Math 554

Hemanshu Kaul

kaul@iit.edu

Entropy, or Information Complexity

Let X be a random variable that takes only finitely many values.

We write $p(x)$ for $\mathbb{P}[X=x]$

and for any event E, we write $p(x|E)$ for $\mathbb{P}[X=x \mid E]$

& similarly $p(x|y)$ for $\mathbb{P}[X=x \mid Y=y]$.

[Claude Shannon, 1948] **Entropy of X** is defined as

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}, \text{ where } x \text{ varies over the range of } X$$

$$= -\sum_x p(x) \log p(x)$$

Convention: $0 \log 0 = 0$

$\log$ is base 2.

Entropy, or Information Complexity

Let X be a random variable that takes only finitely many values.

We write $p(x)$ for $\mathbb{P}[X=x]$

and for any event E, we write $p(x|E)$ for $\mathbb{P}[X=x \mid E]$

& similarly $p(x|y)$ for $\mathbb{P}[X=x \mid Y=y]$.

[Claude Shannon, 1948] **Entropy of X** is defined as

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}, \text{ where } x \text{ varies over the range of } X$$

$$= -\sum_x p(x) \log p(x)$$

Convention: $0 \log 0 = 0$

log is base 2.

→ # bits of information conveyed by X. [Shannon Noiseless Coding Theorem says minimum number of bits needed to encode n iid copies of X is $nH(X) + o(n)$.

→ Expected amount of information contained in a realization of X.

→ Average information content missing we don't know the value of X.

"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

- Claude Shannon, 1971.

$S(p)$ = "Surprise" on seeing an event that occurs with probability p.

Let $S: [0,1] \to \mathbb{R}^+$ be defined so that –

1. $S(1) = 0$ [no surprise on seeing a certain event]
2. If $p < q$, then $S(p) > S(q)$ [rarer event is more surprising]
3. S is a continuous function of P
4. $S(pq) = S(p) + S(q)$ [Suppose two two ind. events E&F occur with probab. p & q, resp. Surprise on seeing E∧F is $S(pq)$ which can be reasonably interpreted as $S(p)$ + Surprise on seeing F given that we have already seen E]

$S(p)$ = "surprise" on seeing an event that occurs with probability p.

Let $S: [0,1] \to \mathbb{R}^+$ be defined so that –

1. $S(1) = 0$ [no surprise on seeing a certain event]
2. If $p < q$, then $S(p) > S(q)$ [rarer event is more surprising]
3. S is a continuous function of P
4. $S(pq) = S(p) + S(q)$ [Suppose two two ind. events E&F occur with probab. p & q, resp. Surprise on seeing E∧F is $S(pq)$ which can be reasonably interpreted as $S(p)$ + surprise on seeing F given that we have already seen E]

Theorem There is a unique function S that satisfies conditions 1-4 as well as a normalizing condition such as $S(\frac{1}{2}) = 1$, given by $S(p) = -\log p$.

Proof Try it!

Binary Entropy function

Let $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$ } Flipping a coin. Bernoulli r.v.

Define $H : [0,1] \to \mathbb{R}$ as $H(p) = -p \log p - (1-p) \log (1-p)$

"Abuse of notation: $H(p)$ vs. $H(X)$"

Binary Entropy function

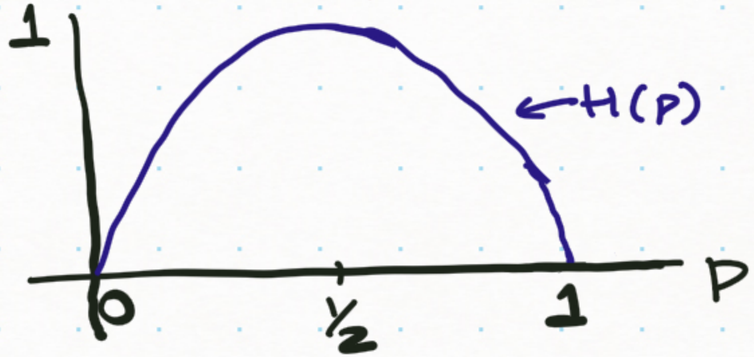
Let $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$ } Flipping a coin. Bernoulli r.v.

Define $H : [0,1] \to \mathbb{R}$ as $H(p) = -p \log p - (1-p)\log(1-p)$

"Abuse of notation: $H(p)$ vs. $H(X)$"

$H(0) = H(1) = 0$

$H(p)$ is maximized when $p = 1/2$



Binary Entropy function

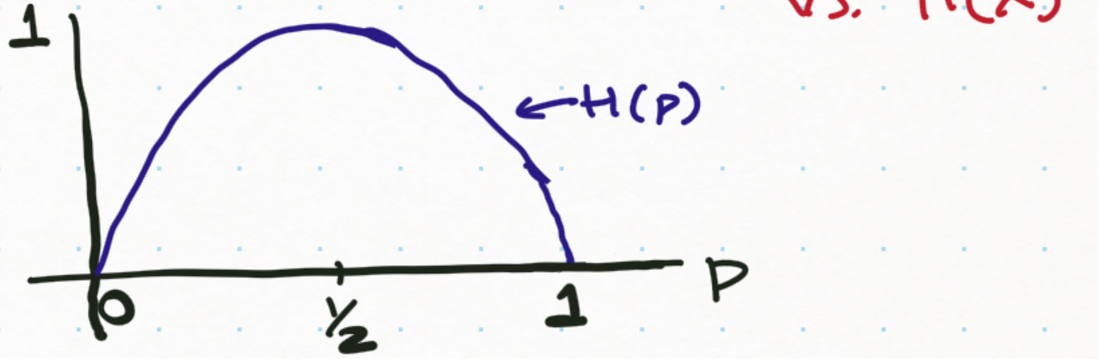
Let $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$ } Flipping a coin. Bernoulli r.v.

