CS258: Information Theory

Fan Cheng
Shanghai Jiao Tong University

http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

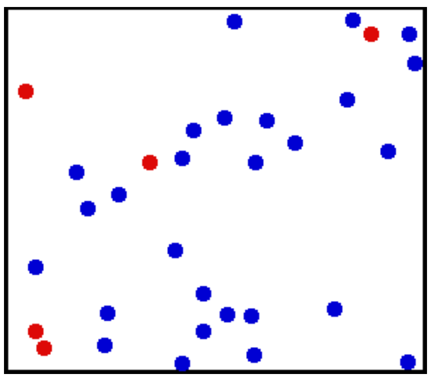
Outline

- Entropy
- Relative entropy
- Mutual information
- Information inequality

Entropy: Brief History



Second law of thermodynamics: one way only



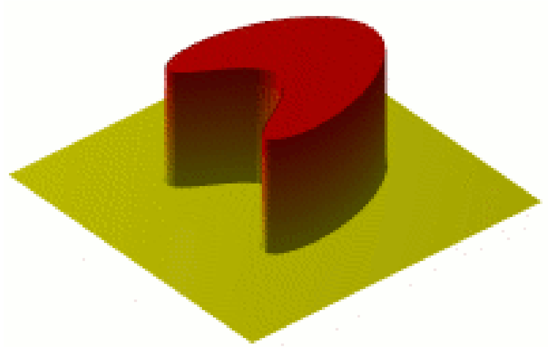
Ludwig Eduard Boltzmann
1844-1906
Vienna, Austrian Empire

- It is hard to analyze the atoms individually
- From the whole system level
- Entropy: quantity for a very complicated system
- Entropy is of great difference from quantities in Newton's law

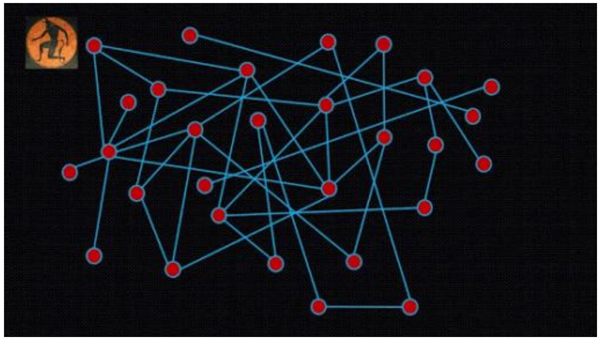
Entropy: Brief History

'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, so it already has a name. In the second place, and more important, **no one really knows what entropy really is,** so in a debate you will always have the advantage.'

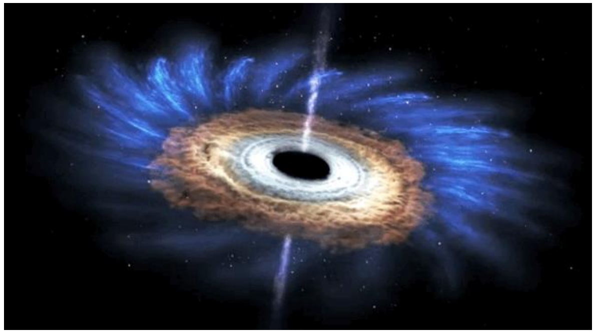
--Shannon explained the name 'entropy'



Thermodynamics



Quantum information



Blackhole

Information is not Matter or Energy. Hard to understand its meaning intuitively.

Entropy: Definition

Notation

- Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \Pr(X = x)\,, x \in \mathcal{X}$.
- For convenience, denote p. m. f. by $p(x)$ rather than $p_X(x)$. Thus $p(x)$ and $p(y)$ are two different p. m. f's.

The entropy of X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

A measure of the uncertainty of a random variable

- $0 \log 0 \to 0\,.\,(x \to 0,\ x \log x \to 0)$
- $H(X)$ only depends on $p(x)$.We also write $H(p)$ for $H(X)$.
- $H(X) \geq 0$
- When X is uniform over $\mathcal{X}$, then $H(X) = \log|\mathcal{X}|$
- $H_b(X) = \log_b a\, H_a(X)$
  - The logarithm is to the base 2 and the unit is bits. If the base of the logarithm is b, we denote of the entropy by $H_b(X)$. If $b = e$, the entropy is measured in nats.
  - Unless otherwise specified, the entropies will be measured in bits.

Entropy: Examples

- Binary entropy function $H(p)$

  Let $X = \begin{cases} 1 & \text{with probablity } p, \\ 0 & \text{with probability } 1-p \end{cases}$

  $$H(X) = -p\log p - (1-p)\log(1-p)$$

  - $H(p)$ is symmetric and concave in p.

