Topics in Ergodic Theory

9. Entropy

Bernoulli Shift : Let $(P_1, P_2, \cdots, P_d)$ be a probability vector. The $(P_1, \cdots, P_d)$**-Bernoulli shift** is the MPS

$$\big(\{1, \cdots, d\}^{\mathbb{Z}}, \mathcal{B}, \mu_{P_1,\cdots,P_d}, \sigma\big)$$

where $\mu_{P_1,\cdots,P_d}$ is the product measure with $(P_1, \cdots, P_d)$ on the coordinates and σ is the shift map.

Definition) The MPS $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ are called **isomorphic**, if there are maps

$$S_1 : X_1 \to X_2, \quad S_2 : X_2 \to X_1$$

such that $S_1 \circ S_2 = id_{X_2}$ a.e., $S_2 \circ S_1 = id_{X_1}$ a.e., $(S_1)_*\mu_1 = \mu_2$, $(S_2)_*\mu_2 = \mu_1$ and $T_1 \circ S_2 = T_1 \circ T_2$ a.e.

The Question : Are the $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli-shift isomorphic?

This question does not seem very difficult, but this had been unsolved for a long time. These two shifts have "the same" Koopman operators, and moreover Meshalkin proved that $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$-Bernoulli shifts are isomorphic. This problem was finally solved by Kolmogorov. He proved that these two systems are not isomorphic by attaching a quantity called *entropy*, which should be preserved by isomorphism, on each system and by showing they are not equal. Later, Ornstein showed that two Bernoulli shifts are isomorphic if and only if they have the same entropy. Actually, the introduction of notion of entropy in measure preserving systems was the starting point of ergodic theory being identified as an independent subject, so the importance of entropy in the field of ergodic theory cannot be overemphasize.

Let us do some actual mathematics now. We define the entropy as the measure of the amount how difficult it is to predict the system.

Definition) Let $(X, \mathcal{B}, \mu)$ be a probability space. A **countable measurable partition** is a collection of measurable sets $A_1, A_2, \cdots$ such that $A_i \cap A_j = \phi$ for all $i \neq j$ and $\bigcup A_i = X$. The sets A_i are called the atoms of partition.

The **join** or **coarsest common refinement** of two countable measurable partiotion ξ, η is

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

Define a function

$$H(p_1, \cdots, p_d) = -\sum_{j=1}^{d} p_j \log(p_j)$$

for all probability vector $(p_1, \cdots, p_d)$ with the convention $0 \cdot \log 0 = 0$. The **entropy** of a countable measurable partition ξ is

$$H_\mu(\xi) = H(\mu(A_1), \mu(A_2), \cdots)$$

where $\xi = \{A_1, A_2, \cdots\}$.

The **conditional entropy** of $\xi = \{A_1, A_2, \cdots\}$ relative to $\eta = \{B_1, B_2, \cdots\}$ is

$$H_\mu(\xi|\eta) = \sum_{n=1}^{\infty} \mu(B_n) \cdot H\Big(\frac{\mu(A_1 \cap B_n)}{\mu(B_n)}, \frac{\mu(A_2 \cap B_n)}{\mu(B_n)}, \cdots\Big)$$

This is the average average of entropy conditioned on each partition of

⋆ Very Useful Interpretations :
Entropy of ξ provides the amount of information we can obtain from an experiment ξ. Conditional entropy of ξ relative to η provides the amount of information we can get from ξ given the information about experiment η. Entropy of join $\xi \vee \eta$ gives the amount of information we can get if we perform both experiments ξ and η.

Lemma)

(1) $H_\mu(\xi) \geq 0$.

(2) The value of $H_\mu(\xi)$ is maximal among partition ξ with k atoms if all atoms have the same measure $\frac{1}{k}$.

(3) $H_\mu(\{A_1, \cdots, A_k\}) = H_\mu(A_{\rho(1)}, \cdots, A_{\rho(k)})$ for all permutations $\rho \in \mathrm{Sym}(\{1, \cdots, k\})$.

(4) $H_\mu(\xi \vee \eta) = H_\mu(\xi) + H_\mu(\eta|\xi)$. This is called *chain rule.*

proof)

- (1) is trivial and (2) is going to be proved shortly using Jensen's inequality.
- (3) and (4) are going to be proved later, in more general setting.

Khinchin) Let $H_\cdot(\cdot) : P(X) \times \mathcal{B}$ be a function satisfying the conditions (1)-(4) of the lemma, where $P(X)$ is the set of Borel probability measures on $(X, \mathcal{B})$. Then $H_\cdot(\cdot)$ is uniquely determined by these properties, up to a multiplication of a scalar factor.

==========

(6th November, Tuesday)
Definition) A function $[a, b] \to \mathbb{R} \cup \{\infty\}$ is **convex**, if $\forall x \in (a, b)$, $\exists \alpha_x \in \mathbb{R}$ such that

$$f(y) \geq f(x) + \alpha_x(y - x) \quad \forall y \in [a, b]$$

f is **strictly convex** if the equality occurs only for $x = y$.

Remark : If f is $C^2([a, b])$ and $f''(x) > 0$ for all $x \in (a, b)$, then f is strictly convex.

Jensen's inequality) Let $f : [a, b] \to \mathbb{R} \cup \{\infty\}$ be a convex function. Let $p_1, p_2, \cdots$ be a probability vector (possibly countably infinite). Let $x_1, x_2, \cdots \in [a, b]$. Then

$$f(p_1x_1 + p_2x_2 + \cdots) \leq \sum_i p_i f(x_i)$$

If f is strictly convex, then equality occurs *iff* those x_i for which $p_i > 0$ coincide.