Define $H : [0,1] \to \mathbb{R}$ as $H(p) = -p \log p - (1-p)\log(1-p)$

"Abuse of notation: $H(p)$ vs. $H(X)$"

$H(0) = H(1) = 0$

$H(p)$ is maximized when $p = 1/2$



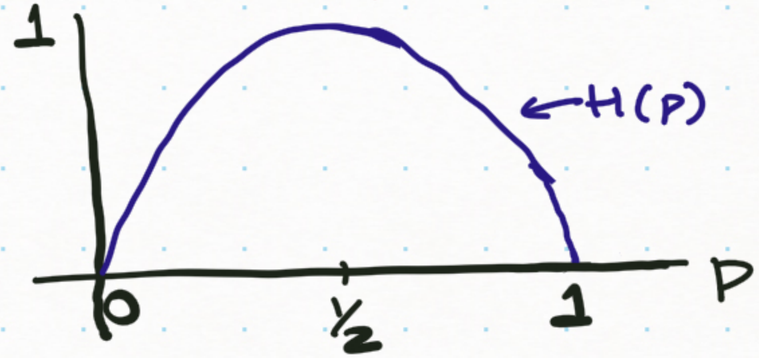
- Flip a 2-sided fair coin, we get $H(X) = H(p) = H(1/2) = 1$ bit of info.
- In general consider an experiment with m equally likely outcomes then entropy is $\sum_{i=1}^{m} \frac{1}{m} \log m = \log m$
- Roll a 8-sided fair die. Outcome is a sequence of 3 bits. and we get 3 bits of info : $H[X] = \log 2^3 = 3$.

Binary Entropy function

Let $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$ } Flipping a coin. Bernoulli r.v.

Define $H : [0,1] \to \mathbb{R}$ as $H(p) = -p \log p - (1-p)\log(1-p)$

"Abuse of notation: $H(p)$ vs. $H(X)$"



$H(0) = H(1) = 0$

$H(p)$ is maximized when $p = 1/2$

We get $H(3/4)$ $(= H(1/4))$ bits of info when a coin that lands head with $p = 3/4$ is flipped.

- Flip a 2-sided fair coin, we get $H(X) = H(p) = H(1/2) = 1$ bit of info.
- In general consider an experiment with m equally likely outcomes then entropy is $\sum_{i=1}^{m} \frac{1}{m} \log m = \log m$
- Roll a 8-sided fair die. Outcome is a sequence of 3 bits. and we get 3 bits of info : $H[X] = \log 2^3 = 3$.

Entropy and Counting

Lemma (Jensen's Inequality)

Let f be a continuous concave function. Then for any reals $x_1, \dots, x_n$,

$$\frac{\sum_{i=1}^{n} f(x_i)}{n} \leq f\left(\frac{\sum x_i}{n}\right),$$

or more generally,

$$\frac{\sum a_i f(x_i)}{\sum a_i} \leq f\left(\frac{\sum a_i x_i}{\sum a_i}\right) \quad \text{for } a_i > 0 \ \forall i.$$

Or, $\int_\Omega f \circ g \, d\mu \leq f\left(\int_\Omega g \, d\mu\right)$

$\mathbb{E}[f(x)] \leq f(\mathbb{E}[X])$

Property (Uniform Bound)

$H(X) \leq \log |\text{Range}(X)|$, where $\text{Range}(X)$ is the set of values X takes with positive probability.

$H(X) = \log |\text{Range}(X)|$ iff X is uniform on its range.

Proof $H(X) = \sum p(x) \log \frac{1}{p(x)} \leq \log\left(\sum p(x) \frac{1}{p(x)}\right) = \log(|\text{Range}(X)|)$

Entropy and Counting

Lemma (Jensen's Inequality)

Let f be a continuous concave function. Then for any reals $x_1, \dots, x_n$,

$$\frac{\sum_{i=1}^{n} f(x_i)}{n} \leq f\left(\frac{\sum x_i}{n}\right),$$

or more generally,

$$\frac{\sum a_i f(x_i)}{\sum a_i} \leq f\left(\frac{\sum a_i x_i}{\sum a_i}\right) \quad \text{for } a_i > 0 \ \forall i.$$

Or, $\int_\Omega f \circ g \, d\mu \leq f\left(\int_\Omega g \, d\mu\right)$

$\mathbb{E}[f(x)] \leq f(\mathbb{E}[X])$

Property (Uniform Bound)

$H(X) \leq \log |\text{Range}(X)|$, where Range(X) is the set of values X takes with positive probability.

$H(X) = \log |\text{Range}(X)|$ iff X is uniform on its range.

Suppose $\mathcal{C}$ is a set whose size we want to estimate.
If X selects each element of $\mathcal{C}$ uniformly at random then
$H(X) = \log |\mathcal{C}|$, i.e. $|\mathcal{C}| = 2^{H(X)}$ So estimate $H(X)$ to find $|\mathcal{C}|$.