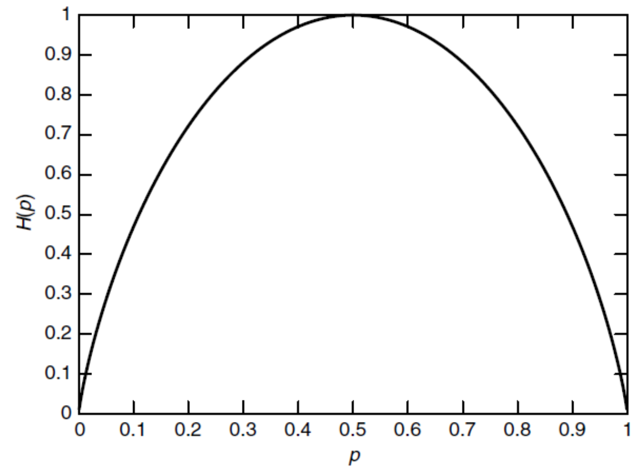


FIGURE 2.1. $H(p)$ vs. p.

- Let

$$X = \begin{cases} a & \text{with prob. } \frac{1}{2} \\ b & \text{with prob. } \frac{1}{4} \\ c & \text{with prob. } \frac{1}{8} \\ d & \text{with prob. } \frac{1}{8} \end{cases}$$

$$H(X) = \frac{7}{4}$$

- We denote expectation by E. If $X \sim p(x)$, the expected value of the random variable $g(X)$ is written

$$E_p g(X) = \sum_{x\in\mathcal{X}} g(x)p(x)$$

$$\boldsymbol{H(X) = E_p \log\frac{1}{p(X)}}$$

Entropy

For a discrete random variable X defined on $\mathcal{X}$,

$$0 \le H(X) \le \log|\mathcal{X}|$$

- When $0 \le x \le 1$, $-x\log x \ge 0$. $x \log x = 0$ iff $x = 0$ or $x = 1$

$$H(X) \ge 0$$

- By definition, we need to prove $\sum_{x\in\mathcal{X}} -p(x)\log p(x) \le \log|\mathcal{X}|$

Facts:

- $f(x) = -x\log x$ is concave in x
- $\sum_x p(x) = 1$

By applying the concavity of $f(x)$,

$$\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} -p(x)\log p(x) \le -\frac{1}{|\mathcal{X}|}\log\frac{\sum_x p(x)}{|\mathcal{X}|} = \frac{1}{|\mathcal{X}|}\log|\mathcal{X}|$$

Equality if and only if $p(x) = 1/|\mathcal{X}|$. (Uniform distribution maximizes entropy)

Convexity (Concavity) is widely applied

$$\sum_i \boldsymbol{p_i f(x_i)} \le \boldsymbol{f}\Big(\sum_i \boldsymbol{p_i\, x_i}\Big)$$

General Roadmap

- Entropy is determined by probability distribution only, and alphabet is not involved

  Probability distribution $\longrightarrow$ Entropy

- For a set of random variables $X_1, X_2, \dots, X_n$ with joint probability distribution $p(x_1, x_2, \dots, x_n)$
  - Joint distribution: $p(x_i, x_j)$
  - Conditional distribution: $p(x_i | \dots)$

  All leads to some "entropy"

- Basic law in Probability theory
  - Chain rule: $p(x_1, x_2, \dots, x_n) = p(x_n)p(x_{n-1}|x_n) \dots p(x_1|x_2, \dots, x_{n-1})$
  - Bayesian rule: $p(y)p(x|y) = p(x)p(y|x)$

  Certain structures exist in "entropies"

Joint entropy, mutual information, chain rule, etc.

Joint Entropy

Facts:

- Two random variables X and Y can be considered to be **a single vector-valued random variable**
- Entropy is defined on probability

The joint entropy $H(X,Y)$ of a pair of discrete random variable (X,Y) with joint distribution $p(x,y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

Entropy and joint entropy

$$H(X) \quad \text{---->} \quad H(X,Y)$$
$$p(x) \quad \text{---->} \quad p(x,y)$$
$$H(X,Y) = -E\log p(X,Y)$$

- $H(X,X) = H(X)$
- $H(X,Y) = H(Y,X)$

For a set of random variables $X_1, \dots, X_n$ with joint distribution $p(x_1, \dots, x_n)$, its joint entropy is defined as

$$H(X_1, X_2, \dots, X_n) = -\sum p(x_1, x_2, \dots, x_n)\log p(x_1, x_2, \dots, x_n) = -E\log p(X_1, \dots, X_n)$$

Conditional Entropy

- When $X = x$ is known, $p(Y|X=x)$ is also a probability distribution

$$\sum_y p(Y=y|X=x) = \sum_y \frac{p(x,y)}{p(x)} = \frac{p(x)}{p(x)} = 1$$

- Entropy for $p(Y|X=x)$

$$H(Y|X=x) = \sum_y -p(y|X=x)\log p(y|X=x) = E\left[-\log p(y|X=x)\right]$$

If $(X,Y) \sim p(x,y)$, the conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= -E\log p(Y|X) \end{aligned}$$

