6. (Metric) Entropy

Let $A_2=\{0,1\}$, $A_3=\{0,1,2\}$. Let p_2 and p_3 denote the uniform measures on A_2 and on A_3 respectively. Denote $\mathbb{X}_2=(A_2^{\mathbb{Z}}, \mathcal{B}(A_2^{\mathbb{Z}}), p_2^{\otimes\mathbb{Z}}, \sigma_2)$ and $\mathbb{X}_3=(A_3^{\mathbb{Z}}, \mathcal{B}(A_3^{\mathbb{Z}}), p_3^{\otimes\mathbb{Z}}, \sigma_3)$, where $\sigma_\bullet$ denotes the left shift. Are $\mathbb{X}_2$ and $\mathbb{X}_3$ isomorphic as dynamical systems? This question was answered by Kolmogorov in 1958 by examining an isomorphism invariant of a dynamical system and showing that the invariant takes value $\log 2$ for $\mathbb{X}_2$ and value $\log 3$ for $\mathbb{X}_3$. This invariant is the entropy and we will discuss the notion of entropy introduced by Kolmogorov's student Sinaĭ in 1959. Before defining the entropy of a dynamical system, we need to discuss more basic notions.

Definition

Let $(X,\mathcal{B})$ a measurable space. A partition of $(X,\mathcal{B})$ is a subset $\xi\subseteq\mathcal{B}$ such that

$$X=\bigsqcup_{A\in\xi} A$$

Remark: Let $(X,\mathcal{B})$ a measurable space. There is a 1-1 correspondence

$$\begin{array}{ccc} \{\xi\subseteq\mathcal{B} : \xi \text{ is a finite partition}\} & \xi & \xi(\mathcal{A}) \\ \updownarrow & \downarrow & \uparrow \\ \{\mathcal{A}\subseteq\mathcal{B} : \mathcal{A} \text{ is a finite } \sigma\text{-algebra}\} & \sigma(\xi) & \mathcal{A} \end{array}$$

and $\xi(\sigma(\xi))=\xi$, $\sigma(\xi(\mathcal{A}))=\mathcal{A}$.

Definition

Let $(X,\mathcal{B})$ a measurable space, ξ,η partitions of $(X,\mathcal{B})$.

(i) ξ is a refinement of η, denoted $\eta<\xi$, if every element in η is a union of elements in ξ.

(ii) The common refinement of ξ and η, denoted $\xi\vee\eta$, is the partition of $(X,\mathcal{B})$ given by
$$\xi\vee\eta=\{A\cap B : A\in\xi, B\in\eta\}.$$

In what follows, we'll define the entropy of a (finite) partition. This notion serves to quantify the information gained when observing which element of a partition a point belongs to. Here it helps to switch to a probabilistic viewpoint. Let $(X,\mathcal{B},\mu)$ be a probability space and

think of an element in X as a realization of a possible state. Now suppose we are given a partition ξ of X, which gives a classification of all states. This classification might be coarse or very fine, depending on ξ. How much information do we gain from knowing the class label of x, i.e., from identifying the unique element of ξ containing x?

For any reasonable quantification, the information gain should be 0 if $\xi=\{X\}$. **Moreover**, if $\xi=\{A_1,\dots,A_k\}$, then the uncertainty should be maximal – and hence the average information gain maximal – if all class labels are equally likely, i.e.,

$$\forall 1\le i\le k \quad \mu(A_i)=\frac{1}{k}.$$

Denote the average information gained by $H_\mu(\xi)$. Then this should only depend on the probabilities assigned to the events, i.e.,

for some $H:\Delta_k\to[0,\infty)$

$$H_\mu(\{A_1,\dots,A_k\})=H(\mu(A_1),\dots,\mu(A_k)).$$

Finally, suppose that ξ,η are two partitions of $(X,\mathcal{B})$. Let $B\in\eta$, then the average information gained from knowing the class label according to ξ given label B (according to η) should be

$$H\left(\frac{\mu(A_1\cap B)}{\mu(B)},\dots,\frac{\mu(A_k\cap B)}{\mu(B)}\right)$$

(where $\xi=\{A_1,\dots,A_k\}$).

Hence the average information gain from knowing the ξ-label conditional on the η-label should be

$$H_\mu(\xi|\eta)=\sum_{B\in\eta}\mu(B)H\left(\frac{\mu(A_1\cap B)}{\mu(B)},\dots,\frac{\mu(A_k\cap B)}{\mu(B)}\right).$$

Theorem (Khintchine?)

