CHAPTER IV ENTROPY

IV.1 INFORMATION CONTENT

4.1 UNCERTAINTY

Let • $\xi = \{A_1, A_2, \dots, A_N\}$ be a measurable partition of $(X, \mathcal{B}, \mu)$

• $T: X \to X$ measure preserving

Let $\xi(x) = i :\Leftrightarrow x \in A_i$

Idea: $\xi(x)$ is all we can observe of the state.

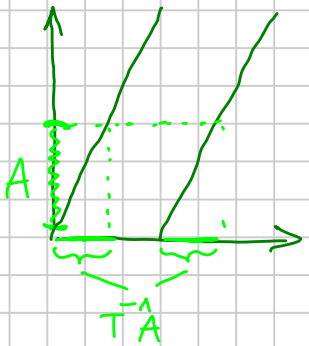
Q: Given $\xi(x), \xi(Tx), \dots, \xi(T^{m-1}x)$ ("itinerary"), how much uncertainty do we have regarding $\xi(T^m x)$? Or: how much "information content" do we gain by learning $\xi(T^m x)$?

Examples:

1) $X = S^1$, μ = Lebesgue, $T(x) = x + \frac{1}{100} \bmod 1$. $\xi = \{A_1, A_2\}$, $A_1 = [0, 1/2)$, $A_2 = [1/2, 1)$

Itinerary: $1, 1, 1, 1, 1$ ↷ If $x \sim \mu$, then the next partition element is A_1 with $\frac{45}{46} \approx 97\%$ probability

2) As 1), only now $T(x) = 2x \bmod 1$.



↷ Here, after any observed itinerary there is a 50% chance of observing both A_1 and A_2 as the next partition element

We would like to be able to describe the difference between examples 1) and 2) quantitatively.

4.2 MEASURABLE PARTITION

In the entire chapter: $(X, \mathcal{B}, \mu)$ — prob. space

- A (finite measurable) partition of $(X, \mathcal{B}, \mu)$, or, in shorthand, of X, is a finite collection $\xi = \{A_1, \dots, A_n\}$ of m'ble sets $A_i \in \mathcal{B}$, s.t.

  (i) the A_i are essentially disjoint: $\mu(A_i \cap A_j) = 0$, $i \neq j$;

  (ii) ξ is a covering: $\mu\left(\bigcup_{i=1}^{n} A_i\right) = 1$

- ξ' is a refinement of ξ, denoted by $\xi \leq \xi'$, if $\xi \subseteq \xi' \pmod 0$, i.e. up to differences of measure zero
- The common refinement of partitions ξ and η is defined as

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

- ξ and η are called independent, denoted by $\xi \perp \eta$, if

$$\mu(A \cap B) = \mu(A)\mu(B) \quad \forall A \in \xi, B \in \eta$$

Note the trivial one-to-one correspondence between finite partitions and finite σ-algebras.

4.3 ENTROPY OF A PARTITION

Suppose $Y \sim \mu$ is a random variable.

Q: How much information do we gain by learning $Y \in A$ for $A \in \mathcal{B}$?

↳ Call this $I(A)$

Natural requirements:

(i) $I(A) \geq 0$, and decreasing in A, i.e. $I(A) \geq I(B)$ if $A \subseteq B$; and $I(A) = 0$ if $\mu(A) = 1$

(ii) I continuous in A, i.e. $I(A_\varepsilon) \to I(A)$ for $A_\varepsilon \downarrow A$ or $A_\varepsilon \uparrow A$

(iii) If A, B independent, i.e. $\mu(A \cap B) = \mu(A)\mu(B)$, then $I(A \cap B) = I(A) + I(B)$

It turns out, the only function $\varphi:[0,1]\to\mathbb{R}_{\geq 0}$ such that $I(A)=\varphi(\mu(A))$ satisfies (i)–(iii) is $\varphi(t)=-c\log t$, $c\geq 0$.