When X is known, the remaining uncertainty of Y: $H(Y|X) \le H(Y)$

Two ways for calculating $H(Y|X)$

Conditional Entropy

Let (X, Y) have the following joint distribution:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

$$H(X,Y) = ?$$
$$H(X) = ?$$
$$H(Y) = ?$$
$$H(Y|X) = ?$$
$$H(X|Y) = ?$$

By $p(x,y)$, one can calculate its marginal distribution:

$$p(x) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$$
$$p(y) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$$
$$p(x|y)$$
$$p(y|x)$$

$$H(X,Y) = \frac{27}{8}$$
$$H(X) = \frac{7}{4}$$
$$H(Y) = 2$$
$$H(X|Y) = \frac{11}{8}$$
$$H(Y|X) = \frac{13}{8}$$

$$\boldsymbol{H(X|Y) \neq H(Y|X)}$$
$$\boldsymbol{H(X|Y) + H(Y) = H(Y|X) + H(X) = H(X,Y)}$$

Make your hands dirty

Chain Rule

Fact: $$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$$

$$\log p(x,y) = \log p(x|y) + \log p(y) = \log p(y|x) + \log p(x)$$

- Probability is not linear, but log function can alleviate it
- Take expectations E:

$$\begin{aligned} &E - \log p(x,y) \\ &= E - \log p(x|y) + E - \log p(y) \\ &= E - \log p(y|x) + E - \log p(x) \end{aligned}$$

> Chain rule
>
> $$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$$

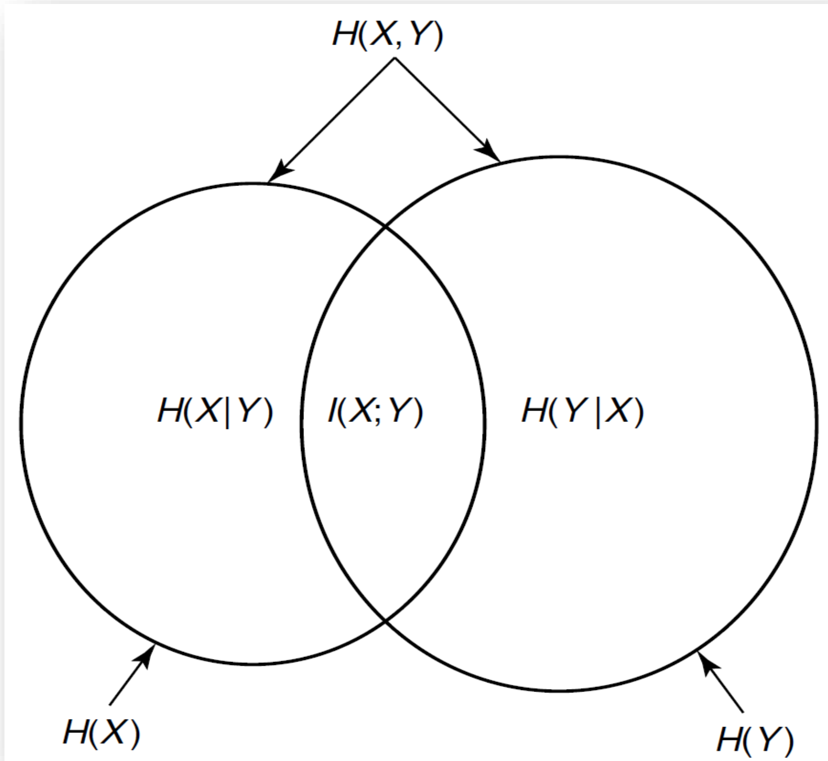
- If X and Y are independent, $H(X,Y) = H(X) + H(Y)$
- If X is a function of Y, $H(X,Y) = H(Y)$
- Bayesian formula
  - $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$. Check $p(x,y|z) = p(x|z)p(y|x,z)$!

> **The underling joint probability $p(x_1, x_2, \dots, x_n)$ determined the relationship of $H(\quad), H(\quad|\quad), etc.$**

Chain Rule: Venn Diagram

Chain rule

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$$



Zero Entropy

Zero conditional entropy: Show that if $H(Y|X) = 0$, then Y is a function of X [i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x,y) > 0$].

Proof sketch:

- When $H(X) = 0$, what is the probability distribution of X?
- Generalize to the condition $H(Y|X = x) = 0$
- Generalize to $H(Y|X) = 0$

Homework: 2.1 2.5 2.7 (Textbook of Cover, Due: 11. p.m., Next Friday)