Given $k\in\mathbb{N}$, let $\Delta_k=\{(p_1,\dots,p_k)\in[0,\infty)^k : \sum_{i=1}^k p_i=1\}$. Suppose

$$H:\bigcup_{k\in\mathbb{N}}\Delta_k\longrightarrow\mathbb{R}$$

satisfies

(i) $\forall k\in\mathbb{N}\,\forall p\in\Delta_k$ $H(p)\ge 0$ and $H(p)=0 \iff \exists 1\le i\le k$ $p_i=1$

(ii) $\forall k\in\mathbb{N}$ $H|_{\Delta_k}$ is continuous

(iii) $\forall k\in\mathbb{N}$ $H|_{\Delta_k}$ is symmetric

(iv) $\forall k\in\mathbb{N}$ $H|_{\Delta_k}$ attains its maximum at $p=(\frac{1}{k},\dots,\frac{1}{k})$

(v) $\forall$ finite partitions ξ,η $H_\mu(\xi\vee\eta)=H_\mu(\xi)+H_\mu(\eta|\xi)$

(vi) $\forall k\in\mathbb{N}\ \forall p\in\Delta_k$ $H(p)=H((p,0))$.

Then $\exists\lambda>0$ such that $\forall k\in\mathbb{N}\ \forall p\in\Delta_k$

$$H(p)=-\lambda\sum_{i=1}^{k}p_i\log p_i=-\lambda\sum_{i=1}^{k}\varphi(p_i)$$

$\varphi(p)=\begin{cases}0 & p=0,\\ -p\log p & \text{if } p\in(0,1]\end{cases}$

Definition

Let $(X,\mathcal{B},\mu)$ a probability space, ξ,η finite partitions of $(X,\mathcal{B})$. The entropy of ξ given η (with respect to μ) is

$$H_\mu(\xi|\eta):=\sum_{\substack{Q\in\eta\\ \mu(Q)\neq 0}}\mu(Q)\left(-\sum_{P\in\xi}\varphi\left(\frac{\mu(P\cap Q)}{\mu(Q)}\right)\right).$$

The entropy of ξ is $H_\mu(\xi):=H_\mu(\xi|\{X\})$.

If $\mathcal{A},\mathcal{C}\subseteq\mathcal{B}$ are finite σ-algebras, then

$$H_\mu(\mathcal{C}|\mathcal{A}):=H_\mu(\xi(\mathcal{C})|\xi(\mathcal{A})),\quad H_\mu(\mathcal{C}):=H_\mu(\xi(\mathcal{C})).$$

Remark: Let $(X,\mathcal{B},\mu)$ be a probability space, $\mathcal{A}\subseteq\mathcal{B}$ a σ-algebra, then for all $f\in L^1(X,\mathcal{B},\mu)$ we have

$$f\geq 0\ \ \mu\text{-a.s.}\implies \mathbb{E}(f|\mathcal{A})\geq 0\ \ \mu\text{-a.s.}$$

Proof: $\max\{\mathbb{E}(f|\mathcal{A}),0\}$ is $\mathcal{A}$-measurable and in $L^2(\mu)$ whenever $f\in L^2_+(\mu)$.

But

$$\|\max\{\mathbb{E}(f|\mathcal{A}),0\}-f\|_2\leq\|\mathbb{E}(f|\mathcal{A})-f\|_2$$

then yields $\mathbb{E}(f|\mathcal{A})=\max\{\mathbb{E}(f|\mathcal{A}),0\}$ μ-a.s. ∎

Lemma:

Let $(X,\mathcal{B},\mu)$ a probability space, $\mathcal{A},\mathcal{C}\subseteq\mathcal{B}$ finite σ-algebras. Then

$$H_\mu(\mathcal{C}|\mathcal{A})=\int_X-\sum_{P\in\xi(\mathcal{C})}\mathbb{1}_P\log(\mathbb{E}(\mathbb{1}_P|\mathcal{A}))\,d\mu.$$

Proof: $\int_X\mathbb{1}_P\log(\mathbb{E}(\mathbb{1}_P|\mathcal{A}))\,d\mu\overset{(*)}{=}\int_X\mathbb{E}(\mathbb{1}_P|\mathcal{A})\log(\mathbb{E}(\mathbb{1}_P|\mathcal{A}))\,d\mu$

$$=\sum_{A\in\xi(\mathcal{A})}\int_A\mathbb{E}(\mathbb{1}_P|\mathcal{A})\log(\mathbb{E}(\mathbb{1}_P|\mathcal{A}))\,d\mu=\sum_{A\in\xi(\mathcal{A})}\mu(A)\varphi\left(\frac{\mathbb{E}_\mu(\mathbb{1}_P\mathbb{1}_A)}{\mathbb{E}_\mu(\mathbb{1}_A)}\right)$$ ∎

(*) extend $g\in L^\infty(X,\mathcal{A},\mu)$ $\Rightarrow\mathbb{E}(f\cdot g|\mathcal{A})=\mathbb{E}(f|\mathcal{A})\cdot g$ to positive functions.

