Entropy

- The entropy of a signal as first defined by Shannon measures its unpredictability or randomness
- For example, a fair coin toss is very unpredictable, no matter what knowledge you have, the next event is not predictable
- On the other hand, a biased 30-70 coin, one can predict that the 70% side will come up more frequently.
- Another standard example is English text, "z" is infrequent, "a" is frequent, a "q" is usually followed by a "u", etc.
- Information comes from lack of predictability, "you learn something new"

- The first step is to define the information of an event

- Think of X as a collection of RESULTS of an experiment, for example, the hair color of all humans on earth

- X is partitioned into a number of events (disjoint) eg brown, black, red, blond

$$X = A_1 \amalg A_2 \amalg \dots \amalg A_k$$

- There is a probability measure on X which indicates the probability of an event $\mu(A_i)$

- The information of an event $I(c)$ needs to satisfy some basic properties

(1) I should be non-negative and a decreasing function of the probability of an event: the lower the probability of an event, the greater the information content

(2) $I(X) = 0$ or $I(\mu(X)) = I(1) = 0$

(3) independent events should have additive inform

$$I(C \wedge D) = I(C) + I(D)$$

This implies that

It is not hard to show that

$$I(C) = \alpha\,(-\log(\mu(C)))$$

is the only such function $\alpha > 0$. Note by the change of base formula, α just indicates the base. Standard choices are base 2, base 10, and in ergodic Theory, natural logs.

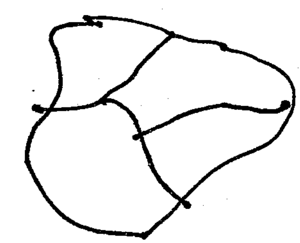
- The next step is to take in account the entire partition
- Recall from probability the expected value
  - Example, say there are 3 outcomes red, blue, green with payoffs \$3, \$4 and \$5 then the expected value is $3\mu(\text{red}) + 4\mu(\text{blue}) + 5\mu(\text{green})$ so if red comes up 30% of the time, blue 50% and green 20%, $EV = 3 \cdot .3 + 4 \cdot .5 + 5 \cdot .2 = 3.9$. So if the game costs \$4, on average you lose 10¢ per game.
- The entropy of the partion is the expected value of the information

$$\mu(A_1) I(A_1) + \dots + \mu(A_k) I(A_k) = -\sum \mu(A_i) \log(A_i)$$

with the convention, $0 \log 0 = 0$

PARTITIONS

- A partition of the probability space is a disjoint collection of measurable sets whose union is $\mathbb{X}$ or $A_i \in \mathcal{B}$, $\coprod A_i = \mathbb{X}$



- Hence forth all partitions will be finite or $\{A_1, \ldots, A_k\}$ with $\coprod_{i=1}^{k} A_i = \mathbb{X}$

- As above a partition $\rho = \{A_1, \ldots, A_k\}$ is a way of dividing the out comes of an experiment with values in $\mathbb{X}$ into a finite number of disjoint out comes.

- $\Sigma_2^{+} = {}_0[00] \coprod {}_0[01] \coprod {}_0[10] \coprod {}_0[11]$

- Given a finite partition ρ, $A(\rho)$ is the collection of all finite unions of elements of ρ. It is a finite subalgebra of $\mathcal{B}$.

- Conversely, given a finite subalgebra $\mathcal{C}$ then all nonempty sets of the form $B_1 \cap \ldots \cap B_\ell$ where $B_i = C_i$ or $B_i = C_i^c$ form a finite partition of X, called $\rho(\mathcal{C})$

- $A(\rho(\mathcal{C})) = \mathcal{C}$ and $\rho(A(\eta)) = \eta$

  so finite partitions and subalgebras bijectively correspond.

- Sometimes one or the other is better to work with

Entropy of a partition

- If $\rho = \{A_1, \ldots, A_k\}$ is a finite partition of the probability space $(X, \mathcal{B}, \mu)$ then its entropy is

$$H(\rho) = -\sum_{i=1}^{k} \mu(A_i) \log \mu(A_i)$$

with the convention (dictated by continuity) that

$$0 \log 0 = 0$$

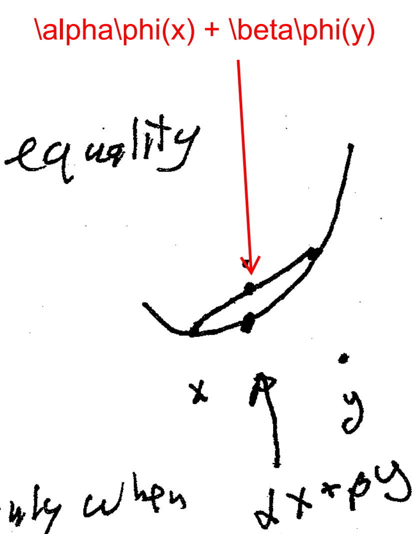
- If $\mathcal{A}$ is a finite subalgebra $\rho(\mathcal{A}) = \{A_1, \ldots, A_k\}$ then $H(\mathcal{A}) = -\sum_{i=1}^{k} \mu(A_i) \log \mu(A_i)$ as above.

