Lecture 3: Entropy

For each probability preserving map T on a probability space $(\Omega, \mathcal{F}, \mu)$ we will define a number $h_\mu(T)$ called the metric entropy (or Kolmogorov-Sinai entropy).

Then we will interpret it in terms of:

- the size of coarse grained phase space
- the information production rate
- the capacity of T to generate purely random signals

The metric entropy is defined using "dynamics of partitions", so we begin this.

Dynamics of Partitions

Def$^{\underline{n}}$. A finite measurable partition of a probability space $(\Omega, \mathcal{F}, \mu)$ is a collection $\alpha = \{A_1, \ldots, A_N\}$ s.t.:

- $A_1, \ldots, A_N$ are measurable sets (called the "atoms")
- $A_i \cap A_j = \emptyset$ for $i \neq j$
- $A_1 \cup \cdots \cup A_N = \Omega$

We let $\alpha(x) :=$ the element of α which contains x

Given two partitions $\alpha = \{A_1, \ldots, A_N\}, \beta = \{B_1, \ldots, B_M\}$

$$\alpha \vee \beta := \left\{ A_i \cap B_j \;\middle|\; \begin{matrix} i = 1, \ldots, N \\ j = 1, \ldots, M \end{matrix} \right\}$$

How to think about this object?

(1) <u>experiment</u> with N possible outcomes

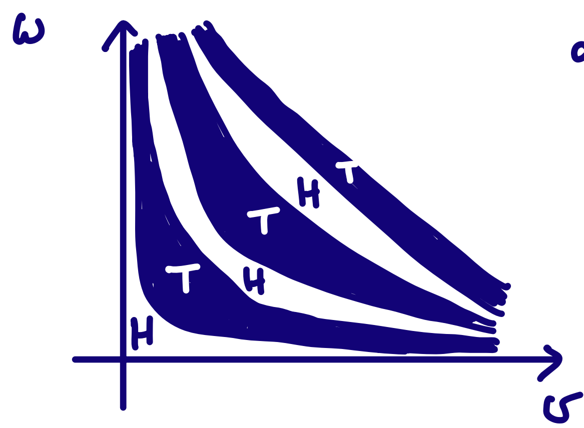
$$A_i = \{\omega \in \Omega : \text{ the experiment gives } i^{\text{th}} \text{ outcome}\}$$

$$\alpha(\omega) := \text{the outcome when the state is } \omega$$

For example, in lecture 1 we saw that when we toss a coin with $\begin{cases} \text{vertical velocity } \sigma \\ \text{angular velocity } \omega \end{cases}$, then

$$\alpha = \{\{(\sigma,\omega) : \text{"heads"}\}, \{(\sigma,\omega) : \text{"tails"}\}\}$$

is the partition



$$\alpha = \left\{ \begin{matrix}\text{union of} \\ \text{white} \\ \text{bands}\end{matrix}, \begin{matrix}\text{union of} \\ \text{blue} \\ \text{bands}\end{matrix} \right\}$$

Given two "experiments" α, β, if we carry out both at the same time, the possib outcomes are

$$(i,j) = \left(\begin{matrix} i^{\text{th}} \text{ outcome} \\ \text{in exp. } \alpha \end{matrix} ; \begin{matrix} j^{\text{th}} \text{ outcome} \\ \text{in exp. } \beta \end{matrix} \right)$$

and we obtain $\alpha \vee \beta$:

$$A_i \cap B_j = \left\{ \omega \in \Omega : \begin{matrix} \alpha \text{ exp. gives } i^{\text{th}} \text{ outcome} \\ \beta \text{ exp gives } j^{\text{th}} \text{ outcome} \end{matrix} \right\}$$

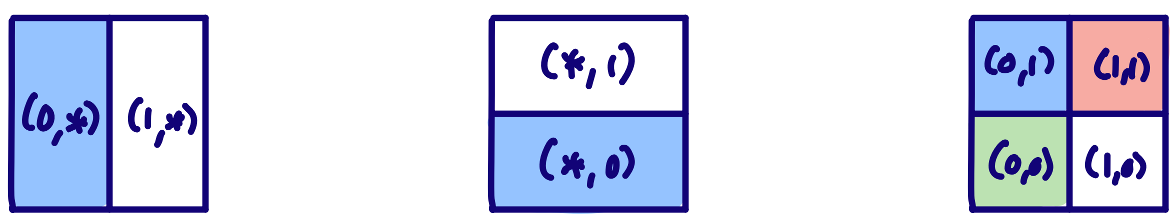
② <u>Information</u>: Imagine we don't know ω fully, we only know $\alpha(\omega)$ = element of α containing ω.

With this info

* <u>can</u> distinguish ω_1, ω_2 in different elements of α
* <u>cannot</u> —— " —— in the same elements of α.

