (P_j^{(n)} \setminus P_j)) \\
&\leq \varepsilon + (i-1)\varepsilon = i\epsilon \\
&\leq r\epsilon
\end{aligned}$$

It's intuitively clear that $H[\mathscr{A} | \mathscr{P}^{(n)}] \leq \varepsilon$ because the log function inside the expression would converge to 0 by our last result; and by Proposition 3.4

$$H[\mathscr{A} | \mathscr{A}_n] \leq H[\mathscr{A} | \mathscr{P}^{(n)}] \leq \varepsilon$$

This is the proof from one side, we will skip the proof from the other side for now. A nice proof for the converse can be found in [4]. □

Corollary 4.5. *If $\mathscr{A}_1 \leq \mathscr{A}_1 \leq \dots$ is a sequence of finite partitions such that $\mathscr{P}$ is the σ-algebra $\bigvee_{i=1}^{\infty} \mathscr{A}_i$, then*

$$h[T] = \sup_n h[T, \mathscr{A}_n]$$

Proof. Let $\mathscr{A}$ be a finite partition. By the theorem and hypothesis that $\bigvee_{i=1}^{\infty} \mathscr{A}_i = \mathscr{P}$, we get

$$\lim_{n \to \infty} H[\mathscr{A}, \mathscr{A}_n] = 0$$

Thus

$$h[T, \mathscr{A}] - h[T, \mathscr{A}_n] \leq h[\mathscr{A} | \mathscr{A}_n]$$

So again,

$$h[T, \mathscr{A}_n] \geq h[T, \mathscr{A}] - h[\mathscr{A} | \mathscr{A}_n]$$

Hence we get,

$$\sup_n h[T, \mathscr{A}_n] = \sup_n h[T, \mathscr{A}] = h[T]$$

Hence proved. □

Now we finally have enough background to prove the Kolgomorov-Sinai Theorem. We will now give a proof to the Kolgomorov-Sinai Theorem.

Proof of Theorem 4.2. By the theorem

$$\begin{aligned} h[T] &= \sup_n h\left[T, \bigvee_{i=-n}^{n} T^{-i}(\mathscr{A})\right] \\ &= \sup_n h\left[T, T^n\left(\bigvee_{i=0}^{2n} T^{-i}(\mathscr{A})\right)\right] \\ &= \sup_n h\left[T, \bigvee_{i=0}^{2n} T^{-i}(\mathscr{A})\right] \\ &= h[T, \mathscr{A}] \end{aligned}$$

The last part is ensured by Proposition 3.8. □

Now that we are done with our first basic theorem of entropy, we can move on to further topics. Now we will be seeing a very famous theorem in ergodic theory, called the *Shannon-McMillan-Breiman Theorem.*

5. Shannon-McMillan-Breiman Theorem

Shannon-McMillan-Breiman Theorem is an important result in ergodic theory. This theorem gives us a new viewpoint of the entropy of a transformation on some partition. It tells us the size of the n^{th} join of $\mathscr{A}$ with it's transformations in terms of it's entropy.

Theorem 5.1 (Shannon-McMillan-Breiman)**.** *Let $(X, \mathcal{A}, \mu, T)$ be a measure-preserving probability system, where μ is ergodic under T, and $\mathscr{A}$ a partition with $H[\mathscr{A}] < \infty$. Let $\mathscr{A}_n = \bigvee_{i=0}^{n} T^i(\mathscr{A})$ for all $n \geq 1$, and $\mathscr{A}_n(x)$ be the atom of $\mathscr{A}_n$ containing x. Then we have*

$$-\lim_{n\to\infty} \frac{1}{n} \log(\mu(\mathscr{A}_n(x))) = h[T, \mathscr{A}]$$

As expected, we would be needing some more results and concepts to have enough background to prove this. First we start with the concept of conditional expectation.

Definition 5.2. *For a measure space $(X, \mathcal{A}, \mu)$, and a function $f \in \mathscr{L}^1$, $f : X \to \mathbb{R}$. The conditional expectation $\mathbb{E}_\mu(f|\mathcal{B})$ of a σ-algebra $\mathcal{B}$ where $\mathcal{B} \leq \mathcal{C}$ is defined as the unique function f' such that*

$$\int_{\mathcal{B}} f' \; d\mu = \int_{\mathcal{B}} f \; d\mu$$

Note that conditional expectation, regardless of it's name, is a function, and not a number.