Claim : Let $(X, \mathcal{B}, \mu)$ be a probability space. Let ξ be a measurable partition with k atoms. Then

$$H_\mu(\xi) \leq \log(k)$$

and equality occurs only if each atom of ξ has measures $\frac{1}{k}$.

proof) Apply Jensen's inequality to the function $x \mapsto x \log(x)$ with weights $p_i = \frac{1}{k}$ at the point $\mu(A_i)$, where A_i are the atoms of ξ. Note,

$$\sum p_i \mu(A_i) = \frac{1}{k} \sum \mu(A_i) = \frac{1}{k}$$
$$\Rightarrow \quad \frac{1}{k} \log(\frac{1}{k}) \leq \sum \frac{1}{k} \mu(A_i) \log(\mu(A_i))$$

so

$$\log k \geq \sum (-1) \mu(A_i) \log \mu(A_i) = H_\mu(\xi)$$

(End of proof) □

Definition) Let $(X, \mathcal{B}, \mu)$ be a probability space and let ξ be a countable measurable partition. The **information function** of ξ is

$$\begin{aligned} I_\mu(\xi) : & X \to \mathbb{R} \cup \{\infty\} \\ & x \mapsto -\log \mu\big([x]_\xi\big) \end{aligned}$$

where $[x]_\xi$ is the atom of ξ where x belongs.

If η is another partition, then the **conditional information function** of ξ relative to x is

$$I_\mu(\xi|\eta)(x) = -\log \frac{\mu\big([x]_{\xi\vee\eta}\big)}{\mu\big([x]_\eta\big)}$$

It is apparent that the information function is related to entropy. This is summarized in the following lemma.

Lemma) With notation as above,

$$\begin{aligned} H_\mu(\xi) &= \int I_\mu(\xi) d\mu \\ H_\mu(\xi|\eta) &= \int I_\mu(\xi|\eta) d\mu \end{aligned}$$

proof) The first equality is direct from the definition. For the second equality,

$$\begin{aligned} I_\mu(\xi|\eta)d\mu &= \sum_{A\in\xi, B\in\eta} \int_{A\cap B} I_\mu(\xi|\eta)d\mu = -\sum_{A\in\xi, B\in\eta} \mu(A\cap B)\log\Big(\frac{\mu(A\cap B)}{\mu(B)}\Big) \\ &= -\sum_{B\in\eta}\mu(B)\cdot\sum_{A\in\xi}\frac{\mu(A\cap B)}{\mu(B)}\log\Big(\frac{\mu(A\cap B)}{\mu(B)}\Big) \end{aligned}$$

(End of proof) □

One reason we use information function is that it is much easier to prove chain rule with information function.

Lemma) *(Chain rule)* Let $(X, \mathcal{B}, \mu)$ be a probability space and let ξ, η, λ be countable measurable partitions. Then

$$\begin{aligned} I_\mu(\xi\vee\eta|\lambda)(x) &= I_\mu(\xi|\lambda)(x) + I_\mu(\eta|\xi\vee\lambda)(x) \quad \forall x \in X \\ H_\mu(\xi\vee\eta|\lambda) &= H_\mu(\xi|\lambda) + H_\mu(\eta|\xi\vee\lambda) \end{aligned}$$

proof) For the first equality,

$$\begin{aligned} I_\mu(\xi\vee\eta|\lambda)(x) &= \log\frac{\mu([x]_\lambda)}{\mu([x]_{\xi\vee\eta\vee\lambda})} \\ I_\mu(\xi|\lambda)(x) &= \log\frac{\mu([x]_\lambda)}{\mu([x]_{\xi\vee\lambda})} \\ I_\mu(\eta|\lambda\vee\xi)(x) &= \log\frac{\mu([x]_{\xi\vee\lambda})}{\mu([x]_{\xi\vee\eta\vee\lambda})} \end{aligned}$$

and this proves the chain rule for information function.

The second equality follows from the first equality by integration the information function (as in the previous lemma).

(End of proof) □

The following inequality is very important in theory of mathematics of information.

Lemma) Let notation be as above. Then

$$H_\mu(\xi|\eta) \geq H_\mu(\xi|\eta\vee\lambda)$$

"The amount of information obtained from ξ given η is larger than information obtained from ξ given η and λ."

proof)

$$H_\mu(\xi|\eta\vee\lambda)=\sum_{A\in\xi,B\in\eta,C\in\lambda}\mu(A\cap B\cap C)\log\Big(\frac{\mu(B\cap C)}{\mu(A\cap B\cap C)}\Big)$$
$$H_\mu(\xi|\eta)=\sum_{A\in\xi,B\in\eta}\mu(A\cap B)\log\Big(\frac{\mu(B)}{\mu(A\cap B)}\Big)$$

It is enough to show that for all fixed $A\in\xi$, $B\in\eta$, we have

$$\mu(A\cap B)\log\Big(\frac{\mu(B)}{\mu(A\cap B)}\Big)\geq\sum_{C\in\lambda}\mu(A\cap B\cap C)\log\Big(\frac{\mu(B\cap C)}{\mu(A\cap B\cap C)}\Big)$$

