Entropy—A Concrete Introduction
Student-Directed Colloquium 4/24/2019

1 Information

- Before talking about entropy, we need to answer a different question: *how to quantify information.*
- If we do an experiment, and the experiment has n possible outcomes, each with probability p_i, how much information do we gain if outcome i occurs?
- Example: The first letter of a word is X. How much information do we have about this word? What if the first letter is T?
- Suppose we have a finite alphabet $E = \{e_1, \dots, e_N\}$, and $X_1, X_2, \dots$ are i.i.d. random variables with values in E.
- Recall this means X_i are measurable functions from a probability space $(\Omega, \mathbb{P})$ to E. For example, we could have $\Omega = E^{\mathbb{N}}$.
- Suppose $\mathbb{P}[X_i = e] = p_e$ for each $e \in E$. Let $p = (p_{e_1}, \dots, p_{e_N})$ be the probability distribution vector for E.
- Question: How do we determine the amount of *information* encoded in the outcome $X_1(\omega), X_2(\omega), \dots, X_n(\omega) \in E$?
- One way we can do this is say a computer is storing this information for us. How should we program the computer to store each piece of information in the letter e_i?
- Associate to each letter $e \in E$ a short string of 0s and 1s. Then the string $e_1, \dots e_n$ becomes a string of 0s and 1s.
- Let's think about what we want for this code:
  1. More probable letters should have fewer digits. This economizes on storage: if we have higher-probability events with too long strings, our storage will fill up faster. (Consider Morse code: the letters e and t are $\cdot$ and $-$, but q is $- - \cdot -$).
  2. Our code needs to be translatable from binary back into something useful. So no code for a letter can also be the beginning of another code.
     - Example: If S is encoded with 0110, we can't encode L with 011011. Since as a computer translates, we want the computer to know when it reaches the end of a letter. Imagine if we also encoded H with 11. Then Ana might try texting Caitlin "Dom's talk is lit", but Caitlin's phone may render this as "Dom's talk is shit"!

- The latter condition makes this a *binary prefix code*, and is actually how computers transcribe letters to binary and back.

- Let $\ell(e)$ be the length of the code for the letter e, and let $c(e) \in \{0,1\}^{\ell(e)}$ be the code of e. Let

$$C = \{c(e) : e \in E\} \subseteq \coprod_{k=1}^{\#E} \{0,1\}^k$$

be the binary code for the alphabet E.

- The first condition in this list can be achieved by minimizing the *expected length* of the code of a random symbol:

$$L_p(C) := \int_E \ell \, dp = \sum_{e \in E} p_e \ell(e)$$

- NOW, we're going to construct a specific code, and show that it's almost optimal.

- Assume E is enumerated so that $p_{e_1} \geq p_{e_2} \geq \cdots \geq p_{e_N}$.

- Define $\lambda : E \to \mathbb{N}$ so that $2^{-\lambda(e)} \leq p_e < 2^{-\lambda(e)+1}$.

- Also define $\widetilde{p}_e = 2^{-\lambda(e)}$ and $\widetilde{q}_k = \sum_{j<k} \widetilde{p}_{e_j}$.

- Then $\lambda(e_m) \leq \lambda(e_k)\ \forall m \leq k$. So the binary representation of the number $\widetilde{q}_k$ has at most $\lambda(e_k)$ digits:

$$\widetilde{q}_k = \sum_{i=1}^{\lambda(e_k)} c_i(e_k) 2^{-i}$$

for uniquely determined $c_1(e_k), \ldots, c_{\lambda(e_k)}(e_k) \in \{0,1\}$. Indeed, the smallest power of $1/2$ that is used in the construction of $\widetilde{q}_k$ is $2^{-\lambda(e_{k-1})}$.

