# 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 $\sigma$ 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 $\xi, \eta$ 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 $\xi$ 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

$\star$ Very Useful Interpretations :

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

**Lemma)**

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

(2) The value of $H_\mu(\xi)$ is maximal among partition $\xi$ 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 $H.(\cdot) : P(X) \times \mathcal{B}$ 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 $H.(\cdot)$ 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

$$f(p_1 x_1 + p_2 x_2 + \cdots) \leq \sum_i p_i f(x_i)$$

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 $\xi$ be a measurable partition with $k$ atoms. Then

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

and equality occurs only if each atom of $\xi$ 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 $\xi$. 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 $\xi$ be a countable measurable partition. The **information function** of $\xi$ 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 $\xi$ where $x$ belongs.

If $\eta$ is another partition, then the **conditional information function** of $\xi$ 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 $\xi, \eta, \lambda$ 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 $\xi$ given $\eta$ is larger than information obtained from $\xi$ given $\eta$ and $\lambda$."

**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 $\xi,\eta$ 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 $H_\mu$ follows by integrating $I_\mu$

*(End of proof)* □

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

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

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

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

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

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

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

**proof sketch)** Need to show that $\limsup_{n\to\infty} \frac{a_n}{n} \le \frac{a_{n_0}}{n_0}$ 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 $a_n \le j(n)a_{n_0} + a_{i(n)}$.

See the online note for the full proof.

**Definition)** Let $(X, \mathcal{B}, \mu, T)$ be an MPS. Let $\xi, \eta$ be countable measurable partitions such that $H_\mu(\xi) < \infty$. The **entropy of the MPS w.r.t.** $\xi$ 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 $\xi$, 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 $\xi$ for each $\xi$. 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 $\xi$ 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 $\xi$ 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 $\sigma$-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 $\xi$ 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)* □

- 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.