To see this, apply Jensen's inequality for $x\mapsto x\log x$ at points $\frac{\mu(A\cap B\cap C)}{\mu(B\cap C)}$ for $C\in\lambda$ with weights $\frac{\mu(B\cap C)}{\mu(B)}$. Write

$$\sum_{C\in\lambda}\frac{\mu(B\cap C)}{\mu(B)}\cdot\frac{\mu(A\cap B\cap C)}{\mu(B\cap C)}=\frac{1}{\mu(B)}\sum_{C\in\lambda}\mu(A\cap B\cap C)=\frac{\mu(A\cap B)}{\mu(B)}$$

and application of Jensen gives

$$\frac{\mu(A\cap B)}{\mu(B)}\cdot\log\Big(\frac{\mu(A\cap B)}{\mu(B)}\Big)\leq\sum_{C\in\lambda}\frac{\mu(B\cap C)}{\mu(B)}\cdot\log\Big(\frac{\mu(A\cap B\cap C)}{\mu(B\cap C)}\Big)$$

and therefore

$$\mu(A\cap B)\cdot\log\Big(\frac{\mu(A\cap B)}{\mu(B)}\Big)\leq\sum_{C\in\lambda}\mu(B\cap C)\cdot\log\Big(\frac{\mu(A\cap B\cap C)}{\mu(B\cap C)}\Big)$$

(End of proof) □

Corollary) $H_\mu(\xi)\leq H_\mu(\xi\vee\eta)\leq H_\mu(\xi)+H_\mu(\eta)$.

proof) Using the chain rule, obtain

$$H_\mu(\xi\vee\eta)=H_\mu(\xi)+H_\mu(\eta|\xi)$$

and form the previous lemma, has $H_\mu(\eta|\xi)\leq H_\mu(\eta)$

(End of proof) □

==========

(8th November, Thursday)

Lemma) Let $(X,\mathcal{B},\mu,T)$ be an MPS. Let ξ,η be countable measurable partitions. Then :

$$I_\mu(T^{-1}\xi|T^{-1}\eta)=I_\mu(\xi|\eta)(Tx)$$
$$H_\mu(T^{-1}\xi|T^{-1}\eta)=H_\mu(\xi|\eta)$$

where $T^{-1}\xi$ is the partition whose atoms are $T^{-1}([x]_\xi)$.

proof) Has

$$I_\mu(T^{-1}\xi|T^{-1}\eta)(x)=-\log\Big(\frac{\mu([x]_{T^{-1}\xi\vee T^{-1}\eta})}{\mu([x])_{T^{-1}\eta}}\Big)$$

Note

$$T^{-1}\xi\vee T^{-1}\eta=T^{-1}(\xi\vee\eta)\quad\text{and}\quad[x]_{T^{-1}\xi\vee T^{-1}\eta}=T^{-1}[Tx]_{\xi\vee\eta}$$

hence $\mu([x]_{T^{-1}\xi\vee T^{-1}\eta})=\mu([Tx]_{\xi\vee\eta})$ by the measure preserving property. Similarly $\mu([x]_{T^{-1}\eta})=\mu([Tx]_\eta)$. Then $I_\mu(T^{-1}\xi|T^{-1}\eta)=-\log\Big(\frac{\mu([x]_{T^{-1}\xi\vee T^{-1}\eta})}{\mu([x])_{T^{-1}\eta}}\Big)=I_\mu(\xi|\eta)(Tx)$

The statement on H_μ follows by integrating I_μ

(End of proof) □

Corollary) Writing $\xi_m^n = T^{-m}\xi \vee T^{-(m+1)}\xi \vee \cdots \vee T^{-n}\xi$, has

$$H_\mu(\xi_0^{n+m-1}) \leq H_\mu(\xi_0^{n-1}) + H_\mu(\xi_0^{m-1})$$

proof) Note that $\xi_0^{n+m-1} = \xi_0^{n-1} \vee \xi_n^{n+m-1}$. So we have

$$\begin{aligned} H_\mu(\xi_0^{n+m-1}) &\leq H_\mu(\xi_0^{n-1}) + H_\mu(\xi_n^{n+m-1}) = H_\mu(\xi_0^{n-1}) + H_\mu(T^{-n}\xi_0^{m-1}) \\ &= H_\mu(\xi_0^{n-1}) + H_\mu(\xi_0^{m-1}) \end{aligned}$$

where the last equality follows from the previous lemma.

(End of proof) □

Lemma) *(Felate's lemma)* Let $(a_n) \subset \mathbb{R}$ be a subadditive sequence, that is

$$a_{n+m} \leq a_n + a_m \quad \forall n, m$$

Then $\lim_{n\to\infty} a_n/n$ exists and equals $\inf_n a_n/n$.

proof sketch) Need to show that $\limsup_{n\to\infty} \frac{a_n}{n} \leq \frac{a_{n_0}}{n_0}$ for all n_0. For each fixed n_0, we can write $n = j(n)n_0 + i(n)$, where $i(n) \in [0, n_0 - 1]$. Iterate sub-additivity to get $a_n \leq j(n)a_{n_0} + a_{i(n)}$.

See the online note for the full proof.

Definition) Let $(X, \mathcal{B}, \mu, T)$ be an MPS. Let ξ, η be countable measurable partitions such that $H_\mu(\xi) < \infty$. The **entropy of the MPS w.r.t.** ξ is :

$$h_\mu(\xi) = \lim_{n\to\infty} \frac{H_\mu(\xi_0^{n-1})}{n} = \inf_n \frac{H_\mu(\xi_0^{n-1})}{n}$$

whose existence of the limit is guaranteed by *Felate's lemma.*(in fact, $\frac{H_\mu(\xi_0^{n-1})}{n}$ is a monotone decreasing sequence - will show in the example sheet)

The **entropy of the MPS** is $h_\mu(T) = \sup_{\xi: H_\mu(\xi)<\infty} h_\mu(T|\xi)$.