Definition

Let $(X,\mathcal{B},\mu)$ a probability space, ξ a finite partition of $(X,\mathcal{B})$, $\mathcal{A}\subseteq\mathcal{B}$ a σ-algebra. The conditional information of ξ given $\mathcal{A}$ is the function

$$I_\mu(\xi|\mathcal{A})=-\sum_{P\in\xi}\mathbb{1}_P\log\circ\mathbb{E}(\mathbb{1}_P|\mathcal{A})$$

and the conditional entropy of ξ given $\mathcal{A}$ is

$$H_\mu(\xi|\mathcal{A})=\int_X I_\mu(\xi|\mathcal{A})\,d\mu.$$

If $\mathcal{C}\subseteq\mathcal{B}$ is a finite σ-algebra, then

$$I_\mu(\mathcal{C}|\mathcal{A}):=I_\mu(\xi(\mathcal{C})|\mathcal{A}),\quad H_\mu(\mathcal{C}|\mathcal{A}):=H_\mu(\xi(\mathcal{C})|\mathcal{A}).$$

Definition

Let $(X,\mathcal{B})$ a measurable space, ξ a partition of $(X,\mathcal{B})$, $T:X\longrightarrow X$ measurable. Then $T^{-k}\xi=\{T^{-k}A:A\in\xi\}$ ($k\in\mathbb{N}$ if T non-invertible, otherwise $k\in\mathbb{Z}$). Similarly, if $\mathcal{A}\subseteq\mathcal{B}$ is a σ-algebra, then $T^{-k}\mathcal{A}=\{T^{-k}A:A\in\mathcal{A}\}$.

Lemma

Let $(X,\mathcal{B},\mu)$ a probability space, $\xi\subseteq\mathcal{B}$ a finite partition, $\mathcal{A}\subseteq\mathcal{B}$ a σ-algebra. Let $T:X\longrightarrow X$ measurable and $\sigma=T^k_*\mu$. Then

$$I_\mu(T^{-k}\xi|T^{-k}\mathcal{A})=I_\sigma(\xi|\mathcal{A})\circ T^k\quad\mu\text{-a.s.}$$

In particular, $H_\mu(T^{-k}\xi|T^{-k}\mathcal{A})=H_\sigma(\xi|\mathcal{A})$.

Proof: Note that $I_\mu(T^{-k}\xi|T^{-k}\mathcal{A})=\sum_{P\in\xi}\mathbb{1}_P\circ T^k\log\circ\mathbb{E}_\mu(\mathbb{1}_P\circ T^k|T^{-k}\mathcal{A})$, so it suffices to show that $\mathbb{E}_\mu(\mathbb{1}_P\circ T^k|T^{-k}\mathcal{A})=\mathbb{E}_\sigma(\mathbb{1}_P|\mathcal{A})\circ T^k$ μ-a.s.

Let $A\in T^{-k}\mathcal{A}$, then $A=T^{-k}A'$ for some $A'\in\mathcal{A}$, hence

$$\int_A\mathbb{E}_\sigma(\mathbb{1}_P|\mathcal{A})\circ T^k\,d\mu=\int_X(\mathbb{E}_\sigma(\mathbb{1}_P|\mathcal{A})\cdot\mathbb{1}_{A'})\circ T^k\,d\mu=\int_{A'}\mathbb{E}_\sigma(\mathbb{1}_P|\mathcal{A})\,d\sigma$$

$$=\int_{A'}\mathbb{1}_P\,d\sigma=\int_X(\mathbb{1}_P\cdot\mathbb{1}_{A'})\circ T^k\,d\mu=\int_A\mathbb{1}_P\circ T^k\,d\mu.$$

Hence $\mathbb{E}_\sigma(\mathbb{1}_P|\mathcal{A})\circ T^k=\mathbb{E}(\mathbb{1}_P\circ T^k|T^{-k}\mathcal{A})$ μ-a.s. ∎

Definition

Given $(X,\mathcal{B},\mu)$ a probability space, $T: X \longrightarrow X$ measurable, ξ a partition of $(X,\mathcal{B})$, and integers $m \le n$ (with $0 \le m$ if T is non-invertible), we define
$$\xi_m^n = \bigvee_{i=m}^{n} T^{-i}\xi = \left\{ \bigcap_{i=m}^{n} T^{-i}A_i : A_m, \dots, A_n \in \xi \right\}$$

Lemma (enter dynamics!)

Let $(X,\mathcal{B},\mu,T)$ a pmps. Suppose ξ is a partition of $(X,\mathcal{B})$. Then for all $n\in\mathbb{N}$
$$H_\mu(\xi_0^{n-1}) = H_\mu(\xi) + \sum_{j=1}^{n-1} H_\mu(\xi \mid \xi_1^j).$$
Moreover, the sequence $\frac{1}{n}H_\mu(\xi_0^{n-1})$ is monotonically decreasing.