Corollary 5.3. *For any atom A, we have*

$$f'(x) = \frac{1}{\mu(A)} \int_A f \; d\mu \quad \text{for all } x \in A$$

The finer the σ-algebra $\mathcal{B}$, the closer f' is to f. This is called the *Martingale Convergence Theorem.*

Theorem 5.4 (Martingale Convergence Theorem)**.** *Let $\{\mathcal{A}_n\}_{n\geq 1}$ be a sequence of σ-algebras, where $\mathcal{A}_{k+1}$ refines $\mathcal{A}_k$. Let $\mathcal{A} = \lim_{n\to\infty} \mathcal{A}_n \coloneqq \bigvee_{i=1}^{\infty} \mathcal{A}_i$, then for any $f \in \mathscr{L}^1$, we have*

$$\mathbb{E}_\mu(f|\mathcal{A}_n) \to \mathbb{E}_\mu(f|\mathcal{A}) \quad \text{as } n \to \infty$$

We skip over the proof. We will now take a look at the relative information function.

Definition 5.5. *Similar to relative entropy, we also define the relative information function $I_{\mathscr{A}|\mathscr{B}}$ as*

$$I_{\mathscr{A}|\mathscr{B}}(x) := -\sum_{A\in\mathscr{A}}\sum_{B\in\mathscr{B}} \mathbb{1}_{A\cap B}(x)\frac{\mu(A\cap B)}{\mu(A)}$$

Note that this is defined, such that

$$\int_X I_{\mathscr{A}|\mathscr{B}}(x)\ d\mu = H[\mathscr{A}|\mathscr{B}]$$

One can check using the previously given properties that

$$I_{\mathscr{A}\vee\mathscr{B}}(x) = I_{\mathscr{A}}(x) + I_{\mathscr{B}|\mathscr{A}}(x)$$

Corollary 5.6. *For any two partitions $\mathscr{A}$ and $\mathscr{B}$, we have*

$$-\log \mathbb{E}_\mu(\mathbb{1}_{\mathscr{A}(x)}|\mathscr{B}) = I_{\mathscr{A}|\mathscr{B}}(x)$$

Proof.

$$\begin{aligned}
-\log \mathbb{E}_\mu(\mathbb{1}_{\mathscr{A}(x)}|\mathscr{B}) &= -\log \mathbb{E}_\mu(\sum_{A\in\mathscr{A}} \mathbb{1}_A(x)|\mathscr{B}) \\
&= -\log \sum_{B\in\mathscr{B}} \frac{1}{\mu(B)} \int_B \sum_{A\in\mathscr{A}} \mathbb{1}_A(x) \\
&= -\log \sum_{A\in\mathscr{A}}\sum_{B\in\mathscr{B}} \mathbb{1}_{A\cap B}\frac{\mu(A\cap B)}{\mu(B)} \\
&= I_{\mathscr{A}|\mathscr{B}}(x)
\end{aligned}$$

□

Now we are ready to prove our theorem.

Proof of Theorem 5.1. Write $g_k(x) = I_{\mathscr{A}}(x)$ for $k = 1$ and $g_k(x) = I_{\mathscr{A}|\bigvee_{i=1}^{k-1} T^{-i}(\mathscr{A})}$ for $k \geq 2$, so we have

$$\begin{aligned}
I_{\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})}(x) &= I_{\bigvee_{i=1}^{n-1} T^{-i}(\mathscr{A})}(x) + I_{\mathscr{A}|\bigvee_{i=1}^{n-1} T^{-i}(\mathscr{A})}(x) \\
&= I_{\bigvee_{i=1}^{n-1} T^{-i}(\mathscr{A})}(x) + g_n(x) \\
&= I_{\bigvee_{i=0}^{n-2} T^{-i}(\mathscr{A})}(T(x)) + g_n(x) \\
&= I_{\mathscr{A}|\bigvee_{i=1}^{n-2} T^{-i}(\mathscr{A})}(x) + I_{\bigvee_{i=1}^{n-2} T^{-i}(\mathscr{A})}(T(x)) + g_n(x) \\
&= I_{\bigvee_{i=1}^{n-2} T^{-i}(\mathscr{A})}(T(x)) + g_{n-1}(x) + g_n(x) \\
&= I_{\bigvee_{i=0}^{n-3} T^{-i}(\mathscr{A})}(T^2(x)) + g_{n-1}(x) + g_n(x) \\
&\vdots \\
&= \sum_{i=0}^{n-1} g_{n-i}(T^i(x))
\end{aligned}$$