$h_\mu(\xi)$ expresses how fast we can learn information from a particular experiment ξ, and $h_\mu(T)$ is the maximal information we can obtain from the system when an appropriate experiment is chosen.

The problem of this definition is that it is generally difficult to find out the supremum $\sup_{\xi: H_\mu(\xi)<\infty} h_\mu(T|\xi)$ - since this requires computing entropy w.r.t ξ for each ξ. The good news is that (at least for the Bernoulli shifts), if we can find a partition that satisfies a particular property(so called **2-sided generator**), then in fact the supremum is achieved by the partition.

Definition) Let $(X, \mathcal{B}, \mu, T)$ be an invertible MPS. Let $\xi \subset \mathcal{B}$ be a countable measurable partitions. We say that ξ is a **2-sided generator** if $\forall A \in \mathcal{B}$ and $\forall \epsilon > 0$, $\exists k \in \mathbb{Z}_{>0}$ such that $\exists A' \in \sigma(\xi_{-k}^k)$ and $\mu(A \triangle A') < \epsilon$.

Theorem) *(Kolmogorov-Sinai)* Let $(X, \mathcal{B}, \mu, T)$ be an *invertible* measure preserving system. Let ξ be a countable measurable partition with $H_\mu(\xi) < \infty$, which is a 2-sided generator. Then

$$h_\mu(T) = h_\mu(T, \xi)$$

We delay the proof of this theorem until next lecture. Instead, we start to compute something useful.

Example : Let $(\{1, 2, \cdots, k\}^{\mathbb{Z}}, \mathcal{B}, \mu, \sigma)$ be the $(p_1, \cdots, p_k)$-Bernoulli shift. Let $X = \{1, 2, \cdots, k\}^{\mathbb{Z}}$.

- **Claim :** The partition $\xi = \{\{x \in X : x_0 = j\} : j = 1, \cdots, k\}$ is a 2-sided generator.

  proof) The collection of sets

  $$\{A \in \mathcal{B} : \forall \epsilon,\ \exists k\ \exists A' \in \xi_{-k}^k \text{ with } \mu(A \triangle A') < \epsilon\} \subset \sigma(\xi) \subset \mathcal{B}$$

  is a σ-algebra, and it contains cylinder sets. Hence it is equal to $\mathcal{B}$, as $\mathcal{B}$ is generated by cylinder sets.

  (End of proof) □

- **Claim :** With ξ defined as above, we have

$$H_\mu(\xi|\xi_1^n) = H(p_1, p_2, \cdots, p_k) = -p_1 \log p_1 - \cdots - p_k \log p_k$$

for all $n \in \mathbb{Z}_{\geq 0}$.

proof) Calculate the information function :

$$I_\mu(\xi|\xi_1^n)(x) = \log\Big(\frac{\mu([x]_{\xi_1^n})}{\mu([x]_{\xi_0^n})}\Big)$$

Note $[x]_{\xi_0^n} = \{y \in X : y_0 = x_0, \cdots, y_n = x_n\}$, so $\mu([x]_{\xi_0^n}) = p_{x_0} \cdots p_{x_n}$. Similarly, has $\mu([x]_{\xi_1^n}) = p_{x_1} \cdots p_{x_n}$, and

$$I_\mu(\xi|\xi_1^n)(x) = -\log p_{x_0}$$

therefore $H_\mu(\xi|\xi_1^n) = \sum_{j=1}^k p_j(-\log(p_j)) = H(p_1, \cdots, p_k)$.

(End of proof) $\square$

- Hence

$$\begin{aligned} H_\mu(\xi_1^{n-1}) &= H_\mu(\xi_{n-1}^{n-1}) + H_\mu(\xi_{n-2}^{n-2}|\xi_{n-1}^{n-1}) + H_\mu(\xi_{n-3}^{n-3}|\xi_{n-2}^{n-1}) + \cdots + H_\mu(\xi|\xi_1^{n-1}) \quad \text{(Chain rule)} \\ &= H_\mu(\xi) + H(\xi|\xi_1^1) + \cdots + H_\mu(\xi|\xi_1^{n-1}) \quad \text{(invariance, first lemma of the day)} \\ &= nH(p_1, \cdots, p_k) \end{aligned}$$

Divide by n and take the limit,

$$h_\mu(T) = h_\mu(T, \xi) = H(p_1, \cdots, p_k)$$

So the entropy of $(1/2, 1/2)$ shift is $\log 2$ and $(1/3, 1/3, 1/3)$ shift is $\log 3$ - which shows that two systems cannot be isomorphic.

==========

(10th November, Saturday)

Theorem) *(Kolmogorov-Sinai)* Let $(X, \mathcal{B}, \mu, T)$ be an *invertible* measure preserving system. Let ξ be a countable measurable partition with $H_\mu(\xi) < \infty$, which is a 2-sided generator. Then

$$h_\mu(T) = h_\mu(T, \xi)$$

We will need three lemmas.

