Information Theory

"A Mathematical Theory of Communication"

by Claude Shannon

1948

"The Mathematical Theory of Communication"

Information Theory

- Source coding
  - compression

- channel coding
  - error-free communication

Information: amount of uncertainity

Ahmedabad	Cherrapunji
Sun: 95%	Sun: 50%
Rain: 5%	Rain: 50%

Which has more uncertainity?

Pass / Fail courses

Course A	Course B
P: 95%	P: 50%
F: 5%	F: 50%

Two coins

Unbiased A	Biased B
H: 0.5	H: 0.8
T: 0.5	T: 0.2

Information: amount of uncertainity

Entropy: Discrete random variable X

$$H(X) = \sum_{i=1}^{n} p_i \log_2\left(\frac{1}{p_i}\right)$$

bits

e.g.: $X \in \{A, B\}$

$$p(X = A) = \frac{1}{2}$$

$$p(X = B) = \frac{1}{2}$$

$$H(X) = \frac{1}{2} \log 2 + \frac{1}{2} \log 2$$

$$= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 \text{ bit}$$

Amount of uncertainity about the outcome of the event before it takes place is 1 bit.

Amount of information gained after the event is 1 bit.

e.g. $X \in \{A, B\}$

$p(X = A) = 0.95$

$p(X = B) = 0.05$

$H(X) = ?$

$$= \sum p_i \log\left(\frac{1}{p_i}\right)$$

$$= 0.95 \log_2\left(\frac{1}{0.95}\right) +$$

$$0.05 \log_2\left(\frac{1}{0.05}\right)$$

$$= 0.286 \text{ bits}$$

if 'e' is the base,
then 'nats' is the unit

---

e.g. You have a fair dice

X = outcome

what is $H(X)$?

$X \in \{1, 2, 3, 4, 5, 6\}$

$p_i \in \{\frac{1}{6}, \ldots, \frac{1}{6}\}$

$$H(X) = \cancel{6} * \frac{1}{\cancel{6}} \log_2 6$$

$$= \log_2 6 = 2.59 \text{ bits}$$

e.g. You have a fair coin
You toss it until you get the first Head. Let X denote the no. of tosses required.
What is $H(X)$?

X	$P(X)$	Event
1	1/2	H
2	1/4	TH
3	1/8	TTH
4	1/16	TTTH
⋮	⋮	

$$H(X) = \sum p_i \log \frac{1}{p_i}$$

$$= \sum \frac{1}{2^i} \log 2^i$$

$$= \sum_{i=1}^{\infty} \frac{i}{2^i} = 2 \text{ bits}$$

---

e.g.

$$\left.\begin{array}{l} A: p \\ B: 1-p \end{array}\right\} X$$

$0 \leq p \leq 1$

Maximize $H(X)$ over p

minimize

maxima is expected at $\frac{1}{2}$, minima at 0 or 1

T H

$$H(X) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$

$$\frac{dH(X)}{dp} = 0$$

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

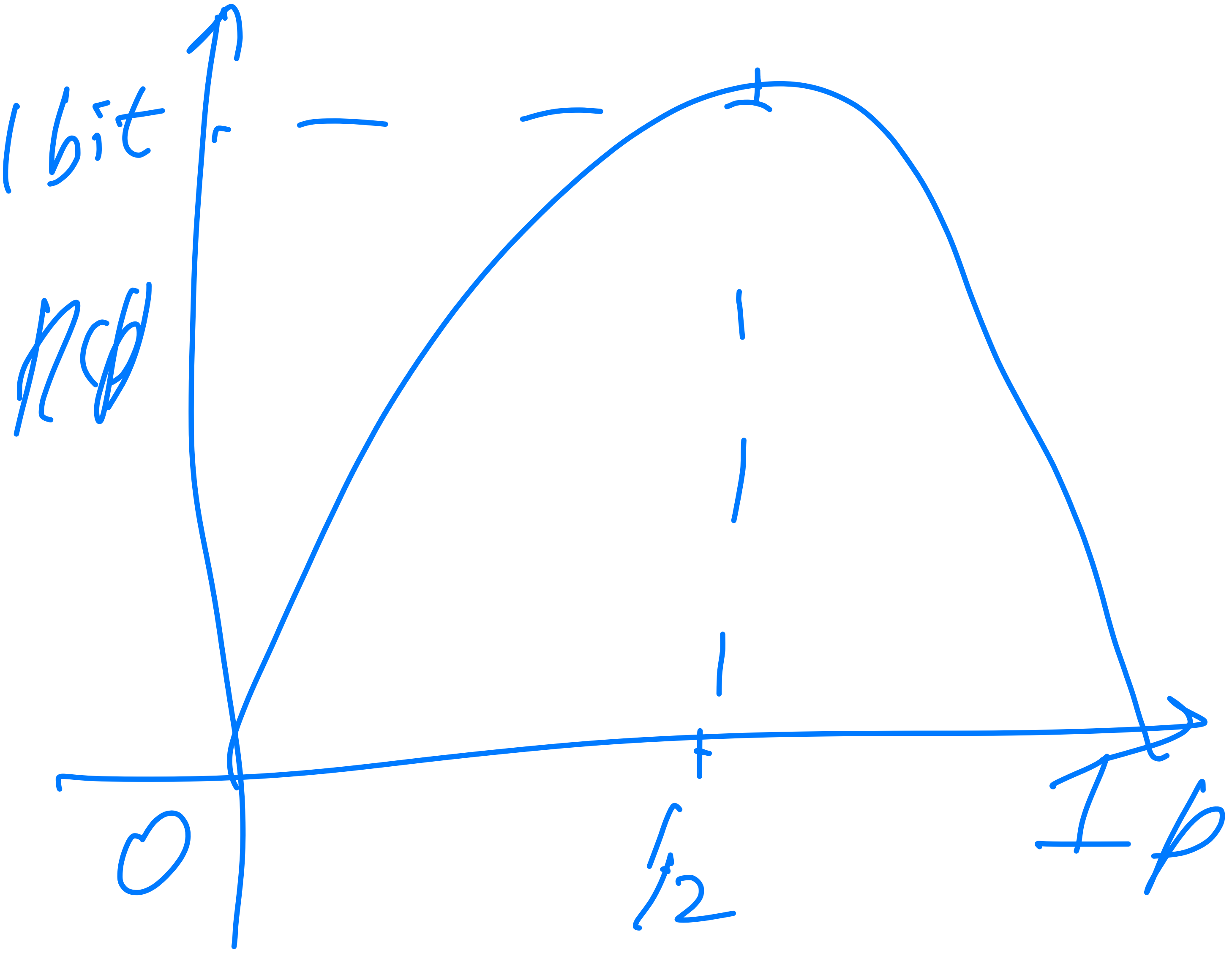
$$\frac{dH(X)}{dp} = -\left[1\cdot\log p + \cancel{p}\cdot\frac{1}{\cancel{p}} + \frac{\cancel{(1-p)}(-1)}{\cancel{1-p}} - 1\cdot\log(1-p)\right]$$