- Observe also that $\widetilde{q}_m \geq \widetilde{q}_k + 2^{-\lambda(e_k)}$ for $m > k$. So,

$$\big(c_1(e_k), \ldots, c_{\lambda(e_k)}(e_k)\big) \neq \big(c_1(e_m), \ldots, c_{\lambda(e_k)}(e_m)\big)$$

Indeed, adding another term of $2^{-\lambda(e_k)}$ would cascade a change in preceding digits c_i.

- UPSHOT: The code $C = \{c(e) : e \in E\}$ is a prefix code, where $c(e) = \big(c_1(e), \ldots, c_{\lambda(e)}(e)\big)$. The length of each code is thus $\ell(e) = \lambda(e)$.

2 Information Entropy

- Now, recall $2^{-\ell(e)} \leq p_e < 2^{-\ell(e)+1}$. So, $-\ell(e) \leq \log_2(p_e) < -\ell(e) + 1$, or $-\log_2(p_e) \leq \ell(e) \leq 1 - \log_2(p_e)$.
- So the expected length is bounded in the following way:

$$-\sum_{e\in E} p_e \log_2(p_e) \leq L_p(C) \leq 1 - \sum_{e\in E} p_e \log_2(p_e)$$

- **Definition.** For a probability distribution $p = (p_e)_{e\in E}$ on a countable set E, the *binary entropy* of p is

$$H_2(p) := -\sum_{e\in E} p_e \log_2(p_e)$$

where we use the convention $0 \log 0 = 0$. If we replace 2 by Euler's constant $e = 2.71...$, then $H_e(p) = H(p)$ is the *Shannon entropy*, or simply the *entropy*:

$$H(p) = -\sum_{e\in E} p_e \log(p_e)$$

Theorem 1. *Let $p = (p_e)_{e\in E}$ be a probability distribution on a finite alphabet E. Then for any binary prefix code $C = \{c(e) : e \in E\}$, we have $L_p(C) \geq H_2(p)$. Furthermore, there is a binary prefix code C with $L_p(C) \leq H_2(p) + 1$.*

Theorem 2. *Let E be a finite set and let p be a probability vector on E. Then the entropy $H(p)$ is minimal if $p = \delta_e$ for some $e \in E$; that is, if $p_{e'} = 0$ if $e' \neq e$, and $p_e = 1$. In this case, $H(p) = 0$.*

On the other hand, $H(p)$ is maximal if $p_e = 1/\#E$ for every $e \in E$ (that is, p is uniformly distributed). In this case, $H(p) = \log(\#E)$.

Proving the second theorem is a simple Lagrange multipliers exercise.

Theorem 3 (Shannon)**.** *Let E be a finite set, and let $X_1, X_2, \ldots : \Omega \to E$ be i.i.d. random variables with $\mathbb{P}[X_i = e] = p_e$ for every $i \geq 1$, so that $p = (p_{e_1}, \ldots, p_{e_N})$ is a probability vector on E. For $\omega \in \Omega$, define*

$$\pi_n(\omega) = \prod_{i=1}^{n} p_{X_i(\omega)}$$

Then $\pi_n(\omega)$ is the probability that the observed sequence $X_1(\omega), \ldots, X_n(\omega)$ occurs. Finally let $Y_n(\omega) = -\log\left(p_{X_n(\omega)}\right)$, the information after the n^{th} experiment. Then,

$$-\frac{1}{n} \log \pi_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{n\to\infty} H(p) \quad a.s.$$

This follows from strong law of large numbers.

- So how is entropy a measurement of disorder? It becomes clear in these three theorems.
- The first one shows that $H_2(p)$ is a good estimate for the expected complexity of encoded information on an experiment.
- The second one shows that if a certain outcome happens 100% of the time, then the entropy is 0. But if every outcome is equally likely, and an experiment is repeated, then we essentially see "randomness" in the experiment.
- The third one shows that the average information received converges to the entropy $H(p)$.
- What does this have to do with entropy in physics? Well, if we have a medium with particles, then we can look at a (finite) number of possible configurations the particles can take on.
- If all configurations are equally probable, the particles are highly random and disordered; this is maximal entropy.