Def: (a) The information content of $A\in\mathcal{B}$ is $I_\mu(A):=-\log\mu(A)$

(b) The information function of a finite measurable partition ξ is

$$I_\mu(\xi)(x):=\sum_{A\in\xi} I_\mu(A)\,\chi_A(x)=-\sum_{A\in\xi}\log\mu(A)\,\chi_A(x)$$

(c) The entropy of a finite measurable partition ξ is its average information content:

$$H_\mu(\xi):=\int I_\mu(\xi)\,d\mu=-\sum_{A\in\xi}\mu(A)\log\mu(A)$$

Convention: $0\cdot\log 0=0$

If μ is unambiguous, we will write $I(A)$, $I(\xi)$, and $H(\xi)$.

Prop: Let ξ,η be finite partitions. Then

(a) $H(\xi)\geq 0$, and $H(\xi)=0$ if and only if $\xi=\{X\}$

(b) If $\xi\leq\eta$, then $H(\xi)\leq H(\eta)$ with equation only if $\xi=\eta$

(c) If $|\xi|=n$ (n elements in ξ), then $H(\xi)\leq\log n$, equality only if $\mu(A)=1/n$ for every $A\in\xi$.

(d) $H(\xi\vee\eta)\leq H(\xi)+H(\eta)$ with equality only if $\xi\perp\eta$.

Useful Lemma: Let $\varphi(t):=-t\log t$, then for every prob. vector $(p_1,\dots,p_m)$ and $x_1,\dots,x_m\in[0,1]$ one has

$$\varphi\Big(\sum_i p_i x_i\Big)\geq\sum_i p_i\,\varphi(x_i),$$

with equality only if all the x_i where $p_i\neq 0$ are equal.

Pf: (a)–(c): Exercise.

(d) Shorthands: $\xi=\{A_1,\dots,A_m\}$, $\eta=\{B_1,\dots,B_m\}$

$$\mu_i=\mu(A_i),\quad \nu_i=\mu(B_i),\quad \kappa_{ij}=\mu(A_i\cap B_j),$$

hence $\sum_i\kappa_{ij}=\nu_j$, $\sum_j\kappa_{ij}=\mu_i$

Lemma (with $x_i = \frac{\kappa_{ij}}{\mu_i}$, $p_i = \mu_i$) gives

$$-\nu_j \log \nu_j \geq -\sum_i \mu_i \frac{\kappa_{ij}}{\mu_i} \log \frac{\kappa_{ij}}{\mu_i} = \sum_i \kappa_{ij} \log \mu_i - \sum_i \kappa_{ij} \log \kappa_{ij}$$

Sum up over j:

$$H(\eta) = -\sum_j \nu_j \log \nu_j \geq \sum_i \mu_i \log \mu_i - \sum_{ij} \kappa_{ij} \log \kappa_{ij}$$

$$= -H(\xi) + H(\xi \vee \eta)$$

Equality only if for fixed j, all the $x_i = \frac{\kappa_{ij}}{\mu_i}$ are the same (without loss: $\mu_i \neq 0 \ \forall i$), say $\frac{\kappa_{ij}}{\mu_i} = c_j$

$$\overset{\Sigma_i}{\Longrightarrow} \quad \underbrace{\sum_i \kappa_{ij}}_{\nu_j} = c_j \sum_i \mu_i = c_j \quad \Rightarrow \quad \kappa_{ij} = \mu_i \nu_j \ \forall i,j, \text{ which is exactly } \xi \perp \eta. \qquad \blacksquare$$

4.4 CONDITIONAL ENTROPY

ξ, η : finite partitions

Q: How much information do we gain by learning $Y \in A \in \xi$, if we know $Y \in B \in \eta$?

We already know $Y \in B \Rightarrow$ new probability space $(B, \mathcal{B}|_B, \mu_B)$, where $\mu_B(E) = \frac{\mu(E \cap B)}{\mu(B)}$. New information $Y \in A$ tells us $Y \in A \cap B$, hence the information content:

$$I_\mu(A \mid B) = -\log \frac{\mu(A \cap B)}{\mu(B)} =: -\log \mu(A|B)$$

(content)

If we are not told $B \in \eta$ yet, just η, the information gain by learning $Y \in A$ is going to depend on Y itself:

