# Entropy

In this lecture we start discussing the notion of entropy. In dynamical systems, there is topological entropy (which is independent of invariant measures), and in ergodic theory there is measure-theoretic entropy, also misnamed as metric entropy. It depends on the choice of invariant measure $\mu$.

Topological and measure-theoretical entropy are related by the Variation Principle which say that

$$h_{top}(T) = \sup\{h_\mu(T) : \mu \text{ is } T\text{-invariant probability measure}\}$$

If $\mu$ is such that $h_\mu(T) = h_{top}(T)$, then $\mu$ is called a measure of maximal entropy.

We give some history first.

# Entropy

- Entropy was first used by 19th century physicist Rudolf Clausius and Ludwig Boltzmann as part of thermodynamics.
- Andrej Kolmogorov introduced a mathematical version in probability in the 1950s, and used it as isomorphism invariant. This is the way we still define it in ergodic theory.
- Around the same time, Claude Shannon used it in information theory.
- Major work in 1960s by Yakov Sinaĭ on measure-theoretic entropy and generators.
- Topological entropy was introduced in 1969 by Roy Adler, Alan Konheim and Harry McAndrew.
- In the early 1970s, Rufus Bowen and Efim Dinaburg independently introduced a user-friendlier version of topological entropy. Around this time, the Variational Principle was proved.
- In 1974 Don Ornstein published his theorem that entropy is complete invariant for two-sided Bernoulli shifts.

# Entropy

The current mathematical definition has very little to do anymore with the original definition from thermodynamics. Only it still expresses the amount of disorder in the system.

- For the circle rotation $(\mathbb{S}^1, \mathcal{B}, \mu, R_\alpha)$ with Lebesgue measure,

$$h_{top}(R_\alpha) = h_\mu(R_\alpha) = 0.$$

- For the doubling map $(\mathbb{S}^1, \mathcal{B}, \mu, T)$ with Lebesgue measure,

$$h_{top}(T) = h_\mu(T) = \log 2.$$

- For the Bernoulli shift $(\{1, \ldots, N\}^{\mathbb{N} \text{ or } \mathbb{Z}}, \mathcal{B}, \mu_p, \sigma)$ with probability vector $p = (p_1, \ldots, p_N)$,

$$h_{top}(\sigma) = \log N \geq h_{\mu_p}(\sigma) = -\sum_{i=1}^{N} p_i \log p_i.$$

# Jensen's Inequality

A function $f : \mathbb{R} \to \mathbb{R}$ is concave if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \tag{1}$$

for all $x, y \in \mathbb{R}$ and $\alpha \in [0,1]$.

If $f : \mathbb{R} \to \mathbb{R}$ is $C^2$ and $f'' \leq 0$, then $f$ is concave.

Theorem: For every strictly concave function $f : [0,\infty) \to \mathbb{R}$, and all $\alpha_i > 0$, $\sum_{i=1}^{n} \alpha_i = 1$ and $x_i \in [0,\infty)$ we have

$$\sum_{i=1}^{n} \alpha_i f(x_i) \leq f(\sum_{i=1}^{n} \alpha_i x_i), \tag{2}$$

with equality if and only if all the $x_i$ are the same.

# Jensen's Inequality

**Proof of Jensen's Inequality:** We prove this by induction on $n$. For $n = 2$ it is simply (1). So assume that (2) holds for some $n$. For $n+1$, take $\alpha_i > 0$ and $\sum_{i=1}^{n+1} \alpha_i = 1$ and write $B = \sum_{i=1}^{n} \alpha_i$.

$$
\begin{aligned}
f(\sum_{i=1}^{n+1} \alpha_i x_i) &= f(B \sum_{i=1}^{n} \frac{\alpha_i}{B} x_i + \alpha_{n+1} x_{n+1}) \\
&\geq Bf(\sum_{i=1}^{n} \frac{\alpha_i}{B} x_i) + \alpha_{n+1} f(x_{n+1}) \qquad \text{by (1)} \\
&\geq B \sum_{i=1}^{n} \frac{\alpha_i}{B} f(x_i) + \alpha_{n+1} f(x_{n+1}) \qquad \text{by (2) for } n \\
&= \sum_{i=1}^{n+1} \alpha_i f(x_i)
\end{aligned}
$$

Equality also carries over by induction. If $x_i$ are all equal for $i \leq n$, then (1) for $n+1$ is an equality only if $x_{n+1} = \sum_{i=1}^{n} \frac{\alpha_i}{B} x_i = x_1$.