Lemma 1) Let $(X, \mathcal{B}, \mu, T)$ be an *invertible* measure preserving system. Let $\xi \subset B$ be a countable partition. Then

$$h_\mu(T, \xi_{-n}^n) = h_\mu(T, \xi)$$

Lemma 2) Let $(X, \mathcal{B}, \mu, T)$ be an MPS. Let $\xi, \eta \subset \mathcal{B}$ be two countable partitions. Then

$$h_\mu(T, \eta) \leq h_\mu(T, \xi) + H_\mu(\eta|\xi)$$

Lemma 3) For any $\epsilon > 0$ and $k \in \mathbb{Z}_{>0}$, $\exists \delta > 0$ such that the following holds : let $(X, \mathcal{B}, \mu)$ be a probability space. Let $\xi \subset \mathcal{B}$ be a countable and $\eta \subset \mathcal{B}$ a finite partition. Suppose that η has k atoms and for each $A \in \eta$, $\exists B \in \sigma(\xi)$ such that $\mu(A \triangle B) < \delta$. Then

$$H_\mu(\eta|\xi) \leq \epsilon$$

Let us prove the theorem assuming the lemmas. We will come hack to the proof of the lemmas later.

proof of Theorem) We first show $h_\mu(T,\xi) = \sup_{\eta \text{ finite}} h_\mu(T,\eta)$. We need to show that for all finite partition $\eta \subset \mathcal{B}$, we have $h_\mu(T,\eta) \le h_\mu(T,\xi)$. Fix $\epsilon > 0$. By **Lemma 3** and definition of 2-sided generator, $\exists n$ such that $H_\mu(\eta|\xi_{-n}^n) \le \epsilon$. Then we have

$$\begin{aligned} h_\mu(T,\eta) &\le h_\mu(T,\xi_{-n}^n) + H_\mu(\eta|\xi_{-n}^n) \le\le h_\mu(T,\xi_{-n}^n) + \epsilon \quad \text{(Lemma 2)} \\ &= h_\mu(T,\xi) + \epsilon \quad \text{(Lemma 1)} \end{aligned}$$

We are done if we take $\epsilon \searrow 0$.

Now it is left to show that $\forall \epsilon > 0$, for every countable partition $\eta \subset \mathcal{B}$, $\exists \tilde{\eta} \subset \mathcal{B}$ finite such that $h_\mu(T,\eta) \le h_\mu(T,\tilde{\eta}) + \epsilon$. By **Lemma 2,** it is enough to show $H_\mu(\eta|\tilde{\eta}) \le \epsilon$. When $\eta = \{A_1, A_2, \cdots\}$, let $\tilde{\eta} = (A_1, A_2, \cdots, A_n, \bigcup_{j>n} A_j)$ for sufficiently large n . Then

$$H_\mu(\eta) = H_\mu(\eta \vee \tilde{\eta}) = H_\mu(\tilde{\eta}) + H_\mu(\eta|\tilde{\eta})$$

Thus

$$\begin{aligned} H_\mu(\eta|\tilde{\eta}) = H_\mu(\eta) - H_\mu(\tilde{\eta}) &= \sum_{j>n} (-1)\mu(A_j)\log\mu(A_j) + \mu\Big(\cup_{j=n+1}^\infty A_j\Big)\log\mu\Big(\bigcup_{j=n+1}^\infty A_j\Big) \\ &\le \sum_{j>n} (-1)\mu(A_j)\log\mu(A_j) \le \epsilon \end{aligned}$$

if n is sufficiently large.

(End of proof) □

Now we prove the lemmas.

proof of Lemma 1)

$$\begin{aligned} h_\mu(T,\xi_{-m}^m) &= \lim_{n\to\infty} \frac{1}{n} H_\mu(\xi_{-m}^{n+m-1}) = \lim_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n+2m-1}) \\ &= \lim_{n\to\infty} \frac{1}{n+2m-1} H_\mu(\xi_0^{n+2m-1}) = h_\mu(T,\xi) \end{aligned}$$

(End of proof) □

proof of Lemma 2)

$$\begin{aligned} h_\mu(T,\eta) &= \lim_{n\to\infty} \frac{1}{n} H_\mu(\eta_0^{n-1}) \\ &\le \lim_{n\to\infty} \frac{1}{n} H_\mu((\xi \vee \eta)_0^{n-1}) \\ &= \lim_{n\to\infty} \frac{1}{n}\Big(H_\mu(\xi_0^{n-1}) + H_\mu(\eta_0^{n-1}|\xi_0^{n-1})\Big) \\ &= h_\mu(T,\xi) + \lim_{n\to\infty} \frac{1}{n}\Big(\sum_{j=0}^{n-1} H_\mu(\eta_j^1|\xi_0^{n-1} \vee \eta_{j+1}^{n-1})\Big) \quad \text{(Chain rule)} \\ &= h_\mu(T,\xi) + \lim_{n\to\infty} \frac{1}{n}\Big(\sum_{j=0}^{n-1} H_\mu(\eta_j^j|\xi_j^j)\Big) \quad \text{(throwing away kwown information increases entropy)} \\ &= h_\mu(T,\xi) + H_\mu(\eta|\xi) \end{aligned}$$

where the last equality follows because $H_\mu(\eta_j^j|\xi_j^j) = H_\mu(\eta|\xi)$.

(End of proof) □

proof of Lemma 3) Write $A_1, \cdots, A_k$ for the atoms of η and for each $i \in \{1, \cdots, k\}$. Let $B_i \in \sigma(\xi)$ such that $\mu(A_i \triangle B_i) < \delta$. We consider the partition λ, which has the following $k+1$ atoms

: $C_0 = \bigcup_{i=1}^k (A_i \cap (B_i \setminus \bigcup_{j\neq i} B_j))$ and $C_i = A_i \setminus C_0$ for $i = 1, \cdots, k$.
They have some useful properties :

- C_0 is big, i.e. $\mu(C_0) \geq \sum_{i=1}^k (\mu(A_i) - k\delta) = 1 - k^2\delta$.
- If $x \in C_0$, then $x \in A_i \Leftrightarrow x \in B_i$.