3 One-Sided Shifts

- But in particular, we're interested in probability theory and dynamical systems.
- We define the **Bernoulli shift**: Let E be a finite alphabet with probability vector $p = (p_{e_0}, \dots, p_{e_{N-1}})$ and let $\Omega^+ := E^{\mathbb{N}_0}$, and let $\mathbb{P}$ be the probability measure defined on *cylinders*:

  $$[x_0, \dots x_{n-1}] := \left\{\omega \in \Omega^+ : \omega_i = x_i \ \forall 0 \leq i \leq n-1\right\}$$

  so the probability is defined as

  $$\mathbb{P}\left[x_0, \dots, x_{n-1}\right] = \prod_{i=0}^{n-1} p_{e_i}$$

  We can interpret Ω^+ as the space of all sequences of experimental outcomes.
- Say the n^{th} outcome is given by $X_n(\omega)$. Then $X_n(\omega) = \omega_n$ is simply the projection on the n^{th} coordinate.
- However, in dynamical systems, we typically treat a stochastic process like $(X_n)_{n\geq 1}$ instead as a composition of an observable function $f : \Omega^+ \to \mathbb{R}$ with a measurable transformation $T : \Omega^+ \to \Omega^+$.

- In this case, we let T be the *shift map*:

$$T(\omega)_i = \omega_{i+1}$$

That is, $T(\omega)$ is the sequence obtained by shifting the sequence ω to the left by one and chopping off the first letter of the sequence.

- So, if $f(\omega) := \omega_1$, the coordinate projection X_n can instead be expressed as $X_n = f \circ T^n$.

- In particular, T is a *measure-preserving transformation.* If we take a cylinder $[x_0, \ldots, x_{n-1}]$, then $T^{-1}[x_0, \ldots, x_{n-1}]$ has measure equal to the measure of the cylinder:

$$\mathbb{P}\left(T^{-1}[x_0, \ldots, x_{n-1}]\right) = \mathbb{P}\left(\coprod_{i=0}^{N-1} [x_i, x_1, \ldots, x_n]\right) = \sum_{i=0}^{N-1} p_{e_i} \prod_{j=0}^{n-1} p_{e_j} = \prod_{j=0}^{n-1} p_{e_j}$$
$$= \mathbb{P}\left[x_0, \ldots, x_{n-1}\right]$$

- Most of the important maps of ergodic theory are these: measurable and measure-preserving transformations. Because the stochastic processes they generate, $X_n = f \circ T^n$, are identically distributed.

4 Metric Entropy of Bernoulli Shift

- For $n \in \mathbb{N}$, denote by P_n the probability measure on E^n given by the projection of $\mathbb{P}$ on $E^{\mathbb{N}}$ onto the first n coordinates. That is:

$$P_n\left(\{e_0, \ldots, e_{n-1}\}\right) := \mathbb{P}\left[e_0, \ldots, e_{n-1}\right]$$

Theorem 4. *Let E^1 and E^2 be finite sets with probability vectors p^1 and p^2. Let p be a probability vector on the finite set $E^1 \times E^2$ with marginals p^1 and p^2:*

$$\sum_{e^2 \in E^2} p_{(e^1, e^2)} = p^1_{e^1} \quad \forall e^1 \in E^1 \quad \textit{(probability of 1st coordinate being } e^1\textit{)}$$

and

$$\sum_{e^1 \in E^1} p_{(e^1, e^2)} = p^2_{e^2} \quad \forall e^2 \in E^2 \quad \textit{(probability of 2nd coordinate being } e^2\textit{)}$$

Then $H(p) \leq H(p^1) + H(p^2)$.