# Jensen's Inequality

For measure-theoretic entropy, the function $\varphi : [0,1] \to \mathbb{R}$ defined as

$$\varphi(x) = -x \log x \quad \varphi(0) = \varphi(1) = 0$$

is important. Compute

$$\varphi'(x) = -1 - \log x, \qquad \varphi''(x) = -\frac{1}{x} < 0,$$

so $\varphi$ is concave.

By Jensen's Inequality (with all $\alpha_i = \frac{1}{N}$)

$$-\sum_{i=1}^{N} p_i \log p_i = N \sum_{i=1}^{N} \frac{1}{N} \varphi(p_i) \leq N\varphi(\frac{1}{N} \sum_{i=1}^{N} p_i) = \log N,$$

with equality if and only if all $p_i$ are the same.

# Entropy

Let $(X, \mathcal{B}, \mu)$ be a probability measure space. We call a collection

$$\mathcal{P} = \{P_i\} \qquad P_i \text{ measurable}$$

a (measurable) partition if all $P_i$'s are disjoint, and $X = \cup_i P_i$.

We say that a partition $\mathcal{P}$ is finer than $\mathcal{Q}$ (written as $\mathcal{P} \preceq \mathcal{Q}$) if every $P \in \mathcal{P}$is contained in some $Q \in \mathcal{Q}$. For example, the finest possible partition is when if $P_i = \{i\}$, $i \in X$; this is the point partition. It is (uncountably) infinite if $X$ is. The coarsest partition $\mathcal{P} = \{X\}$ is called the trivial partition.

Given two partitions $\mathcal{P}$ and $\mathcal{Q}$, the joint is

$$\mathcal{P} \vee \mathcal{Q} := \{P \cap Q : P \in \mathcal{P}, Q \in \mathcal{Q}\};$$

This joint is finer than both $\mathcal{P}$ and $\mathcal{Q}$.

# Entropy

If $T : X \to X$ is measurable, then the $n$-th joint of $\mathcal{P}$ is defined as

$$\mathcal{P}_n = \begin{cases} \bigvee_{k=-n}^{n-1} T^{-k}\mathcal{P}. & \text{if } T \text{ is invertible,} \\ \bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}. & \text{if } T \text{ is non-invertible,} \end{cases}$$

For example, if $T : \mathbb{S}^1 \to \mathbb{S}^1$ is the doubling map (non-invertible), and

$$\mathcal{P} = \{[0, \frac{1}{2}), [\frac{1}{2}, 1)\}$$

then

$$\mathcal{P}_n = \{[i/2^n, (i+1)/2^n), i = 0, \ldots, 2^n - 1\} \quad \text{so} \quad \#\mathcal{P}_n = 2^n.$$

A partition is generating for $T$ if for almost all $x \neq y \in X$, there is $n$ such that $x$ and $y$ lie in different elements of $\mathcal{P}_n$.

# Entropy

Given a finite partition $\mathcal{P}$ of a probability space $(X, \mu)$, let

$$H_\mu(\mathcal{P}) = \sum_{P \in \mathcal{P}} \varphi(\mu(P)) = -\sum_{P \in \mathcal{P}} \mu(P) \log(\mu(P)), \tag{3}$$

where we can ignore the partition elements with $\mu(P) = 0$ because $\varphi(0) = 0$. For a $T$-invariant probability measure $\mu$ on $(X, \mathcal{B}, T)$, and a partition $\mathcal{P}$, define the entropy of $\mu$ w.r.t. $\mathcal{P}$ as

$$h_\mu(T, \mathcal{P}) = \lim_{n \to \infty} \frac{1}{n} H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}). \tag{4}$$