<u>For example</u>: Suppose $\Omega = [0,1]^2$

- α = information on 1^{st} binary digit of x
- β = —— " —— 1^{st} —— " —— y



α $\qquad\qquad$ β $\qquad\qquad$ $\alpha \vee \beta$

③ <u>Coarse Graining</u>:

α = Coarse graining of $\boxed{\text{observers info}}$ on $\omega \in \Omega$

We're not coarse graining the system!

We're coarse graining the observer.

- bad for simulations
- good for theory (the dynamics stays the same)

A calculation: Suppose we repeat an "experiment" $\alpha = \{A_1, A_2, \dots\}$ at times $t = 0, 1, 2, \dots$.
What is the information we have at time n?

time	state	new information	accumulated information
0	ω	$\omega \in A_{i_0} := \alpha(\omega)$	$\omega \in A_{i_0}$
1	$T(\omega)$	$T(\omega) \in A_{i_1} := \alpha(T\omega)$	$\omega \in A_{i_0} \cap T^{-1}A_{i_1}$
2	$T^2(\omega)$	$T^2(\omega) \in A_{i_2} := \alpha(T^2\omega)$	$\omega \in A_{i_0} \cap T^{-1}A_{i_1} \cap T^{-2}A_{i_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	$T^{n-1}(\omega)$	$T^{n-1}(\omega) \in A_{i_{n-1}} := \alpha(T^{n-1}(\omega))$	$\omega \in \bigcap_{j=0}^{n-1} T^{-j}A_{i_j}$

Let $T^{-j}\alpha = \{T^{-j}A_i : A_i \in \alpha\}$, then

$$(T^{-j}\alpha)(\omega) = \alpha(T^j\omega) \qquad \text{(check!)}$$

The information at time n is coded by the partition

$$\alpha_n = \bigvee_{j=0}^{n-1} T^{-j}\alpha = \left\{ \bigcap_{j=0}^{n-1} T^{-j}A_{i_j} : A_{i_0}, \dots, A_{i_{n-1}} \in \alpha \right\}.$$

Notice that the information grows
the cardinality of the partition grows
but the elements —— " —— decrease

<u>Example</u>: $T : [0,1] \to [0,1]$, $T(\omega) = 2\omega \pmod 1$.
In binary expansions, $2 \times 0.\omega_1 \omega_2 \cdots = \omega_1 . \omega_2 \omega_3 \cdots$ so

$$T(0.\omega_1 \omega_2 \omega_3 \dots) = 0.\omega_2 \omega_3 \cdots .$$

Suppose we can only measure the first binary digit:

$$\alpha = \{ [0, 1/2) ; [1/2, 1] \}$$

time	state	new info	total info
0	$\omega = 0.\omega_1 \omega_2 \cdots$	ω_1	ω_1
1	$T(\omega) = 0.\omega_2 \omega_3 \cdots$	ω_2	ω_1, ω_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	$T^{n-1}(\omega) = 0.\omega_n \omega_{n+1} \cdots$	ω_{n-1}	$\omega_1, \dots, \omega_{n-1}$

Thus: $\alpha_n = \bigvee_{i=1}^{n-1} T^{-i} \alpha =$ partition into 2^n dyadic intervals of length $1/2^n$

The Entropy of a Partition

<u>Def^n</u>. Let $\alpha = \{A_1, \dots, A_N\}$ be a finite measurable partition of a probability space $(\Omega, \mathcal{F}, \mu)$.

$$H_\mu(\alpha) := -\sum_{A \in \alpha} \mu(A) \ln \mu(A)$$

$$h_\mu(T, \alpha) := \lim_{n \to \infty} \frac{1}{n} H_\mu(\alpha_n), \quad \alpha_n = \bigvee_{j=0}^{n-1} T^{-j} \alpha.$$

$$h_\mu(T) := \sup \left\{ h_\mu(T, \alpha) : \begin{array}{l} \alpha \text{ a finite measurable} \\ \text{partition of } \Omega \end{array} \right\}$$

<u>Thm (Krieger)</u>: If $h_\mu(T) < \infty$, then there exist finite partitions α s.t. $h_\mu(T, \alpha) = h_\mu(T) =$ maximal possible.

<u>Example</u> : Let

- $T(\omega) = 2\omega \bmod 1$ on $[0,1]$
- μ = length measure (Lebesgue measure)
- $\alpha = \{[0, \tfrac{1}{2}); [\tfrac{1}{2}, 1)\}$. Then:

α_n = n^{th} dyadic partition

$$H_\mu(\alpha_n) = \underset{\# \text{ intervals}}{2^n} \times \underbrace{\left(-\frac{1}{2^n} \ln \frac{1}{2^n}\right)}_{2^{-n} \cdot n \ln 2} = n \ln 2.$$

So $h_\mu(T, \alpha) = \lim_{n \to \infty} \frac{1}{n} H_\mu(\alpha_n) = \ln 2$.