- In particular, this implies the entropies $H(P^{m+n})$, $H(P^m)$, and $H(P^n)$ for the finite probability spaces E^{m+n}, E^m, and E^n respectively satisfy:

$$H(P^{m+n}) \leq H(P^m) + H(P^n)$$

- It is an exercise in real analysis that the following limit exists:

$$h := h_{\mathbb{P}}(T) := \lim_{n\to\infty} \frac{1}{n} H(P^n) = \inf_{n \geq 1} \frac{1}{n} H(P^n)$$

We call this the *entropy* of the system.

5 Metric Entropy in Ergodic Theory

- Now suppose $(\Omega, \mathcal{A}, \mu)$ is a general probability space, and let $T : \Omega \to \Omega$ be a measurable transformation. Let $f : \Omega \to \mathbb{R}$ be an observable.
- A collection of measurable subsets $\xi = \{C_i\}_{i \in I}$ of $(\Omega, \mathcal{A}, \mu)$ is called a **measurable partition** if $\mu(C_i \cap C_j) = 0$ for $i \neq j$, and $\mu\left(\bigcup_{i \in I} C_i\right) = 1$.
- We can consider each of these C_is to be one of finitely many outcomes in an experiment—letters in an alphabet, for example.
- With this interpretation, the *entropy* of the partition is

$$H(\xi) = H_\mu(\xi) = -\sum_{C \in \xi} \mu(C) \log \mu(C)$$

- What if we want to consider not just events at the first reading of the experiment, but after a second reading at time 1?
- Well now there are more possibilities: we have to consider the events right now, but we also have to consider the events of the next stage in the experiment. That is, we not only consider events $C \in \xi$, but also $T^{-1}(C) \in T^{-1}(\xi)$.
- A measurable partition ξ' is a **refinement** of a measurable partition ξ if $\mu(C_i' \setminus C_j) = 0$ for every $C_i' \in \xi'$, $C_j \in \xi$; that is, every element of ξ' is contained (up to a set of measure 0) in an element of ξ.
- Given two partitions ξ and η, the **common refinement** $\xi \vee \eta$ is the smallest partition that is a refinement of both ξ and η. That is, the partition of intersections:

$$\xi \vee \eta := \{C_i \cap C_j : C_i \in \xi,\ C_j \in \eta\}.$$

- In particular, if we consider events that happen both now and will happen at the next stage, we consider the common refinement of ξ and $T^{-1}(\xi)$:

$$T^{-1}(\xi) \vee \xi = \left\{T^{-1}(C_i) \cap C_j : C_i, C_j \in \xi\right\}$$

- Of course, we can then ask what happens at the stage after the next one, and take three common refinements (since common refining is obviously associative and commutative):

$$T^{-2}(\xi) \vee T^{-1}(\xi) \vee \xi = \left\{T^{-2}(C_i) \cap T^{-1}(C_j) \cap C_k : C_i, C_j, C_k \in \xi\right\}$$

- And on, and on. As with the one-sided shift, we get:

$$H\left(\xi^{m+n}\right) \leq H\left(\xi^m\right) + H\left(\xi^n\right), \quad \text{where} \quad \xi^n = \bigvee_{k=0}^{n-1} T^{-k}(\xi)$$

whence the following limit exists:

$$h_\mu(T, \xi) := \lim_{n\to\infty} \frac{1}{n} H_\mu \left(\bigvee_{k=0}^{n-1} T^{-k}(\xi) \right)$$

- That's the *entropy of* T *with respect to the partition* ξ. And it looks confusing, but actually it's surprisingly simple: it's the long-term asymptotically observed disorder after repeating an experiment while observing a finite number of possible outcomes.

- But a partition ξ is generally not part of the structure of a dynamical system. So there's one more step in the construction of entropy. This is to eliminate the consideration of a partition altogether.

- **Definition.** The *Kolmogorov-Sinai Entropy* (a.k.a. the *metric entropy*) of the measure-preserving dynamical system $(\Omega, \mathcal{A}, \mu, T)$ is

$$h_\mu(T) = \sup_\xi h_\mu(T, \xi),$$

where the supremum is over all finite measurable partitions of Ω.