Finally, the measure theoretic entropy of $\mu$ is

$$h_\mu(T) = \sup\{h_\mu(T, \mathcal{P}) \ : \ \mathcal{P} \text{ is a finite partition of } X\}. \tag{5}$$

# Fekete's Lemma

For this definition to make sense, we need to verify that the limit in (4) exists. For this we need:

Definition: We call a real sequence $(a_n)_{n\geq 1}$ subadditive if

$$a_{m+n} \leq a_m + a_n \quad \text{for all } m, n \in \mathbb{N}.$$

A positive sequence $(a_n)_{n\geq 1}$ submultiplicative if

$$a_{m+n} \leq a_m \cdot a_n \quad \text{for all } m, n \in \mathbb{N}.$$

NB: If $(a_n)$ is submultiplicative, then $(\log a_n)$ is subadditve.

Fekete's Lemma: If $(a_n)_{n\geq 1}$ is subadditive, then

$$\lim_n \frac{a_n}{n} = \inf_{r\geq 1} \frac{a_r}{r}.$$

# Fekete's Lemma

**Proof of Fekete's Lemma:** Every integer $n$ can be written uniquely as $n = p \cdot q + r$ for $0 \leq r < q$. Therefore

$$\limsup_{n\to\infty} \frac{a_n}{n} = \limsup_{p\to\infty} \frac{a_{p\cdot q+r}}{p\cdot q + r} \leq \limsup_{p\to\infty} \frac{pa_q + a_r}{p\cdot q + r} = \frac{a_q}{q}.$$

This holds for all $q \in \mathbb{N}$, so we obtain

$$\inf_q \frac{a_q}{q} \leq \liminf_n \frac{a_n}{n} \leq \limsup_n \frac{a_n}{n} \leq \inf_q \frac{a_q}{q},$$

as required.

# Fekete's Lemma

Call $a_n = H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$. Then (Prop. 13 of the Class Notes)

$$\begin{aligned} a_{m+n} &= H_\mu(\bigvee_{k=0}^{m+n-1} T^{-k}\mathcal{P}) \\ &\leq H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{P}) + H_\mu(\bigvee_{k=m}^{m+n-1} T^{-k}\mathcal{P}) \\ \text{by } T\text{-invariance of } \mu \quad &= H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{P}) + H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}) \\ &= a_m + a_n. \end{aligned}$$

Therefore $H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$ is subadditive, and the existence of the limit $h_\mu(T,\mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$ follows.

# Entropy

The second natural question about computing entropy:

*How can one possibly consider* **all** *partitions of $X$?*

By the next theorem, which we state without proof (see Theorem 22 in the Class Notes), we can reduce "all partitions" to "a single generating partition":

Theorem (Kolmogorov-Sinaĭ): Let $(X, \mathcal{B}, T, \mu)$ be a measure-preserving dynamical system. If partition $\mathcal{P}$ is such that

$$\begin{cases} \bigvee_{j=0}^{\infty} T^{-k}\mathcal{P} \text{ generates } \mathcal{B} & \text{if } T \text{ is non-invertible,} \\ \bigvee_{j=-\infty}^{\infty} T^{-k}\mathcal{P} \text{ generates } \mathcal{B} & \text{if } T \text{ is invertible,} \end{cases}$$

then $h_\mu(T) = h_\mu(T, \mathcal{P})$.

# Entropy

Now a good property of entropy:

Theorem: Two isomorphic measure preserving systems have the same entropy.

Indeed, let $(X, \mathcal{B}, T, \mu)$ and $(Y, \mathcal{C}, S, \nu)$ have full-measured sets $X' \subset X$, $Y' \subset Y$ and a bi-measurable invertible measure-preserving map $\phi : X' \to Y'$ such that

$$\begin{array}{ccc} (X', \mathcal{B}, \mu) & \xrightarrow{T} & (X', \mathcal{B}, \mu) \\ \phi \downarrow & & \downarrow \phi \\ (Y', \mathcal{C}, \nu) & \xrightarrow{S} & (Y', \mathcal{C}, \nu) \end{array}$$

commutes, then $h_\mu(T) = h_\nu(S)$.