Now let $g = \lim_{n\to\infty} g_n$, which exists and belongs to $\mathscr{L}^1$ by the Martingale Convergence Theorem. We can write the equality as

$$\frac{1}{n} I_{\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})}(x) = \frac{1}{n}\sum_{i=0}^{n-1} g(T^i(x)) + \frac{1}{n}\sum_{i=0}^{n-1}(g_{n-i} - g)(T^i(x))$$

Since μ is ergodic, we can apply Birkhoff's Ergodic Theorem to get

$$\frac{1}{n} I_{\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})}(x) = \int_X g\ d\mu + \frac{1}{n}\sum_{i=0}^{n-1}(g_{n-i} - g)(T^i(x))$$

Now as we defined conditional information function, we have

$$\frac{1}{n}I_{\bigvee_{i=0}^{n-1}T^{-i}(\mathscr{A})}(x)=H\left[\mathscr{A}|\bigvee_{i=1}^{\infty}T^{-i}(\mathscr{A})\right]+\frac{1}{n}\sum_{i=0}^{n-1}(g_{n-i}-g)(T^{i}(x))$$

which is just

$$\frac{1}{n}I_{\bigvee_{i=0}^{n-1}T^{-i}(\mathscr{A})}(x)=h[T,\mathscr{A}]+\frac{1}{n}\sum_{i=0}^{n-1}(g_{n-i}-g)(T^{i}(x))$$

For the second sum, we will define

$$G_n=\sup_{k\geq N}|g_k-g| \text{ and } g^*=\sup_{n}g_n$$

We have that $0\leq G_n\leq g^*+g$ and $g^*+g\in\mathscr{L}^1$ because $\int_X g_n=H\left[\mathscr{A}|\bigvee_{i=1}^{n-1}\mathscr{A}\right]$ is a decreasing function as n grows. Also, moreover, $G_n\to 0$, so by the Dominated Convergence Theorem

$$\lim_{n\to\infty}\int_X G_n\ d\mu=\int_X\lim_{n\to\infty}G_n\ d\mu\to 0$$

now we can split the second sum into two parts

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}(g_{n-i}-g)(T^{i}(x))&=\frac{1}{n}\sum_{i=0}^{n-N-1}(g_{n-i}-g)(T^{i}(x)))+\frac{1}{n}\sum_{i=n-N}^{n-1}(g_{n-i}-g)(T^{i}(x))\\&\leq\frac{1}{n}\sum_{i=0}^{n-N-1}G_N(T^{i}(x)))+\frac{1}{n}\sum_{i=n-N}^{n-1}(g_{n-i}-g)(T^{i}(x))\end{aligned}$$

The second one clearly tends to 0, and the first one tends to zero by dominated convergence theorem when we take $N\to\infty$. Hence we have our desired result. □

6. Lochs' Theorem: An interesting application

Lochs' Theorem is an interesting and a very strong result in the study of continued fractions. Lochs' Theorem is an application of entropy and the Shannon-McMillan-Breiman Theorem. It gives us an explicit relation between the number of decimal places which are expressed by using a certain part of our continued fraction expansion. It is stated as,

Theorem 6.1. *For $\mathbb{R}\setminus\mathbb{Z}$, if $c(d)$ is the number of continued fraction terms needed to represent the first d digits of the decimal expansion of a number, then for almost all numbers in $\mathbb{R}\setminus\mathbb{Z}$, we have*

$$\lim_{d\to\infty}\frac{c(d)}{d}=\frac{6\log 2\log 10}{\pi^2}$$

The proof of this theorem majorly relies on the *Gauß* map, which is $G:x\to\frac{1}{x} \pmod 1$ and the map $T:x\to 10x \pmod 1$.