It can be shown that $h_\mu(T, \alpha) = h_\mu(T)$.

Entropy and "the Volume of Coarse Grained Space"

Think of $\alpha_n = \{\alpha_n(\omega) : \omega \in \Omega\}$ as of a coarse-graining of the observer's information on Ω.

Shannon-McMillan-Breiman Thm: Let T be an ergodic probability preserving map on a probability space $(\Omega, \mathcal{F}, \mu)$. For every finite measurable partition α,

(1) for μ-a.e. $\omega \in \Omega$,

$$-\frac{1}{n} \log \mu(\alpha_n(\omega)) \xrightarrow[n\to\infty]{} h_\mu(T, \alpha)$$

(2) for every $\epsilon > 0$, for all n large enough we can decompose

$$\Omega = \Omega_{main} \cup \Omega_{negligible}$$

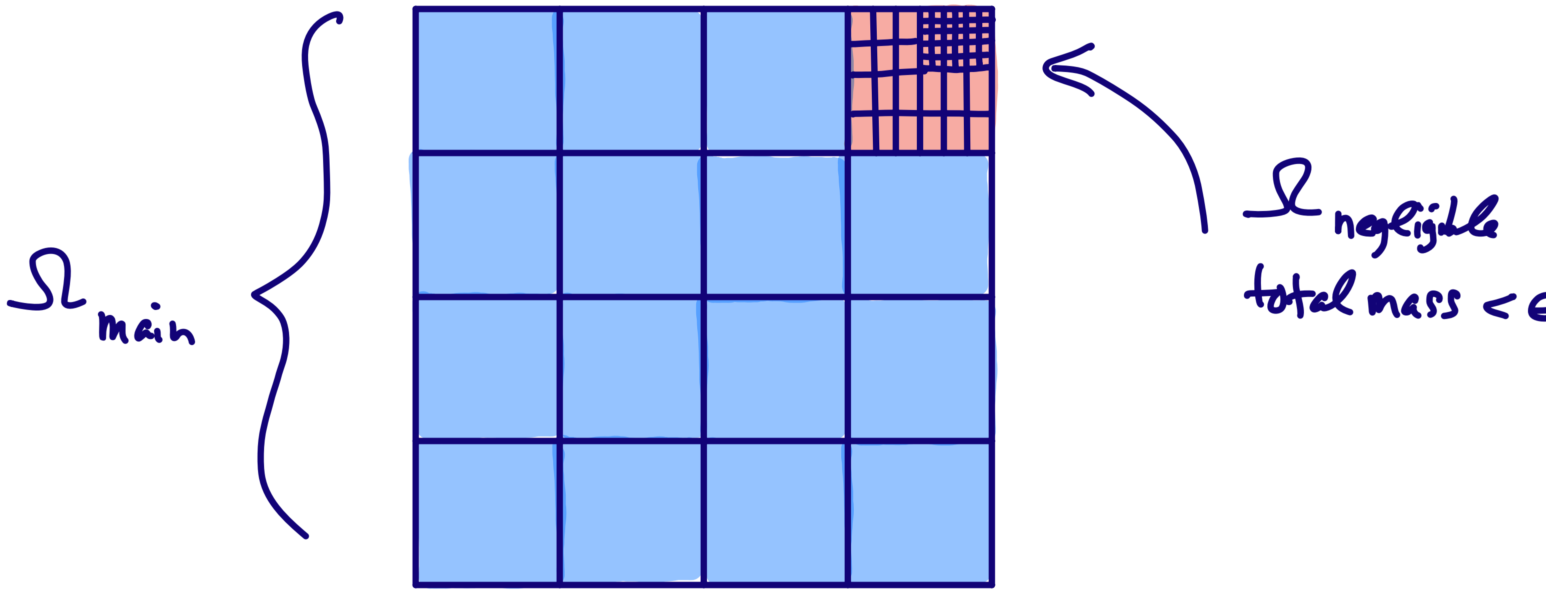
so that:

(a) $\mu(\Omega_{negligible}) < \epsilon$

(b) Ω_{main} = union of $e^{n(h_\mu(T,\alpha) \pm \epsilon)}$ elements of α_n, each of size $e^{-n(h_\mu(T,\alpha)\pm\epsilon)}$

Thus the cardinality of coarse grained space is approximately $\exp(n \cdot (h_\mu(T,\alpha) \pm \epsilon))$.

"Approximately" $=$ $\left(\begin{array}{l}\text{after removing a set} \\ \text{of measure } < \epsilon\end{array}\right)$



Entropy and Information Production Rate

We wish to quantify the information content $I_\mu(E)$ of the statement "ω belongs to E."