- Basic properties of H follow from those of $\phi : [0, \infty) \to \mathbb{R}$ given by $\phi(x) = \begin{cases} 0 & , \text{if } x = 0 \\ x \log x & , \text{if } x \neq 0 \end{cases}$

8

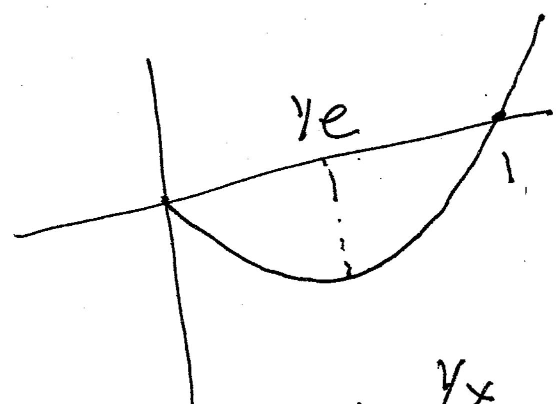
Lemma • ϕ is continuous at zero, strictly convex

or $\phi(\alpha x+\beta y)\le \alpha\phi(x)+\beta\phi(y)$, $x,y\in[0,\infty)$, $\alpha,\beta\ge 0$, $\alpha+\beta=1$; with equality only when $x=y$ or $\alpha=0$ or $\beta=0$.



• Further, $\phi\left(\sum_{i=1}^{k}\alpha_i x_i\right)\le\sum_{i=1}^{k}\alpha_i\phi(x_i)$

$x_i\in[0,\infty)$, $\alpha_i\ge 0$, $\sum\alpha_i=1$ with equality only when all x_i corresponding to nonzero α_i are equal.



Proof $\lim_{x\to 0} x\log x=\lim_{x\to 0}\frac{\log x}{1/x}=\lim_{x\to 0}\frac{1/x}{-1/x^2}=0$

by LHR.

on $(0,\infty)$ $\phi'(x) = 1 + \log x$ and $\phi''(x) = \frac{1}{x} > 0$

which using a standard result implies the strict convexity.

The further follows from an induction, here is the inductive step for $n=3$. Given $\alpha x + \beta y + \gamma z$

Let $w = \frac{\alpha x + \beta y}{1-\gamma}$ since $\alpha + \beta + \gamma = 1$ $\alpha + \beta = 1 - \gamma$ and so $(1-\gamma) w + \gamma z$ satisfies

$\frac{\alpha}{1-\gamma} + \frac{\beta}{1-\gamma} = 1$. And also

$(1-\gamma) + \gamma = 1$ Thus $\phi(\alpha x + \beta y + \gamma z) = \phi((1-\gamma) w + \gamma z)$

$\leq (1-\gamma)\phi(w) + \gamma \phi(z) = (1-\gamma)\phi\left(\frac{\alpha}{1-\gamma} x + \frac{\beta}{1-\gamma} y\right) + \gamma \phi(z)$

$\leq \alpha \phi(x) + \beta \phi(y) + \gamma (\phi(z))$.

COR: If $\rho = \{A_1, \dots, A_k\}$ then $H(\rho) \le \log k$ and $H(\rho) = \log k \iff \mu(A_i) = 1/k$ for all k

Recall $H(\rho) = -\sum \mu(A_i) \log \mu(A_i)$

Proof: $= -\sum \phi(A_i) = -k \sum \frac{1}{k} \phi(A_i)$. Since $\sum 1/k = 1$ we get by the opposite of the theorem $\ge -k\, \phi(\sum 1/k\, \mu(A_i))$

$= -k\, \phi(\frac{1}{k} \sum \mu(A_i)) = -k\, \phi(1/k) = -k \cdot 1/k \log k = \log k$

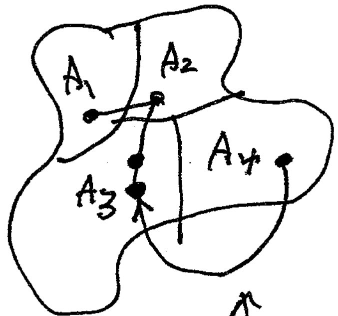
ρ is a partition so $\sum \mu(A_i) = 1$

And note the MAX is achieved when $\mu(A_i) = 1/k$ for then $H(\rho) = -\sum 1/k \log(1/k) = \log k$. $\boxtimes$

The rough idea behind entropy in dynamics is to use a partition to generate a signal or think of the dynamics as repeating an experiment

Define $\sigma : X \to \Sigma_4^+$ via

$$(\sigma(x))_i = j \iff f^i(x) \in A_j.$$



$$\sigma(x) = .4\,3\,3\,2\,1\,\ldots$$

and analyze the entropy of the signal. But one must make this independent of the choice of partition and the influence of the dynamics in more detail.