Also note,

$$\begin{aligned} H_\mu(\lambda) &= -\mu(C_0)\log(\mu(C_0)) - \sum_{i=1}^k \mu(C_i)\log\mu(C_i) \\ &\leq -\mu(C_0)\log(\mu(C_0)) - \sum_{i=1}^n \mu(C_i)\log\Big(\frac{\sum_{i=1}^k \mu(C_i)}{k}\Big) \quad \text{(Jensen to } x \mapsto x\log x\text{)} \end{aligned}$$

If δ is sufficiently small, then $\mu(C_0)$ can be made as close to 1 as we want and $\sum_{i=1}^k \mu(C_i)$ is as small as we want. Then $H_\mu(\lambda) < \epsilon$ if δ is sufficiently small. Now, we may write

$$\begin{aligned} H_\mu(\eta|\xi) &\leq H_\mu(\eta \vee \lambda|\xi) \leq H_\mu(\lambda|\xi) + H_\mu(\eta|\xi \vee \lambda) \\ &\leq H_\mu + H_\mu(\eta|\xi \vee \lambda) < \epsilon + H_\mu(\eta|\xi \vee \lambda) \end{aligned}$$

so it is enough to show that

$$H_\mu(\eta|\xi \vee \lambda) = 0$$

However, this trivially holds because $\sigma(\eta) \subset \sigma(\xi \vee \lambda)$, i.e. $[x]_{\xi\vee\lambda} \subset [x]_\eta$.

: formally, this can be deduced from the fact

$$\begin{aligned} A_i &= C_i \cup (C_0 \cap A_i) = C_i \cup (A_i \cap (B_i \setminus \cup_{j\neq i} B_j)) \\ &= C_i \cup (C_0 \cap (B_i \setminus \cup_{j\neq i} B_j)) \in \sigma(\xi \vee \lambda) \end{aligned}$$

(End of proof) □

==========

(13th November, Tuesday)

Theorem) *(Shannon-McMillan-Breiman)* Let $(X, \mathcal{B}, \mu, T)$ be an ergodic MPS. Let $\xi \subset \mathcal{B}$ be a countable measurable partition. with $H\mu(\xi) < \infty$. Then

$$\frac{1}{n} I(\xi_0^{n-1}) \xrightarrow{n\to\infty} h_\mu(T, \xi)$$

pointwise μ-a.e. and in L^1.

Recall : $I(\xi_0^{n-1})(x) = -\log\mu([x]_{\xi_0^{n-1}})$, so get $\mu([x]_{\xi_0^{n-1}}) \approx \exp(-nh_\mu(T, \xi))$ approximately.

A lecture and a half would be devoted in proving this theorem.

Idea : Using the chain rule and then invariance,

$$\begin{aligned} \frac{1}{n} I_\mu(\xi_0^{n-1})(x) &= \frac{1}{n}\sum_{j=0}^{n-1} I_\mu(\xi_j^j|\xi_{j+1}^{n-1})(x) \\ &= \frac{1}{n}\sum_{j=0}^{n-1} I_\mu(\xi|\xi_1^{n-j-1})(T^j x) \end{aligned}$$

so this looks as if we can apply pointwise ergodic theorem. However, we still the notion of conditional entropy to develop this idea.

Conditional expectation

Let $(X, \mathcal{B}, \mu)$ be a probability space, and let $\mathcal{A} \subset \mathcal{B}$ be a sub-σ-algebra. The for all $f \in L^1(X, \mathcal{B}, \mu)$, $\exists f^* \in L^1(X, \mathcal{B}, \mu)$ such that

(1) f^* is $\mathcal{A}$-measurable.

(2) $\int_A f d\mu = \int_A f^* d\mu$ for all $A \in \mathcal{A}$.

If f_1^* and f_2^* both satisfy (1) and (2) in the role of f^*, then $f_1^* = f_2^*$ μ-a.e. The function f^* is called the **condition expectation** of f relative to $\mathcal{A}$ and it is denoted by $\mathbb{E}[f|\mathcal{A}]$.

Remarks :

- Writing $V_{\mathcal{A}}$ for the closed subspace of $L^2(X, \mathcal{B}, \mu)$ consisting of $\mathcal{A}$-measurable functions, $\mathbb{E}[f|\mathcal{A}]$ is the orthogonal projection of f to $V_{\mathcal{A}}$ provided $f \in L^2$.
- If $\mathcal{A}$ is generated by a countable partition ξ, then

$$\mathbb{E}[f|\mathcal{A}](x) = \mu([x]_\xi)^{-1} \int_{[x]_\xi} f d\mu$$

Theorem) Let $(X, \mathcal{B}, \mu)$ be the probability space, let $f, f_1, f_2 \in L^1(X, \mathcal{B}, \mu)$, $\mathcal{A}, \mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{B}$ be sub-σ-algebra. Then

(1) $\mathbb{E}[f_1 + f_2|\mathcal{A}] = \mathbb{E}[f_1|\mathcal{A}] + \mathbb{E}f_2|\mathcal{A}$.

(2) If f_1 is $\mathcal{A}$-measurable, then $\mathbb{E}[f_1 f|\mathcal{A}] = f_1 \mathbb{E}[f|\mathcal{A}]$.

(3) If $\mathcal{A}_1 \subset \mathcal{A}_2$ then $\mathbb{E}[f|\mathcal{A}_1] = \mathbb{E}[\mathbb{E}[\mathcal{A}|\mathcal{A}_2]|\mathcal{A}_1]$.