Proof: The formula is certainly true if $n=1$ (defining an empty sum as 0). We proceed by induction, i.e., suppose $n\in\mathbb{N}$ and that the formula is valid for n. Then
$$H_\mu(\xi_0^n) = H_\mu(\xi \vee \xi_1^n) = H_\mu(\xi_1^n) + H_\mu(\xi \mid \xi_1^n)$$
previous lemma + $T_*\mu=\mu$
$$\overset{\downarrow}{=} H_\mu(\xi_0^{n-1}) + H_\mu(\xi \mid \xi_1^n) = H_\mu(\xi) + \sum_{j=1}^{n-1} H_\mu(\xi \mid \xi_1^j).$$

Exercise: Let ϱ, η finite partitions of $(X,\mathcal{B})$. Then
$$\eta < \varrho \Rightarrow H_\mu(\xi \mid \varrho) \le H_\mu(\xi \mid \eta)$$
For monotonicity, use that by the exercise
$$H_\mu(\xi) + \sum_{j=1}^{n-1} H_\mu(\xi \mid \xi_1^j) \ge n H_\mu(\xi \mid \xi_1^n)$$
and, therefore,
$$\frac{1}{n} H_\mu(\xi_0^{n-1}) \ge H_\mu(\xi \mid \xi_1^n).$$
Thus
$$H_\mu(\xi_0^n) = H_\mu(\xi_0^{n-1}) + H_\mu(\xi \mid \xi_1^n) \le \frac{n+1}{n} H_\mu(\xi_0^{n-1}).$$
$\square$

Definition

Let $(X,\mathcal{B},\mu,T)$ a pmps, ξ a finite partition of $(X,\mathcal{B})$. Then the entropy of T with respect to ξ is
$$h_\mu(T,\xi) := \lim_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1})$$

Theorem

Let $(X,\mathcal{B},\mu,T)$ a pmps, ξ a finite partition of $(X,\mathcal{B})$. Then
$$h_\mu(T,\xi) = \lim_{n\to\infty} H_\mu(\xi \mid \xi_1^n).$$

Proof: $\forall n\in\mathbb{N}$ $H_\mu(\xi \mid \xi_1^n) \ge H_\mu(\xi \mid \xi_1^{n+1}) \ge 0$. Hence $\lim_{n\to\infty} H_\mu(\xi \mid \xi_1^n)$ exists.

Thus $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n-1} H_\mu(\xi \mid \xi_1^j) = \lim_{n\to\infty} H_\mu(\xi \mid \xi_1^n)$ and
$$h_\mu(T,\xi) = \lim_{n\to\infty} \frac{1}{n}\left(H_\mu(\xi) + \sum_{j=1}^{n-1} H_\mu(\xi \mid \xi_1^j)\right) = \lim_{n\to\infty} H_\mu(\xi \mid \xi_1^n).$$

Theorem

Let $(X,\mathcal{B},\mu,T)$ a pmps, ξ a finite partition of $(X,\mathcal{B})$. Then

(i) $h_\mu(T,\xi) \le H_\mu(\xi)$ and

(ii) $h_\mu(T,\xi) = h_\mu(T,\xi_0^k)$ for all $k\in\mathbb{N}$.

If η is a finite partition of $(X,\mathcal{B})$ and $\xi < \eta$, then $h_\mu(T,\xi) \le h_\mu(T,\eta)$.

Proof: Exercise: Let ϱ a partition of $(X,\mathcal{B})$, $n\in\mathbb{N}$, then
$$H_\mu(\varrho_0^{n-1}) \le n H_\mu(\varrho).$$
Using the exercise, we have
$$\frac{1}{n} H_\mu(\xi_0^{n-1}) \le H_\mu(\xi),$$
hence (i) follows.

Let $k\in\mathbb{N}$, then
$$\frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi_0^k\right) = \frac{1}{n} H_\mu(\xi_0^{n+k-1}) \xrightarrow{n\to\infty} h_\mu(T,\xi).$$
Let now $\xi < \eta$, then $\forall n\in\mathbb{N}$ $\xi_0^n < \eta_0^n$

Exercise: Suppose $\varsigma < \chi$, then
$$H_\mu(\varsigma) \le H_\mu(\chi)$$
Using the exercise, we have
$$\frac{1}{n} H_\mu(\xi_0^{n-1}) \le \frac{1}{n} H_\mu(\eta_0^{n-1})$$
for all $n\in\mathbb{N}$. Hence $h_\mu(T,\xi) \le h_\mu(T,\eta)$.