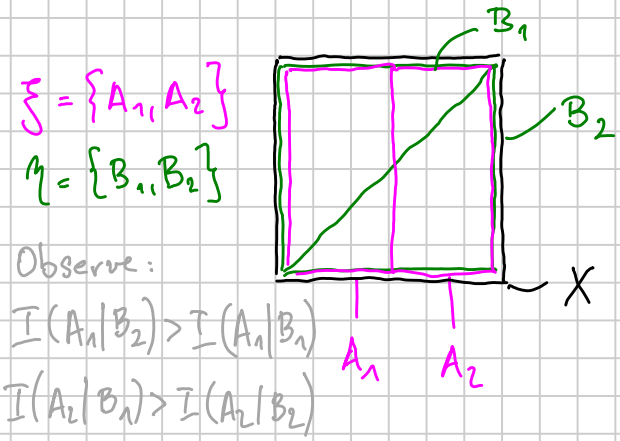
$$I_\mu(A \mid \eta)(x) = \sum_{B \in \eta} I_\mu(A|B) \chi_B(x) = -\sum_{B \in \eta} \log \mu(A|B) \chi_B(x) = -\log \mathbb{E}[\chi_A \mid \eta](x)$$

conditional expectation; recall the analogy between η and $\sigma(\eta)$

If, in addition, we are only told ξ, η, the information content is further going to depend on $\xi(x)$:

$$I_\mu(\xi|\eta) = \sum_{A\in\xi} I_\mu(A|\eta)\,\chi_A$$

"(conditional) information function"



Observe:
$I(A_1|B_2) > I(A_1|B_1)$
$I(A_2|B_1) > I(A_2|B_2)$

This motivates:

Def: The conditional entropy of ξ given η is the average information gain:

$$H_\mu(\xi|\eta) = \int I_\mu(\xi|\eta)\, d\mu$$

If $\xi(x) = A :\Leftrightarrow x \in A \in \xi$, we can write

$$H(\xi|\eta) = \int -\log \mu(\xi(x)|\eta(x))\, d\mu(x)$$

Prop: We have

$$H(\xi \vee \eta) = H(\eta) + H(\xi|\eta)$$

Proof:

$$\begin{aligned} H(\xi|\eta) &= \int -\log \mu(\xi(x)|\eta(x))\, d\mu(x) \\ &= \int -\log \mu(\xi(x)\cap\eta(x)) + \log \mu(\eta(x))\, d\mu(x) \\ &= \underbrace{\int -\log\mu(\xi(x)\cap\eta(x))\, d\mu(x)}_{= H(\xi\vee\eta)} - \underbrace{\int -\log\mu(\eta(x))\, d\mu(x)}_{= H(\eta)} \end{aligned}$$

∎

4.5 METRIC ENTROPY

Throughout: let $(X, \mathcal{B}, \mu, T)$ be a ppt.

Q: What is the information content in "$x \in A_0, Tx \in A_1, \dots, T^{n-1}x \in A_{n-1}$"?

If $A_0, \dots, A_{n-1} \in \xi$ then this is the same information as

$$x \in A_0 \cap T^{-1}A_1 \cap \dots \cap T^{-n+1}A_{n-1}$$

↝ The dynamics induces a partition

Let

- $T^{-1}\xi := \{T^{-1}A \mid A \in \xi\}$ — this is a measurable partition
- $\xi^{(n)} := \bigvee_{k=0}^{n-1} T^{-k}\xi = \xi \vee T^{-1}\xi \vee \dots \vee T^{-n+1}\xi$

Since T is a ppt, we have $H(T^{-k}\xi) = H(\xi) \quad \forall k \geq 0$

From Prop 4.3: $H(\xi \vee \eta) \leq H(\xi) + H(\eta)$

$$\Rightarrow \quad H(\xi^{(n+m)}) = H(T^{-n}\xi^{(m)} \vee \xi^{(n)}) \leq H(\xi^{(m)}) + H(\xi^{(n)})$$

Thus, the sequence $\{H(\xi^{(n)})\}_{n \in \mathbb{N}}$ is sub-additive.