Axioms:

(a) $I_\mu(E) =$ decreasing function of $\mu(E)$

(b) $I_\mu(E) =$ continuous function of $\mu(E)$

(c) If A, B are independent, i.e. $\mu(A \cap B) = \mu(A)\mu(B)$, then $I_\mu(A \cap B) = I_\mu(A) + I_\mu(B)$.

All such functions are positively proportional to

$$I_\mu(E) := -\ln \mu(E)$$

(in computer science, people use $\log_2$).

Notice that

$$H_\mu(\alpha) = -\sum_{A\in\alpha} \mu(A) \ln \mu(A)$$

$$= \int_\Omega \sum_{A\in\alpha} 1_A(\omega) \cdot I_\mu(A) \, d\mu$$

$$= \int_\Omega I_\mu(\alpha)(\omega) \, d\mu$$

where $I_\mu(\alpha)(\omega) = I_\mu(\alpha(\omega))$. Observe:

- $I_\mu(\alpha)(\omega) =$ the information we gain in performing the experiment when the system is at state ω

- $H_\mu(\alpha) =$ $\boxed{\text{mean}}$ information we gain when performing the experiment (over all $\omega \in \Omega$)

If we repeat an experiment α at times $t=0,1,\dots$ then at time $n-1$ we know $\alpha_n(\omega)$, and the amount of information we have is $I_\mu(\alpha_n(\omega)) = -\log \mu(\alpha_n(\omega))$.

<u>Corollary</u>. Under the conditions of the Shannon–McMillan-Breiman theorem, for μ-a.e. ω

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} I_\mu(\alpha_n(\omega))}_{\text{"information production rate"}} = h_\mu(T,\alpha)$$

<u>Entropy and Unpredictability</u>

Define $H_\mu(\alpha|\beta) := H_\mu(\alpha \vee \beta) - H_\mu(\beta)$

= mean additional information in α given we already know β.

<u>Rokhlin Formula</u>: Under the assumptions of SMB thm, for every finite measurable partition α,

$$h_\mu(T,\alpha) = \lim_{n\to\infty} H_\mu(T^{-n}\alpha \mid \alpha_n)$$

learning $\alpha(T^n\omega)$

given that we already know $\alpha(\omega), \alpha(T\omega), \dots, \alpha(T^{n-1}\omega)$.

If $h_\mu(T,\alpha) > 0$, we still have much to learn, even at time $n \gg 1$!

Entropy and Deterministic Chaos

Deterministic Chaos: The ability of deterministic systems to produce behavior which appears random.

The "most random stochastic process":

Def$^{\underline{n}}$. Let $(p_1, \dots, p_N)$ be a probability vector, i.e. $0 \leq p_i \leq 1$ and $p_1 + \dots + p_N = 1$.
A Bernoulli process $B(p_1, \dots, p_N)$ is a bi-infinite sequence of independent random variables $\{X_i\}_{i \in \mathbb{Z}}$ s.t. $\text{Prob}(X_i = \xi) = p_\xi \quad (\xi = 1, \dots, N)$

Def$^{\underline{n}}$. Let T be an invertible map on a probability space. We say that T simulates a Bernoulli process $B(p_1, \dots, p_N)$, if $\exists$ measurable function

$$f: \Omega \longrightarrow \{1, \dots, N\}$$

s.t. $\{f \circ T^k\}_{k \in \mathbb{Z}}$ is a Bernoulli process $B(p_1, \dots, p_N)$

In other words, not knowing ω, we cannot distinguish using the methods of statistics alone the time series $\{f(T^n(\omega)) : n \in \mathbb{Z}\}$ from a purely random signal.

Recall:

<u>Def$^{\underline{n}}$</u>. The <u>metric entropy</u> of a probability preserving map T on a probability space $(\Omega, \mathcal{F}, \mu)$ is

$$h_\mu(T) = \sup\left\{ h_\mu(T,\alpha) : \begin{array}{l}\alpha \text{ finite} \\ \text{measurable partition}\end{array} \right\}$$

<u>Sinai's Factor Thm</u>. Let T be an invertible probability preserving map with entropy $h_\mu(T)$. Let $p = (p_1, \dots, p_N)$ be a probability vector.

(1) If $-\sum_i p_i \ln p_i \leq h_\mu(T)$, then T simulates a Bernoulli process $B(p_1, \dots, p_N)$

(2) If $-\sum_i p_i \ln p_i > h_\mu(T)$, then it doesn't.

In summary:

$$\left(\begin{array}{c}\text{positive metric} \\ \text{entropy}\end{array}\right) \iff \left(\begin{array}{c}\exists \text{ experiment } \alpha \\ \text{with completely} \\ \text{random looking} \\ \text{time series}\end{array}\right)$$

Take $\alpha = \{ \{\omega \in \Omega : f(\omega) = k\} \mid k = 1, \dots, N \}$