This holds, because the bi-measurable measure-preserving map $\phi$ preserves all the quantities involved in (3)-(5), including the class of partitions for both systems.

# Entropy

For two-sided (i.e., invertible) Bernoulli shifts $(X = \{0, \ldots, N\}^{\mathbb{Z}}, \mathcal{B}, \mu_p)$ based on the probability vector $p = (p_1, \ldots, p_N\}$, the cylinder partition

$$\mathcal{P} = \{[i] : i = 0, \ldots, N\} \text{ is generating.}$$

Lemma: For the cylinder partition $\mathcal{P}$

$$h_{\mu_p}(\sigma, \mathcal{P}) = -\sum_i p_i \log p_i.$$

By the Kolmogorov-Sinaĭ Theorem, this generating partition suffices to compute the entropy.

Theorem (Ornstein 1974): Two two-sided Bernoulli shifts $(X, \mu_p, \sigma)$ and $(X', \mu_{p'}, \sigma)$ are isomorphic if and only if $h_{\mu_p}(\sigma) = h_{\mu_{p'}}(\sigma)$.

# Entropy

Let us now compute that $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}. = -\sum_i p_i \log p_i$, just for two symbols, and using the cylinder partition $\mathcal{P} = \{[0], [1]\}$, which can denote "head" and "tail" in coin-flips.

$$\mathcal{P}(k \text{ heads in } n \text{ flips}) = \binom{n}{k} p^k (1-p)^{n-k},$$

so by full probability:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Here $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ are the binomial coefficients, and

$$\begin{cases} k\binom{n}{k} = \frac{n!}{(k-1)!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1} \\ (n-k)\binom{n}{k} = \frac{n!}{(k)!(n-k-1)!} = n\frac{(n-1)!}{k!(n-k-1)!} = n\binom{n-1}{k} \end{cases} \tag{6}$$

# Entropy

We compute.

$$
\begin{aligned}
H_\mu\Big(\bigvee_{k=0}^{n-1}\sigma^{-k}\mathcal{P}\Big) &= -\sum_{x_0,\ldots,x_{n-1}=0}^{1}\mu([x_0,\ldots,x_{n-1}])\log\mu([x_0,\ldots,x_{n-1}]) \\
&= -\sum_{x_0,\ldots,x_{n-1}=0}^{1}\prod_{j=0}^{n-1}\rho(x_j)\log\prod_{j=0}^{n-1}\rho(x_j) \\
&= -\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\log\left(p^k(1-p)^{n-k}\right) \\
&= -\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}k\log p \\
&\qquad -\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}(n-k)\log(1-p)
\end{aligned}
$$

In the first sum, the term $k=0$ gives zero, as does the term $k=n$ for the second sum.

# Entropy

Thus we leave out these terms and rearrange by (6):

$$
\begin{aligned}
H_\mu\Big(\bigvee_{k=0}^{n-1}\sigma^{-k}\mathcal{P}\Big) &= -p\log p\sum_{k=1}^{n}k\binom{n-1}{k}p^{k-1}(1-p)^{n-k} \\
&\quad -(1-p)\log(1-p)\sum_{k=0}^{n-1}(n-k)\binom{n}{k}p^{k}(1-p)^{n-k-1} \\
&= -p\log p\sum_{k=1}^{n}n\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&\quad -(1-p)\log(1-p)\sum_{k=0}^{n-1}n\binom{n-1}{k}p^{k}(1-p)^{n-k-1} \\
&= n\left(-p\log p-(1-p)\log(1-p)\right).
\end{aligned}
$$

$\mathcal{P}$ is generating, so by the Kolomogorov-Sinaĭ Theorem,

$$
h_\mu(\sigma)=h_\mu(\sigma,\mathcal{P})=\lim_n\frac{1}{n}H_\mu\Big(\bigvee_{k=0}^{n-1}\sigma^{-k}\mathcal{P}\Big)=-p\log p-(1-p)\log(1-p).
$$