Remark: & $f(0) \geq 0$
concave $\Rightarrow$ sub-additive
$\not\Leftarrow$
"$\Rightarrow$" draw or compute
"$\not\Leftarrow$" take $|\sin(x)|$

Lemma: For a sub-additive sequence $\{a_n\}_{n \in \mathbb{N}}$, i.e. where $a_{n+m} \leq a_n + a_m$ $\forall n, m \in \mathbb{N}$, one has

$$\lim_{n \to \infty} \frac{1}{n} a_n = \inf_n \frac{a_n}{n}$$

Pf: Fix n. For arbitrary $m \in \mathbb{N}$, write $m = kn + r$ with $r \in \{0, \dots, n-1\}$

Then $a_m \leq k a_n + a_r$ by sub-additivity, thus

$$\frac{a_m}{m} \leq \frac{k a_n + a_r}{kn + r} \leq \frac{a_n}{n} + \frac{a_r}{m}$$

$$\Rightarrow \limsup_{m \to \infty} \frac{a_m}{m} \leq \frac{a_n}{n} \quad \forall n \quad \Rightarrow \limsup_{m \to \infty} \frac{a_m}{m} \leq \inf_n \frac{a_n}{n}$$

Also, clearly, $\liminf_{m\to\infty} \frac{a_m}{m} \geq \inf_m \frac{a_m}{m}$

$\Rightarrow \lim_{m\to\infty} \frac{a_m}{m} = \inf_m \frac{a_m}{m}$ ∎

By this, it makes sense to speak of the average information content generated by T in one step:

Def: The metric (or measure-theoretic) entropy of T wrt the partition ξ is defined as

$$h(T,\xi) := \lim_{n\to\infty} \frac{1}{n} H(\xi^{(n)})$$

(Supposed $H(\xi) < \infty$.)

Recall Q: How much information does $\xi(T^n x)$ contain, given $\xi(x), \dots, \xi(T^{n-1}x)$?

Prop [BS, Prop 9.3.1]:

$$h(T,\xi) = \lim_{n\to\infty} H(\xi \mid T^{-1}\xi^{(n)}) = \lim_{n\to\infty} H(\xi \mid \bigvee_{i=1}^{n} T^{-i}\xi)$$

Proof: It holds $H(\xi|\zeta) \leq H(\xi|\eta)$ if $\eta \leq \zeta$, where ξ, η, ζ are finite partitions (exercise). Hence, $H(\xi \mid T^{-1}\xi^{(n)})$ is a non-increasing sequence in n.

$$H(\xi^{(n)}) = H(\xi \vee T^{-1}\xi^{(n-1)})$$

Prop 4.4

$$= H(T^{-1}\xi^{(n-1)}) + H(\xi \mid T^{-1}\xi^{(n-1)})$$

T pres. $\Rightarrow H(T^{-1}\eta) = H(\eta)$

$$= H(\xi^{(n-1)}) + H(\xi \mid T^{-1}\xi^{(n-1)})$$

$$= \dots = H(\xi) + \sum_{k=1}^{n-1} H(\xi \mid T^{-1}\xi^{(k)})$$

Divide by n, and let $n \to \infty$ to obtain the claim ∎

Def: The metric (or measure-theoretic) entropy of T is defined as

$$h(T) = \sup_{\xi} h(T,\xi),$$

where the sup is over all finite measurable partitions.

4.6 KOLMOGOROV-SINAI GENERATOR THEOREM

$d(\xi,\eta) := \min_{\sigma} \sum_{i=1}^{n} \mu(A_i \Delta B_{\sigma(i)})$ is a metric for finite measurable partitions

(σ: permutation)

$\xi = \{A_1,\dots,A_m\}$, $\eta = \{B_1,\dots,B_n,\phi,\dots,\phi\}$ — add empty sets until $|\xi| = |\eta|$

Def: (a) The sequence $\{\xi_n\}_n$ of finite partitions is refining if $\xi_n \le \xi_{n+1}$ $\forall n$

(b) The sequence $\{\xi_n\}_n$ of finite partitions is generating if for every finite partition η and $\delta > 0$ there is a $\tilde{\eta} \le \bigvee_{i=1}^{k} \xi_i$ for a sufficiently large n such that $d(\eta,\tilde{\eta}) < \delta$.