Theorem) *(Martingale theorems)* Let $(X, \mathcal{B}, \mu)$ be a probability space, let $f \in L^1$ and let $\mathcal{A}, \mathcal{A}_1, \mathcal{A}_2, \cdots \subset \mathcal{B}$ be sub-σ-algebras. Assume that *either* (1) $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots$ and $\mathcal{A} = \sigma(\mathcal{A}_j : j \geq 1)$ *or* (2) $\mathcal{A}_1 \supset \mathcal{A}_2 \supset \cdots$ and $\mathcal{A} = \cap \mathcal{A}_j$. Then

$$\lim_{n \to \infty} \mathbb{E}[f|\mathcal{A}_n] = \mathbb{E}[f|\mathcal{A}]$$

both pointwise μ-a.e. and in L^1.

If you have not seen these theorems but too lazy to go over the proofs, you can see these easily for $f \in L^2$ case.

Conditional information and entropy and entropy relative to σ-algebra

Let $(X, \mathcal{B}, \mu)$ be a probability space, let $\eta \subset \mathcal{B}$ be a countable partition let $\mathcal{A} \subset \mathcal{B}$ be a sub-σ-algebra. Then the **conditional information function of η relative to $\mathcal{A}$** is

$$I_\mu(\eta|\mathcal{A})(x) = \sum_{A \in \eta} -\chi_A \cdot \log \Big(\mathbb{E}[\chi_A|\mathcal{A}] \Big)$$

Example : Suppose that $\mathcal{A}$ is generated by a countable partition $\xi \subset \mathcal{B}$. Then

$$\begin{aligned} I_\mu(\eta|\mathcal{A})(x) &= \sum_{A \in \eta} -\chi_A \log \mathbb{E}[\chi_A|\mathcal{A}](x) \\ &= -\log \mathbb{E}[\chi_{[x]_\eta}|\mathcal{A}](x) \\ &= -\log \frac{1}{\mu([x]_\xi)} \int_{[x]_\xi} \chi_{[x]_\eta} d\mu \\ &= -\log \frac{\mu([x]_\eta \cap [x]_\xi)}{\mu([x]_\xi)} = -\log \frac{[x]_{\eta \vee \xi}}{\mu([x]_\xi)} \end{aligned}$$

Definition) The **conditional entropy of η relative to $\mathcal{A}$** is defined as $H_\mu(\eta|\mathcal{A}) = \int I_\mu(\eta|\mathcal{A}) d\mu$.

Theorem) *(Maximal inequality)* Let $(X, \mathcal{B}, \mu)$ be a probability space, and let $\xi_1, \xi_2, \cdots$ be countable measurable partitions such that $\sigma(\xi_1) \subset \sigma(\xi_2) \subset \cdots$. Let $\mathcal{A} = \sigma(\xi_j : j \geq 1)$. Let $\eta \subset \mathcal{B}$ be another countable partition. Then

$$I_\mu(\eta|\xi_n) \xrightarrow{n\to\infty} I_\mu(\eta|\mathcal{A}) \quad \text{pointwise a.e. and in } L^1$$

Moreover, I^* defined by $I^* = \sup_{n\in\mathbb{Z}_{>0}} I_\mu(\eta|\xi_n)$ is in $L^1(X, \mathcal{B}, \mu)$.

The theorem gives a useful tool to deal with $I_\mu(\eta|\mathcal{A})$ when $\mathcal{A}$ can be approximated by countable partitions.

==========

(15th November, Thursday)

(??? Why are we using the different version of Maximal inequality?)

Theorem) *(Maximal inequality)* Let $(X, \mathcal{B}, \mu)$ be a probability space. Let $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots$ be a sequence of sub-σ-algebras of $\mathcal{B}$, and let $\eta \subset \mathcal{B}$ be a countable partition with $H_\mu(\eta) > \infty$. Then :

$$I_\mu(\eta|\mathcal{A}_n) \to I_\mu(\eta|\mathcal{A}) \quad \text{as } n \to \infty$$

pointwise μ-a.e. and in L^1, where $\mathcal{A} = \sigma(\cup_n \mathcal{A}_n)$.
Moreover, $I^*(x) = \sup_{n\in\mathbb{Z}_{>0}} I_\mu(\eta|\mathcal{A})(x) \in L^1$.

proof for η finite) By definition and martingale convergence,

$$I_\mu(\eta|\mathcal{A}_n)(x) = -\log \mathbb{E}[\chi_{[x]_\eta}|\mathcal{A}_n](x) \xrightarrow{n\to\infty} -\log \mathbb{E}[\chi_{[x]_\eta}|\mathcal{A}](x) = I_\mu(\eta|\mathcal{A})(x)$$

This proves pointwise convergence. By dominated convergence, L^1 convergence follows if assuming $I^* \in L^1$.

Now it is enough to show $I^* \in L^1$

: fix $A \in \eta$, fix $\alpha > 0$. For $x \in X_1$, let $n(x)$ be defined by

$$n(x) = \min\{n : \log(\mathbb{E}[\chi_A|\mathcal{A}_n](x)) > \alpha\}$$

If the above does not hold for any n, the we write $n(x) = \infty$. Define $B_n = \{x \in X : n(x) = n\}$. Note that B_n is $\mathcal{A}_n$-measurable as we may write

$$B_n = \{x \in X : \mathbb{E}[\chi_A|\mathcal{A}_n](x) < \exp(-\alpha) \text{ but } \mathbb{E}[\chi_A|\mathcal{A}(x) \geq \exp(-\alpha)\ \forall j < n]\}$$

Then

$$\mu(B_n)\exp(-\alpha) \geq \int_{B_n} \mathbb{E}[\chi_A|\mathcal{A}_n]d\mu = \int_{B_n} \chi_A d\mu = \mu(B_n \cap A)$$