Definition 6.2. Gauß map *is the map $G:x\to\frac{1}{x} \pmod 1$, this transformation is measure preserving under the Gauß measure, which is*

$$\mu(A)=\frac{1}{\log 2}\int_A\frac{1}{x+1}\ dx$$

We also define the density of this map, as

$$\frac{d\mu(x)}{dx}=\frac{1}{\log 2}\frac{1}{x+1}$$

Proof. We know that $c(d)$ continued fraction terms represent, d decimal digits, this means that if $Z_c(x)$ is the continued fractions cylinder, and $Z_d(x)$ is the decimal cylinder, then $Z_c(x)$ under the Gauß map is contained under $Z_d(x)$ of the (Lebesgue measure preserving) $T : x \to 10x \pmod 1$ map, but not in the Z_{d+1} cylinder. We have

$$\frac{\log 2}{10}\lambda(Z_d(x)) \leq \mu(Z_c(x)) \leq 2\log 2\lambda(Z_d(x))$$

By the Shannon-McMillan-Breiman Theorem, we have

$$\begin{aligned}\frac{h_\lambda(T)}{h_\mu(G)} &= \lim_{d\to\infty} \frac{c(d)}{-\log\mu(Z_c(x))}\frac{-\log\lambda(Z_d(x))}{d}\\ &= \lim_{d\to\infty}\frac{c(d)}{d}\lim_{d\to\infty}\frac{\log\lambda(Z_d(x))}{\log\mu(Z_c(x))}\end{aligned}$$

Combining this with our first inequality, we get

$$\begin{aligned}\frac{h_\lambda(T)}{h_\mu(G)} &\leq \lim_{d\to\infty}\frac{c(d)}{d}\frac{d\log 10}{d\log 10 - \log(2\log 2)}\\ &= \lim_{d\to\infty}\frac{c(d)}{d}\left(1 + \frac{\log(2\log 2)}{d\log 10 - \log(2\log 2)}\right)\end{aligned}$$

by the same inequality, we get

$$\frac{h_\lambda(T)}{h_\mu(G)} \geq \lim_{d\to\infty}\frac{c(d)}{d}\left(1 - \frac{\log(\log 2)}{d\log 10 - 2\log 2}\right)$$

As these are sandwiching, we get the limit

$$\frac{h_\lambda(T)}{h_\mu(G)} = \lim_{d\to\infty}\frac{c(d)}{d}$$

The entropy $h_\lambda(T) = \log 10$ because the map T is isomorphic to the $\left(\frac{1}{10}, \frac{1}{10}, \ldots, \frac{1}{10}\right)$ bernoulli shift. The second entropy $h_\mu(G)$ is trickier to compute, because it requires something called the *Rokhlin Formula*, which says that for absolutely continous measures, we have

$$h_\mu(T) = \int_X \log|T'|\ d\mu$$

We already know the derivative (density) of the Gauß measure, which is $\frac{1}{\log 2}\frac{1}{x+1}$, plugging that in

$$h_\mu(T) = -\frac{2}{\log 2}\int_0^1 \frac{\log\frac{1}{x}}{x+1}\ dx$$

This integral is fairly easy to compute with elementary integration with parts, so are going to skip the steps

$$h_\mu(T) = -\frac{2}{\log 2}\frac{-\pi^2}{12} = \frac{\pi^2}{6\log 2}$$

Plugging this in our equation completes our proof. □

References

[1] Henk Bruin. Shannon-McMillan-Breiman Theorem. `https://www.mat.univie.ac.at/~bruin/ET1_lect1.pdf`, 2020.

[2] Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.

[3] A Ya Khinchin. *Mathematical foundations of information theory.* Courier Corporation, 2013.

[4] Ricardo Mañé. *Ergodic theory and differentiable dynamics*, volume 8. Springer Science & Business Media, 2012.

[5] Çagri Sert Manfred Einsiedler, Menny Akka Ginosar. Measure-Theoretic Entropy, Introduction. `https://metaphor.ethz.ch/x/2017/hs/401-3375-67L/sc/Chapter1.pdf`, 2017.

[6] Peter Walters. *An introduction to ergodic theory*, volume 79. Springer Science & Business Media, 2000.