(c) A generator for a ppt T is a finite partition ξ such that $\{\xi^{(n)}\}_n$ is generating.

Note: Part (b) is a quantitative version of $\sigma\left(\bigcup_{n=1}^{\infty} \xi_n\right) = \mathcal{B} \bmod \mu$

Example: $\xi = \{[0,1/2), [1/2,1)\}$, $T(x) = 2x \bmod 1$.

ξ is a generator for T, since $\bigcup_{n=0}^{\infty} \bigvee_{i=0}^{n} T^{-i}\xi$ contains all dyadic intervals, hence $\sigma\left(\bigcup_{n=1}^{\infty} \xi^{(n)}\right) = \mathcal{B}$, the Borel σ-algebra (cf. §1.17)

Prop: If $\{\xi_n\}_n$ is a refining and generating sequence of partitions, then $h(T) = \lim_{n\to\infty} h(T,\xi_n)$. [BS, Prop. 9.3.4]

Theorem (Kolmogorov-Sinai generator theorem):

Let ξ be a generator for T. Then $h(T) = h(T,\xi)$. [BS, Thm 9.3.5]

4.7 THE SHANNON-MCMILLAN-BREIMAN THEOREM

Theorem: Let T be an ergodic ppt and ξ a finite partition. Then

$$\frac{1}{n} I_\mu(\xi^{(n)}) \xrightarrow[n\to\infty]{} h(T,\xi) \qquad \mu\text{-a.e. and in } L^1$$

[Sa, Thm. 4.3]

Consequence: Recalling Def 4.3 (information function), the theorem states that at a typical point $x \in X$, the size of the partition element $\xi^{(n)}(x)$, $\mu(\xi^{(n)}(x))$ decays as $\exp(-n\, h(T,\xi))$

4.8 EXAMPLES

1) **Circle rotation:** $T(x) = x + \alpha \bmod 1$ on S^1

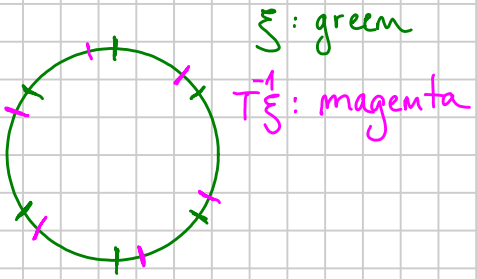
(i) $\alpha \in \mathbb{Q}$: $\exists k \in \mathbb{N}$ s.t. $T^{-k} = id$, hence $T^{-k}\xi = \xi$ for every partition ξ, and thus $\xi \leq \bigvee_{i=1}^{n} T^{-i}\xi \quad \forall n \geq k$.

But from Prop 4.4: $H(\xi \vee \eta) = H(\eta) + H(\xi|\eta)$, and if $\xi \leq \eta$, then $H(\xi|\eta) = 0$ follows.

$\Rightarrow H(\xi | \bigvee_{i=1}^{n} T^{-i}\xi) = 0 \quad \forall n \geq k \overset{\text{Prop 4.5}}{\Rightarrow} h(T,\xi) = 0$

$\Rightarrow h(T) = \sup_{\xi} h(T,\xi) = 0$

(ii) $\alpha \notin \mathbb{Q}$: ξ_N : partition of S^1 into N equal intervals



Since $\alpha \notin \mathbb{Q}$, endpoints of $T^{-i}\xi$ are mutually distinct

$\Rightarrow \xi^{(n)}$ consists of $n \cdot N$ intervals, and the endpoints of $\xi^{(0)}, \xi^{(1)}, \ldots$ are dense in S^1

$\Rightarrow \xi_N$ is generator (every interval can be arbitrary well approximated)

Prop 4.3(c): $H(\xi_N^{(n)}) \leq \log(nN) = \log n + \log N$

$\Rightarrow h(T) = h(T,\xi_N) = 0$

↑ Kolmogorov-Sinai thm

2) **Expanding maps:** $T(x) = kx \bmod 1$ on S^1

Let $\xi = \{[0, 1/k), [1/k, 2/k), \ldots, [k-1/k, 1)\}$

Easy to see that

$$\xi^{(n)} = \left\{ \left[\frac{i}{k^n}, \frac{i+1}{k^n} \right) \;\middle|\; i = 0, \dots, k^n - 1 \right\}$$

and thus ξ is a generator for T.