Define $A^* = \{x \in A : I^*(x) > \alpha\} \subset A \cap (\cup_n B_n)$. Then since B_n's are disjoint,

$$\mu(A^*) \leq \sum_{n=1}^{\infty} \mu(A \cap B_n) \leq \exp(-\alpha) \sum_{n=1}^{\infty} \mu(B_n) \leq \exp(-\alpha)$$

and summing these over all elements of η, we obtain

$$\mu\big(\{x \in X : I^*(x) > \alpha\}\big) \leq |\eta| \exp(-\alpha)$$

Then

$$\int I^*(x)d\mu \leq \sum_{n=1}^{\infty} \mu(x \in X : I^*(x) > n-1) \cdot n \leq \sum_{n=1}^{\infty} |\eta| \exp(-(n-1)) \cdot n < \infty$$

(End of proof) $\square$

Lemma) Let $(X, \mathcal{B}, \mu, T)$ be an MPS. Let $\xi \subset \mathcal{B}$ be a countable partition with $H_\mu(\xi) < \infty$. Then

$$h_\mu(T, \xi) = \lim_{n\to\infty} H_\mu(\xi|\xi_1^n)$$

proof) By chain rule and invariance of T,

$$\frac{1}{n}H_\mu(\xi_0^{n-1}) = \frac{1}{n}\sum_{j=0}^{n-1} H_\mu(\xi|\xi_1^j)$$

hence

$$h_\mu(T,\xi) = C\text{-}\lim_{n\to\infty} H_\mu(\xi|\xi_1^n)$$

Observe that $\mu(\xi|\xi_1^n)$ is a monotone decreasing sequence, hence it converges, so in fact the Cesàro limit implies strong limit.

(End of proof) □

We are now ready to prove Shannon-McMillan-Breiman theorem.

Theorem) Let $(X,\mathcal{B},\mu,T)$ be an ergodic MPS. Let $\xi \subset \mathcal{B}$ be a countable partition with $H_\mu(\xi) < \infty$. Then

$$\frac{1}{N}I_\mu(\xi_0^{N-1})(x) \to h_\mu(\xi,T)$$

pointwise a.e. and in L^1.

proof)

$$\begin{aligned}\frac{1}{N}I_\mu(\xi_0^{N-1})(x) =& \frac{1}{N}\sum_{n=0}^{N-1} I_\mu(\xi|\xi_1^{N-n-1})(T^n x)\\ &+ \frac{1}{N}\sum_{n=0}^{N-1} I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^n x)\\ &+ \frac{1}{N}\sum_{n=0}^{N-1}\Big(I_\mu(\xi|\xi_1^{N-n-1})(T^n x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty)(T^n x)\Big)\end{aligned}$$

where $\mathcal{B}(\xi_1^\infty)$ is the σ-algebra generated by $\bigcup_{n=1}^\infty \xi_1^n$.

By the pointwise ergodic theorem, has

$$\begin{aligned}\frac{1}{N}\sum_{n=0}^{N-1} I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^n x) \to& \int I_\mu(\xi|\mathcal{B}(\xi_1^\infty))d\mu\\ &= \lim_{n\to\infty}\int I_\mu(\xi|\xi_1^n)d\mu\\ &= \lim_{n\to\infty} H(\xi|\xi_1^n) = h_\mu(T,\xi)\end{aligned}$$

Define

$$I_K^*(x) = \sup_{k\geq K}\Big|I_\mu(\xi|\xi_1^k)(x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(x)\Big|$$

By the maximal inequality, $I_K^* \in L^1$ for all x and $I_K^*(x) \to 0$ for μ-a.e. x. I_K^* are pointwise monotone decreasing, so we have

$$\int I_K^* d\mu \to 0 \quad \text{as } K\to\infty$$

Now again by pointwise ergodic theorem,

$$\begin{aligned}&\Big|\frac{1}{N}\sum_{n=0}^{N-1}\Big(I_\mu(\xi|\xi_1^{N-n-1})(T^n x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty)(T^n x)\Big)\Big|\\ \leq& \frac{1}{N}\sum_{n=0}^{N-K-1} I_K^*(T^n x) + \frac{1}{N}\sum_{n=N-K}^{N-1} I_0^*(T^n x) \xrightarrow{N\to\infty} \int I_K^* d\mu\end{aligned}$$

as

$$\frac{1}{N}\sum_{n=N-K}^{N-1} I_0^*(T^n x) = \frac{1}{N}\sum_{n=0}^{N-1} I_0^*(T^n x) - \frac{N-K}{N}\frac{1}{N-K}\sum_{n=0}^{N-K-1} I_0^*(T^n x) \to \int I_0^* d\mu - \int I_0^* d\mu = 0$$

Hence,

$$\limsup_{N\to\infty}\Big|\frac{1}{N}\sum_{n=0}^{N-1}\Big(I_\mu(\xi|\xi_1^{N-n-1})(T^n x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty)(T^n x)\Big)\Big| \le \int I_K^* d\mu$$

Since K was arbitrary, and $\int I_K^* d\mu \xrightarrow{K\to\infty} = 0$, so

$$\limsup_{N\to\infty}\Big|\frac{1}{N}\sum_{n=0}^{N-1}\Big(I_\mu(\xi|\xi_1^{N-n-1})(T^n x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty)(T^n x)\Big)\Big| = 0$$

so we have pointwise convergence.

Moreover, if we observe carefully, we have L^1 convergence at each line of the proof, so we also have the L^1 convergence.

(End of proof) □