$$= 0$$

$$\log p = \log 1-p$$

$$p = 1-p$$

$$\Rightarrow p = \frac{1}{2}$$



In general $H(X)$ is maximized when all the outcomes are equally likely.

Joint Entropy

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

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

Conditional Entropy

$H(Y|X) =$

$$-\sum_{x}\sum_{y} p(x,y) \log p(y|x)$$

$$= \sum_{x} p(x)\, H(Y|X=x)$$

Both are equivalent definitions

Amount of information left in Y, given X

Y \ X	1	2	3	4	
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
	1/2	1/4	1/8	1/8	

$$H(X) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 = \frac{7}{4}$$

$$H(Y) = \frac{1}{4}\cdot 2 * 4 = 2$$

$$H(X|Y) =$$

$$H(Y|X) =$$

$$H(X,Y) =$$

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

$$= \underbrace{\left(\frac{1}{8}\cdot 2 + \frac{1}{16}\cdot 3 + \frac{1}{16}\cdot 3 + \frac{1}{4}\cdot 1\right)}_{x=1} +$$

$$\underbrace{\left(\qquad \right)}_{x=2} +$$

$$\underbrace{\left(\qquad \right)}_{x=3} +$$

$$\underbrace{\left(\qquad \right)}_{x=4}$$

$$=$$

$$H(X|Y) = \frac{11}{8} \text{ bits}$$

$$H(Y|X) = \frac{13}{8}$$

$$H(X,Y) = \frac{27}{8}$$

$$H(X) = \frac{7}{4} = \frac{14}{8}$$

$$H(Y) = 2 = \frac{16}{8}$$

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

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

Chain Rule

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

The amount of information in X,Y is equal to the amount of information in X plus the amount of information in Y given X.

$$H(X_1, X_2, \ldots, X_n) =$$

$$H(X_1) + H(X_2 | X_1) +$$

$$H(X_3 | X_1, X_2) +$$

$$H(X_4 | X_1, X_2, X_3) + \cdots +$$

$$H(X_n | X_1, X_2, \ldots, X_{n-1})$$

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

$$H(X) - H(X|Y)$$

$$= I(X; Y)$$

Mutual information of X and Y

Source Coding:

Source code C for a random variable X is a mapping from the range of X to $\{0,1\}^*$

$= $ all possible bit strings

$= \epsilon, 0, 1, 00, 01, 10, 11, \ldots$

x	$p(x)$	$C_1(x)$	$l_1(x)$	$C_2(x)$	$l_2(x)$
1	1/2	00	2	000	3
2	1/4	01	2	010	3
3	1/8	10	2	1	1
4	1/8	11	2	11111	5

$L(C)$ = average length of a code C

$= \sum p_i l_i = 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8}$

$= 2$ bits. $= l(C_1)$

$l(C_2) = 3 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 5 \cdot \frac{1}{8}$

$= 3$ bits

x	$C_3(x)$	$l_3(x)$	$p(x)$
1	0	1	1/2
2	10	2	1/4
3	110	3	1/8
4	111	3	1/8

$$l_3 = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8}$$

$$= \frac{7}{4} \text{ bits}$$

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3$$

$$= \frac{7}{4} \text{ bits}$$

$$= 1.75 \text{ bits}$$

x	$p(x)$	$C_4(x)$	l
1	1/2	0	1
2	1/4	1	1
3	1/8	10	2
4	1/8	01	2

would you use this code?

$$l_4 = 1\cdot\frac{1}{2} + 1\cdot\frac{1}{4} + 2\cdot\frac{1}{8} + 2\cdot\frac{1}{8}$$

$$= \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{5}{4} \text{ bits}$$

$$= 1.25 \text{ bits}$$