$$H(\xi^{(n)}) = -\sum_{i=1}^{k^n} \frac{1}{k^n} \log\left(\frac{1}{k^n}\right) = n \log k \quad \overset{\text{Kolmogorov-Sinai}}{\Rightarrow} \quad h(T) = \log k$$

3) $\left(\frac{1}{2}, \frac{1}{2}\right)$-Bernoulli shift:

Recall cylinders: $[a_0, \dots, a_k] = \{ \{x_i\}_{i \in \mathbb{N}} \mid x_j = a_j,\ j = 0, \dots, k \}$

$\xi = \{[0], [1]\}$ is generating, since $T^{-i}[a] = [\underbrace{*, \dots, *}_{i}, a]$, hence

$\bigvee_{i=0}^{k} T^{-i}\xi$ contains exactly the cylinders $[a_0, \dots, a_k]$, $a_i \in \{0,1\}$, $i = 0, \dots, k$

Thus

$$H(\xi^{(n)}) = -\sum_{a_0, \dots, a_n \in \{0,1\}} \mu[a_0, \dots, a_n] \cdot \log \mu[a_0, \dots, a_n]$$

$$= \sum_{i=1}^{2^{n+1}} \frac{1}{2^{n+1}} \log\left(2^{n+1}\right) = (n+1)\log 2$$

$$\Rightarrow h(T) = \log 2$$

4) (p, P)-Markov shift:

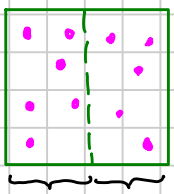
Similar computation as in 3) (cf [Wa, Thm 4.27], [BS, Sec 9.4], [Sa, Prop. 4.7])

yields $h(T) = -\sum_{ij} p_i P_{ij} \log P_{ij}$

4.9 HISTORICAL REMARKS

- Boltzmann (1872) / Gibbs (1878): statistical mechanics & thermodynamics

N particles



microstate: each particle left/right

macrostate: # particles left/right

Q: Given macrostate, how much uncertainty about microstate?

Macro: $(n, N-n) \Rightarrow$ # possible micro $= \binom{N}{n}$ $\rightsquigarrow$ max if $n = N/2$

$\rightsquigarrow$ pressure equilibrium

- von Neumann (~1940): isomorphy of mpts

Q: Are the $(\frac{1}{2},\frac{1}{2})$-Bernoulli and $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$-Bernoulli shifts isomorph?

- Shannon (1948): entropy & information theory
- Kolmogorov, Sinai (1958): entropy of mpts

Thm: Entropy is invariant under measure-theoretic isomorphism

A: $\log 2 \neq \log 3$, hence they are not isomorph

- Ornstein (1970): Bernoulli shifts with same entropy are isomorph

4.10 TOPOLOGICAL ENTROPY

- X: compact metric space
- $T: X \to X$ continuous
- ξ: open cover of X

$$N(\xi) := \min\left\{ |\eta| \;\middle|\; \eta \subseteq \xi,\ X \subseteq \bigcup_{U \in \eta} U \right\}$$

smallest cardinality of finite subcover

It holds $N(\xi \vee \eta) \leq N(\xi) \cdot N(\eta)$ (*)

Q: At which rate is T refining ξ? ↝ Look at $N\left(\bigvee_{k=0}^{n-1} T^{-k}\xi\right)$, $n \geq 0$

By (*), growth is subexponential in n.

Topological entropy

- of T relative to ξ: $h_{top}(T,\xi) = \lim_{n\to\infty} \frac{1}{n} \log N\left(\bigvee_{k=0}^{n-1} T^{-k}\xi\right)$
- of T: $h_{top}(T) = \sup_{\xi} h_{top}(T,\xi)$

↖ finite open covers of X

Theorem (Variational principle)

$T: X \to X$ continuous on the compact metric space X. Then

$$h_{top}(T) = \sup\{ h_\mu(T) \mid \mu \text{ is invariant Borel measure} \}$$